UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Contribuciones a la metodología para estimar efectos causales en estudios longitudinales observacionales

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires en el área Ciencias Matemáticas

Lucía Babino

Director de tesis: Dra. Andrea G. Rotnitzky
Consejero de estudios: Dra. Mariela R. Sued

Lugar de trabajo: Instituto de Cálculo, FCEyN, UBA

Buenos Aires, junio de 2019

Contribuciones a la metodología para estimar efectos causales en estudios longitudinales observacionales

Esta tesis contribuye a la estimación de efectos causales de tratamientos variantes en el tiempo en presencia de variables confusoras variantes en el tiempo que se ven afectadas por el tratamiento recibido en el pasado. La tesis consta de dos capítulos.

El primer capítulo contribuye a la estimación múltiple robusta paramétrica de modelos estructurales marginales. Específicamente, hacemos propuestas de estimación, en base a datos recogidos de estudios longitudinales observacionales, de los parámetros de los modelos marginales estructurales para la media (MMEM) para variables de respuesta no acotadas. Actualmente, los métodos populares utilizados en las aplicaciones para estimar los parámetros de los MMEM incluyen a los estimadores "inverse probability of treatment weighted" y a los estimadores paramétricos doble robustos (DR). Bajo la metodología paramétrica DR, el investigador postula una secuencia de modelos de trabajo paramétricos, un modelo para la media de la variable de respuesta contrafactual dado el historial de covariables y tratamientos hasta cada instante de tiempo de exposición (que, a lo largo de este resumen, denominamos media contrafactual del instante de tiempo específico) y otra secuencia de modelos de trabajo, un modelo para la probabilidad de tratamiento en cada instante de tiempo condicional a los tratamientos y las covariables del pasado que, a lo largo de este resumen, denominamos probabilidad de tratamiento del instante de tiempo específico. Los estimadores DR de los parámetros de los MMEM son consistentes y asintóticamente normales siempre y cuando o bien la secuencia de modelos de trabajo para las medias contrafactuales de cada instante de tiempo específico sea correcta o bien la secuencia de modelos para las probabilidades de tratamiento sea correcta, pero no necesariamente ambas secuencias de modelos sean correctas.

Una dificultad con la estimación DR paramétrica es que la mayoría de los modelos naturales para las medias contrafactuales de cada instante de tiempo específico son usualmente incompatibles. Robins, Rotnitzky y Scharfstein (2000) propusieron una parametrización de la verosimilitud que implica modelos paramétricos compatibles para dichas medias. Esta parametrización no se ha explotado para construir estimadores DR y uno de los objetivos del primer capítulo es llenar este vacío. Más importante aún, al explotar esta parametrización, proponemos un estimador múltiple robusto (MR) de los parámetros de un MMEM que otorga una protección aún mayor contra la especificación errónea de los modelos que los estimadores DR, ya que el estimador tiene la propiedad múltiple robusta de ser consistente y asintóticamente normal siempre y cuando, en cada instante de tiempo, o bien el modelo de trabajo para la media contrafactual o bien el modelo de trabajo para la probabilidad de tratamiento sea correcto, pero no necesariamente ambos lo sean. Nuestros métodos son de fácil implementación ya que se basan en el ajuste iterativo de una secuencia de regresiones ponderadas.

El segundo capítulo explora y contrasta los méritos relativos de los estimadores no paramétricos doble y múltiple robustos de la media de una variable de respuesta contrafactual medida al final de un estudio longitudinal. Cuando hablamos de estimador no paramétrico doble robusto (o múltiple robusto) nos referimos a uno que se calcula siguiendo un procedimiento que produciría un estimador con la propiedad doble (o múltiple) robusta si las medias contrafactuales y las probabilidades de tratamiento de cada tiempo específico se hubieran estimado a tasas paramétricas, pero en el que estas funciones desconocidas se estiman de manera no paramétrica utilizando, por ejemplo, estimación por series, núcleo, spline o, más generalmente, cualquier estimador de aprendizaje automático. Las

contribuciones centrales de este capítulo son (1) la derivación de expresiones novedosas para el sesgo asintótico de los estimadores DR y MR no paramétricos y (2) el cálculo de cotas para las tasas de convergencia de estos sesgos cuando asumimos que las medias contrafactuales y las probabilidades de tratamiento desconocidas pertenecen a bolas Hölder y son estimadas mediante estimación por series. Nuestros análisis sugieren que, en lo que respecta a conseguir estimadores $\sqrt{n}-$consistentes de la media contrafactual al final del estudio, nunca es contraproducente y, bajo algunos procesos de generación de datos, es preferible realizar estimación MR no paramétrica que realizar estimación DR.

Palabras claves: modelos compatibles, modelos marginales estructurales para la media, estimación doble robusta, estimación múltiple robusta, g-fórmula, estimación no paramétrica.

Contributions to methods for estimating causal effects from longitudinal observational studies

This thesis makes contributions to the estimation of causal effects of time-dependent exposures in the presence of time-dependent confounders that are themselves affected by previous treatments. The thesis is comprised of two chapters.

The first chapter makes contributions to the parametric multiple robust estimation of marginal structural models. Specifically, we consider estimation, from longitudinal observational data, of the parameters of marginal structural mean models (MSMM) for unconstrained outcomes. Currently popular methods used in applications for estimating parameters of MSMM include inverse probability of treatment weighted and parametric doubly robust (DR) estimators. Under the parametric DR methodology the investigator postulates a sequence of parametric working models, one model for the mean of the counterfactual outcome given the covariate and treatment history up to each exposure time point -throughout this abstract referred to as the time specific counterfactual mean- and another sequence of working models, one model for the conditional probability of treatment at each time given past treatments and covariates -throughout referred to as the time specific propensity score-. The DR estimators of the parameters of MSMM have the doubly robust property that they are consistent and asymptotically normal so long the sequence of working models for the time specific counterfactual means are correct or the sequence of models for the propensity scores are correct, but not necessarily both sequences of models are correct.

A difficulty with parametric DR estimation is that most natural models for the time specific counterfactual means are often incompatible. Robins, Rotnitzky and Scharfstein (2000) proposed a parameterization of the likelihood which implies compatible parametric models for such means. Their parameterization has not been exploited to construct DR estimators and one goal of the first chapter is to fill this gap. More importantly, exploiting this parameterization we propose a multiple robust (MR) estimator of the parameters of a MSMM that confers even more protection against model misspecification than DR estimators in that the estimator has the multiple robust property that it is consistent and asymptotically normal so long at each time, either the working model for the counterfactual mean or the working model for the propensity score is correct, but not necessarily both. Our methods are easy to implement as they are based on the iterative fit of a sequence of weighted regressions.

The second chapter explores and contrasts the relative merits of non-parametric doubly and multiply robust estimators of the mean of a counterfactual outcome measured at the end of a longitudinal study. By a non-parametric doubly robust (multiply robust) estimator we mean one that is computed following a procedure which would yield an estimator with the double (multiple) robust property if the time specific counterfactual means and the time specific propensity scores had been estimated at parametric rates, but in which these unknowns functions are instead estimated non-parametrically, e.g. using series, kernel, spline or more generally arbitrary machine learning estimators. The key contributions of this chapter are (1) the derivation of novel expressions for the asymptotic bias of the non-parametric DR and MR estimators and (2) the calculation of bounds on the rates of convergence of these biases when the unknown time specific counterfactual means and propensity scores are assumed to belong to Hölder balls and are estimated by series estimation. Our analyses suggest that as far as achieving $\sqrt{n}-$consistent estimators of the counterfactual mean at the end of the study is concerned, it never hurts and, under some data generating processes,

it sometimes helps to conduct non-parametric MR estimation as opposed to non-parametric DR estimation.

Key words: compatible models, marginal structural mean models, doubly robust estimation, multiply robust estimation, g-formula, non-parametric estimation.

Agradecimientos

A Andrea, por haber aceptado dirigir esta tesis, por contagiarme su entusiasmo por la Inferencia Causal, por todo lo que aprendí de ella.

A Víctor por haberme dado el puntapié inicial para comenzar el doctorado. A Víctor, a Ana y a Andrea, por los cursos dictados, que fueron fundamentales para mi formación doctoral.

A mis compañeros del Instituto de Cálculo, a Mariela, Daniela, Marina, Maru, Flor F, Flor S, Inés, Manuel, Gonzalo, Julieta, Agustín, porque es un privilegio trabajar con un grupo tan solidario, por las charlas, los almuerzos y porque todos me ayudaron en distintas instancias de mi doctorado. En particular a Maru, por compartir la oficina casi hasta el final, por su generosidad y sus consejos.

A los integrantes del Instituto de Ecología Genética y Evolución de Buenos Aires, por la calidez con la que me recibieron, por todo lo que aprendí trabajando con ellos.

A mis amigas de siempre, a Naty, Meli, Ana, Male, Ale y Coty, por estar siempre dispuestas a escuchar y a dar buenos consejos, por tantos momentos lindos compartidos.

A mis hijos, Sofía y Vicente, por el amor y la alegría de todos los días.

A Isabel y Ramón, por haber cuidado tantas horas a Sofía, con tanto amor, cuando esta tesis recién comenzaba.

A mis papás, por su apoyo incondicional en este proyecto como en todos los que emprendí en mi vida. A mi hermano Andy, por estar siempre que lo necesito.

Y en especial a Marcelo, por el amor, el sostén, la confianza y el enorme esfuerzo que hizo para que esta tesis fuera posible.

Contents

Chapter 1

Parametric Multiple Robust Estimation of Marginal Structural Mean Models for an unbounded outcome

1.1 Introduction

Marginal Structural Mean Models (MSMM) are popular tools to model the causal effect of a time-dependent exposure in the presence of time-dependent confounders that are themselves affected by previous treatment. Since they were first proposed by Robins ([30]), MSMMs have been applied to analyze numerous health-related studies. For example, studies of, the effect of highly active antiretroviral therapy on CD4 count ([9]), the effect of pillbox organizer use on adherence to antiretroviral medications and viral load ([28]) and the effect of loneliness on depressive symptoms ([59]).

Currently popular methods used in applications to estimate the parameters of MSMMs include inverse probability of treatment weighted (IPTW) estimation ([34]; [35] and [38]) and doubly robust (DR) estimation ([36]; [25]; [1]; [62];[27]; [56]).

IPTW estimation requires that the analyst postulates a sequence of models, each model parameterizing the dependence of each occasion-specific propensity score (PS), i.e., of the probability of treatment assignment at each time point, on past treatments and covariates. Consistency of IPTW estimators is guaranteed only when all the postulated models are correct. On the other hand, DR estimators require that the analyst postulates two sequences of models, one sequence being the sequence of PS models. The second sequence of models parameterizes, for each time point, the mean of the counterfactual outcome given the covariate and treatment history up to that time point. The estimators are consistent provided one, but not necessarily both, of these sequences of models is correct.

A difficulty with DR estimation is that a model for a counterfactual mean given covariate and treatment history up to the given exposure time point usually imposes restrictions on the

counterfactual mean given the covariate history up to any earlier exposure time point. A practical implication of this technicality is that often it is difficult to postulate compatible models for the sequence of counterfactual means. To our knowledge, for DR estimation based on parametric models for the counterfactual means, no general approach to address model incompatibility exists. One goal of this chapter is to fill this gap in the special case in which the outcome is continuous and unbounded.

To arrive at our proposed DR estimator we exploit a parameterization of the likelihood discussed in [42] which implies compatible parametric models for the counterfactual means. In fact, exploiting this parameterization we additionally propose a multiply robust (MR) estimator that confers even more protection against model misspecification than DR estimators. Specifically, letting K denote the total number of exposure time points, we propose an estimator that is consistent and asymptotically normal (CAN) so long as for some k in $\{0, 1, ..., K\}$, regardless of which k, the models for the first k counterfactual means and the last $K - k$ PS models are correctly specified. Our MR estimator is simple to implement. Its computation requires, simply, the fit of a sequence of regressions.

Our multiply robust estimator is inspired by the work of Tchetgen Tchetgen ([53]) and Molina et al. ([24]). The former article describes an augmented inverse probability weighted estimator of the mean of a missing at random outcome that is multiply robust. The latter article develops theory for the construction of MR estimators in factorized likelihood models. However, these articles do not discuss remedies for model incompatibility. In addition, our proposal differs from the ones in these articles in that it is based on the iterative fit of a sequence of regressions.

Our proposal contributes to the growing literature on MR estimation that includes recent MR estimators for methods for natural indirect and direct causal effects ([54]; [64] and [55]), for statistical interactions ([61]) and for missing data ([60]; [53]; [6]; [19]; [17];[16] and [18]).

Sections 1.3 and 1.4 review MSMMs and existing estimators respectively. Section 1.5 discusses compatible models for the sequence of counterfactual means and Section 1.6 presents DR and MR estimators that use these models, for the case in which the number of exposure time points is $K = 2$. In Section 1.7, we generalize our proposal for the case of arbitrary K. In Section 1.8, we illustrate our methods with an analysis of data from the National Heart Lung and Blood Institute Growth and Health Study. In Section 1.9 we present a simulation study. In Section 1.10, we prove the consistency and asymptotic normality of our MR estimator. Finally, in Section 1.11, we generalize our proposal to the case of repeated outcomes.

1.2 Notation

In this section we summarize the notation that will be used through the chapter.

For $k \in \{1, ..., K\}$, $j \in \{1, ..., k\}$ and any $\{v_r\}_{1 \leq r \leq K}$, we let

$$\overline{v}_k \equiv (v_1, \ldots, v_k), \underline{v}_k \equiv (v_k, \ldots, v_K) \text{ and } \overline{v}_j^k \equiv (v_j, \ldots, v_k).$$

Also, for $k \in \{1, ..., K\}$ and any collection of functions $\{f_j\}_{1 \leq j \leq K}$, we let

$$\overline{f}_k \equiv (f_1, \ldots, f_k), \underline{f}_k \equiv (f_k, \ldots, f_K).$$

Likewise, for $k \in \{1, ..., K\}$, $j \in \{1, ..., k\}$ and any collection of sets $\{\mathcal{C}_r\}_{1 \leq r \leq K}$, we let

$$\overline{\mathcal{C}}_k \equiv \mathcal{C}_1 \times \cdots \times \mathcal{C}_k, \underline{\mathcal{C}}_k \equiv \mathcal{C}_k \times \cdots \times \mathcal{C}_K \text{ and } \overline{\mathcal{C}}_j^k \equiv \mathcal{C}_j \times \cdots \times \mathcal{C}_k$$

For any vector $v = [v_i]_{1 \leq i \leq p} \in R^p$, $\|v\|$ denotes its Euclidean norm $\left(\sum_i v_i^2\right)^{1/2}$.

For any matrix $A = [a_{ij}]_{1 \leq i \leq p, 1 \leq j \leq p} \in R^{p \times q}$, $\|A\|$ denotes its Euclidean norm $\sup\{\|Av\| : v \in R^p \text{ with } \|v\| = 1\}$

For any function $f : \mathcal{X} \subseteq R^p \to R^q$, $\|f\|_\infty$ denotes $\sup_{x \in \mathcal{X}} \|f(x)\|$.

For $X_1, \ldots, X_n$ i.i.d. copies of a random vector X with law P and range in $\mathcal{X} \subseteq R^p$, $\mathbb{P}_n(X)$ denotes the sample average $\frac{1}{n}\sum_{i=1}^n X_i$. Also, for any function $f : \mathcal{X} \to R^q$, $\mathbb{P}_n(f)$ and $E_P(f)$ denote $\mathbb{P}_n\{f(X)\}$ and $E_P\{f(X)\}$ respectively.

1.3 Marginal structural mean model for an unconstrained outcome.

Suppose that i.i.d. copies of

$$O \equiv (L_1, A_1, \ldots, L_K, A_K, Y)$$

are collected on n subjects. Variable Y is an outcome of interest at time t_{K+1} which is unconstrained, i.e. with range in the real line. For $k = 1, \ldots, K$, A_k is the treatment given at time t_k taking values in a finite set $\mathcal{A}_k$ and L_k is a vector of covariates measured at time t_k^- , i.e. an instant prior to t_k $(t_{k-1} < t_k)$, taking values in a subset $\mathcal{L}_k$ of a Euclidean space.

For each treatment history $\overline{a}_K = (a_1, \ldots, a_K)$, let $Y_{\overline{a}_K}$ be the subject's response if, possibly contrary to fact, treatment regime $\overline{a}_K$ is followed. Under the assumptions of

(1) consistency:

$$Y_{\overline{a}_K} = Y \text{ if } \overline{A}_K = \overline{a}_K$$

(2) no unmeasured confounding (NUC): for all $\overline{a}_K$ and k,

$$Y_{\overline{a}_K} \amalg A_k | \overline{L}_k, \overline{A}_{k-1} = \overline{a}_{k-1}$$

and

(3) positivity: for all k and $\overline{a}_k$, if $f(\overline{a}_{k-1}, \overline{l}_k) > 0$ then $f(a_k | \overline{a}_{k-1}, \overline{l}_k) > 0$,

it is well known ([32]) that $E(Y_{\overline{a}_K}|Z)$ is identified, where Z is a subvector of L_1. Throughout, we shall refer to (1)-(3) as the *identifying assumptions*. In this chapter, we make the identifying assumptions and consider estimation of the parameter $\psi^* \in R^p$ of the MSMM ([30])

$$E\left(Y_{\overline{a}_K}|Z\right) = m\left(\overline{a}_K, Z; \psi^*\right) \quad \text{for all } \overline{a}_K, \tag{1.1}$$

where $m\left(\cdot, \cdot; \cdot\right)$ is specified.

Throughout, we write $L_1 = (Z, V)$. Also, we say that an estimator $\widehat{\psi}$ of ψ^* is consistent and asymptotically normal (CAN) under a given model $\mathcal{M}$ if $\sqrt{n}\left(\widehat{\psi} - \psi^*\right)$ converges to a mean zero Normal distribution under any law that satisfies model $\mathcal{M}$.

1.4 Existing estimators.

1.4.1 Inverse probability weighted estimators

Under the identifying assumptions, Robins ([35]) proved that (1.1) is equivalent to a model for the observed data O defined by

$$E\left[\pi^p\left(\overline{A}_K,\overline{L}_K\right)^{-1}\left\{Y-m\left(\overline{A}_K,Z;\psi^*\right)\right\}\middle|\,\overline{A}_K,Z\right]=0,$$

where

$$\pi^p\left(\overline{a}_K,\overline{l}_K\right)\equiv\prod_{j=1}^{K}\pi_j\left(\overline{a}_j,\overline{l}_j\right)$$

with

$$\pi_1\left(a_1,l_1\right)\equiv\Pr\left(A_1=a_1|L_1=l_1\right),$$
$$\pi_j\left(\overline{a}_j,\overline{l}_j\right)\equiv\Pr\left(A_j=a_j|\overline{A}_{j-1}=\overline{a}_{j-1},\overline{L}_j=\overline{l}_j\right),$$

$2\leq j\leq K$. This observation gave rise to the so-called IPTW estimator $\widehat{\psi}_{IPTW}$ which is obtained by fitting a weighted least squares regression with outcome Y and covariates $(\overline{A}_K,Z)$. The weights are given by the inverse of the maximum likelihood estimator (MLE) $\widehat{\pi}^p\left(\overline{A}_K,\overline{L}_K\right)$ of $\pi^p\left(\overline{A}_K,\overline{L}_K\right)$ under a parametric model for the treatment probabilities $\pi_j,1\leq j\leq K$. Under regularity conditions, the estimator $\widehat{\psi}_{IPTW}$ is CAN under the assumed parametric model for $\pi_j,1\leq j\leq K$. However, it may not even converge in probability to ψ^* if any of the π_j is incorrectly modeled.

1.4.2 Iterative conditional expectation estimators

An alternative estimator of ψ^* can be obtained from the following observation. Let

$$\eta_K(\overline{a}_K,\overline{l}_K)\equiv E\left(Y|\overline{A}_K=\overline{a}_K,\overline{L}_K=\overline{l}_K\right)$$

and, for $k=K-1,K-2,\ldots,1$, let

$$\eta_k(\overline{a}_K,\overline{l}_k)\equiv E\left\{\eta_{k+1}(\overline{a}_K,\overline{L}_{k+1})|\overline{A}_k=\overline{a}_k,\overline{L}_k=\overline{l}_k\right\}.$$

Also let

$$\eta_0\left(\overline{a}_K,z\right)\equiv E\left\{\eta_1(\overline{a}_K,L_1)|Z=z\right\}.$$

It follows from Theorem 3.2 of [29], (see also, [33]) that under the identifying assumptions,

$$\eta_k\left(\overline{a}_K,\overline{L}_k\right)=E\left(Y_{\overline{a}_K}|\overline{A}_k=\overline{a}_k,\overline{L}_k\right),$$

$k=1,\ldots,K$, and

$$\eta_0\left(\overline{a}_K,Z\right)=E\left(Y_{\overline{a}_K}|Z\right).$$

Hence, under these assumptions, model (1.1) is equivalent to a model for the observed data O defined by the sole restriction

$$\eta_0\left(\overline{a}_K,Z\right)=m\left(\overline{a}_K,Z;\psi^*\right)\text{ for all }\overline{a}_K. \tag{1.2}$$

This observation suggests postulating parametric models $\eta_k(\overline{a}_K, \overline{L}_k) = \eta_k(\overline{a}_K, \overline{L}_k; \delta_k^*), 1 \leq k \leq K$, and computing an iterated conditional expectation (ICE) estimator $\widehat{\psi}_{ICE}$ of ψ^* as follows. First compute $\widehat{\delta}_K$ solving

$$\mathbb{P}_n \left[\frac{\partial}{\partial \delta_K} \eta_K(\overline{A}_K, \overline{L}_K; \delta_K) \left\{ Y - \eta_K(\overline{A}_K, \overline{L}_K; \delta_K) \right\} \right] = 0.$$

Then, for $k = K-1, K-2, \ldots, 1$ iteratively compute $\widehat{\delta}_k$ solving

$$\mathbb{P}_n \left[\sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{\partial}{\partial \delta_k} \eta_k(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \delta_k) \left\{ \eta_{k+1}(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1}; \widehat{\delta}_{k+1}) - \eta_k(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \delta_k) \right\} \right] = 0.$$

Here recall that, $\underline{v}_k = (v_k, \ldots, v_K)$ for any $\{v_j\}_{1 \leq j \leq K}$ and $\underline{\mathcal{A}}_j = \mathcal{A}_j \times \cdots \times \mathcal{A}_K$, $j = 1, \ldots, K$. Finally, $\widehat{\psi}_{ICE}$ solves

$$\mathbb{P}_n \left[\sum_{\underline{a}_1 \in \underline{\mathcal{A}}_1} \frac{\partial}{\partial \psi} m(\underline{a}_1, Z; \psi) \left\{ \eta_1(\underline{a}_1, L_1; \widehat{\delta}_1) - m(\underline{a}_1, Z; \psi) \right\} \right] = 0.$$

Under regularity conditions, this estimator is CAN under the assumed model for all the $\eta_k, 1 \leq k \leq K$. However, it may not even converge in probability to ψ^* if any of the η_k is incorrectly modeled.

1.4.3 Doubly robust estimators

Bang and Robins ([1]) discussed a DR estimator of ψ^* which weakens reliance on modelling assumptions by offering the opportunity to avoid committing to one specific modelling strategy. In Sections 1.6.3 and 1.7.2 we describe a slightly different DR estimator. Other DR estimators were described in [36]; [25]; [14] and [27]. For computing a DR estimator the data analyst postulates a model for the treatment probabilities π_k and also a model for the functionals $\eta_k, 1 \leq k \leq K$. The estimator is CAN for ψ^* under the union model that assumes that either the model for the $\pi_k, 1 \leq k \leq K$, or the model for the $\eta_k, 1 \leq k \leq K$, is true, but not necessarily both are.

1.5 Compatible parametric working models for the η_k for the special case $K = 2$

A model for η_{k+1} implies restrictions on the range of possible $\eta_k's$ because $\eta_k(\overline{a}_K, \overline{l}_k) = E\left\{\eta_{k+1}(\overline{a}_K, \overline{L}_{k+1})|\overline{A}_k = \overline{a}_k, \overline{L}_k = \overline{l}_k\right\}, k = 1, \ldots, K-1$. Likewise, a model for η_1 implies restrictions on η_0. This makes the task of simultaneously modeling all the η_k challenging. For instance, if $K = 2$, the linear model that assumes that $\eta_2(\overline{a}_2, \overline{l}_2) = \alpha_0 + \alpha_1 a_1 + \alpha_2 a_2 + \beta_1' l_1 + \beta_2' l_2$ implies that $\eta_1(\overline{a}_2, l_1) = \alpha_0 + \alpha_1 a_1 + \alpha_2 a_2 + \beta_1' l_1 + \beta_2' E\left(L_2|A_1 = a_1, L_1 = l_1\right)$. If L_2 is binary and L_1 is unbounded, then $E\left(L_2|A_1, L_1\right)$ must be a non-linear function of L_1 unless it is independent of it. Thus, if the linear model for η_2 is correct, except in the extreme scenarios in which either $\beta_2 = 0$ or $E\left(L_2|A_1, L_1\right)$ is independent of L_1, the linear model $\eta_1\left(\overline{a}_2, l_1\right) = \gamma_0 + \gamma_1 a_1 + \gamma_2 a_2 + \delta_1' l_1$ that many analysts would naturally postulate is incorrectly specified.

Robins et al. ([42]) proposed a parameterization of the likelihood which implies compatible parametric models for all the $\eta_k's$ with some shared parameters. By compatible η_k models with shared parameters, we mean that for every combination of parameter values indexing the models for the distinct η_k and agreeing on the shared parameters, it is possible to find at least one distribution for the observed data O that satisfies all models. For instance, when $K = 2$, L_2 is binary and L_1 is unbounded, Robins et al. parameterization results in compatible parametric models for $\eta_1(\overline{a}_2, L_1)$ and $\eta_2(\overline{a}_2, \overline{L}_2)$ that, in fact, does not exclude the possibility of simultaneous dependence of $E\left(L_2|A_1, L_1\right)$ on L_1 and of $\eta_2(\overline{a}_2, \overline{L}_2)$ on L_2. As far as we know, Robins et al.'s parameterization has not been exploited to construct DR estimators. One goal of this chapter is to fill this gap. More importantly, exploiting this parameterization we propose an MR estimator of ψ^* whose implementation requires but just a slight modification of the procedure for computing the ICE estimator. Our MR estimator confers even more protection against model misspecification than DR estimators. For didactical reasons we describe our proposal when $K = 2$, i.e. under a follow-up study with observed data

$$O = (L_1, A_1, L_2, A_2, Y)$$

where $L_1 = (Z, V)$. The case of arbitrary K is discussed in Section 1.7. Note that when $K = 2$,

$$\eta_2\left(\overline{a}_2, \overline{L}_2\right) \equiv E\left(Y|\overline{A}_2 = \overline{a}_2, \overline{L}_2\right),$$

$$\eta_1\left(\overline{a}_2, L_1\right) \equiv E\left\{\eta_2\left(\overline{a}_2, \overline{L}_2\right)|A_1 = a_1, L_1\right\}$$

and

$$\eta_0\left(\overline{a}_2, Z\right) \equiv E\left\{\eta_1\left(\overline{a}_2, L_1\right)|Z\right\}.$$

When $K = 2$, Robins et al.'s parameterization depends on $f\left(V|Z\right), f\left(L_2|L_1, A_1\right), \eta_0\left(\overline{a}_2, z\right)$,

$$\rho_1\left(\overline{a}_2, l_1\right) \equiv \eta_1\left(\overline{a}_2, l_1\right) - \eta_1\left(\overline{a}_2, z, v = v_0\right)$$

and

$$\rho_2\left(\overline{a}_2, \overline{l}_2\right) \equiv \eta_2\left(\overline{a}_2, \overline{l}_2\right) - \eta_2\left(\overline{a}_2, l_1, l_2 = l_{2,0}\right),$$

where v_0 and $l_{2,0}$ are any baseline levels of V and L_2 respectively and $l_1 \equiv (z, v)$. In Appendix A.1, we show that η_0, ρ_1, $\rho_2, f\left(V|Z\right)$ and $f\left(L_2|L_1, A_1\right)$ are variation independent functions in the sense that fixing one or several of them does not restrict the range of the remaining ones. Note that, under the identifying assumptions,

$$\rho_1\left(\overline{a}_2, l_1\right) = E\left(Y_{\overline{a}_2}|A_1 = a_1, L_1 = l_1\right) - E\left(Y_{\overline{a}_2}|A_1 = a_1, Z = z, V = v_0\right)$$

and

$$\rho_2\left(\overline{a}_2,\bar{l}_2\right)=E\left(Y_{\overline{a}_2}|\overline{A}_2=\overline{a}_2,L_1=l_1,L_2=l_2\right)-E\left(Y_{\overline{a}_2}|\overline{A}_2=\overline{a}_2,L_1=l_1,L_2=l_{2,0}\right).$$

Thus, $\rho_1\left(\overline{a}_2,l_1\right)$ quantifies the extent to which the mean of $Y_{\overline{a}_2}$ differs across strata of the baseline covariate V among subjects that received treatment $A_1=a_1$ and had baseline level $Z=z$. Likewise, $\rho_2\left(\overline{a}_2,\bar{l}_2\right)$ quantifies the dependence of the counterfactual outcome mean on L_2 among subjects that received the treatment sequence $\overline{A}_2=\overline{a}_2$ and had baseline covariates $L_1=l_1$. Straightforward calculations yield

$$\eta_1\left(\overline{a}_2,l_1\right)=\eta_0\left(\overline{a}_2,z\right)+\rho_1\left(\overline{a}_2,l_1\right)-E\left\{\rho_1\left(\overline{a}_2,L_1\right)|Z=z\right\}, \tag{1.3}$$

$$\begin{aligned}\eta_2\left(\overline{a}_2,\bar{l}_2\right)&=\eta_0\left(\overline{a}_2,z\right)+\rho_1\left(\overline{a}_2,l_1\right)-E\left\{\rho_1\left(\overline{a}_2,L_1\right)|Z=z\right\}\\&\quad+\rho_2\left(\overline{a}_2,\bar{l}_2\right)-E\left\{\rho_2\left(\overline{a}_2,\overline{L}_2\right)|A_1=a_1,L_1=l_1\right\}.\end{aligned} \tag{1.4}$$

These identities imply that parametric models for ρ_1,ρ_2, $E\left\{\rho_1\left(\overline{a}_2,L_1\right)|Z\right\}$ and $E\left\{\rho_2\left(\overline{a}_2,\overline{L}_2\right)|A_1,L_1\right\}$ and the MSMM $\eta_0\left(\overline{a}_2,z\right)=m\left(\overline{a}_2,z;\psi^*\right)$ determine compatible parametric models for η_0,η_1 and η_2 with shared parameters. For reasons explained in Section 1.6.5, we propose modeling ρ_1 and ρ_2 as

$$\rho_1\left(\overline{a}_2,L_1\right)=g_1(\overline{a}_2,Z;\gamma_1^*)'t_1\left(V\right)\text{ and }\rho_2\left(\overline{a}_2,\overline{L}_2\right)=g_2(\overline{a}_2,L_1;\gamma_2^*)'t_2\left(L_2\right), \tag{1.5}$$

where g_1 and g_2 are user specified vector-valued functions, γ_1^* and γ_2^* are finite dimensional parameters, and t_1 and t_2 are user-specified conformable vector-valued functions verifying $t_1\left(v_0\right)=0$ and $t_2\left(l_{2,0}\right)=0$ so that the definitional restrictions $\rho_1\left(\overline{a}_2,z,v=v_0\right)=0$ and $\rho_2\left(\overline{a}_2,l_1,l_2=l_{2,0}\right)=0$ are respected. Note that models (1.5) include polynomials in the variables intervening in ρ_1 and ρ_2 with unknown coefficients. For instance, the polynomial model $\gamma_{11}v+\gamma_{12}v^2+\gamma_{13}zv+\gamma_{14}zv^2+\gamma_{15}a_1v+\gamma_{16}a_2v$ for $\rho_1\left(\overline{a}_2,l_1\right)$ is obtained by taking $t\left(v\right)=\left(v,v^2,v,v^2,v,v\right)'$ and $g_1\left(\overline{a}_2,z;\gamma_1\right)=\left(\gamma_{11},\gamma_{12},\gamma_{13}z,\gamma_{14}z,\gamma_{15}a_1,\gamma_{16}a_2\right)'$.

Under models (1.5), in order to specify models for $E\left\{\rho_1\left(\overline{a}_2,L_1\right)|Z\right\}$ and $E\left\{\rho_2\left(\overline{a}_2,\overline{L}_2\right)|A_1,L_1\right\}$, it suffices to specify parametric conditional mean models

$$E\left\{t_1(V)|Z\right\}=e_1(Z;\tau_1^*), \tag{1.6}$$

$$E\left\{t_2(L_2)|A_1,L_1\right\}=e_2\left(A_1,L_1;\tau_2^*\right). \tag{1.7}$$

where e_1 and e_2 are user specified conformable vector-valued functions and τ_1^* and τ_2^* are finite dimensional parameters.

By (1.3) and (1.4), the models (1.2),(1.5), (1.6) and (1.7) imply the following compatible, shared parameter, models for η_1 and η_2,

$$\begin{aligned}\eta_1\left(\overline{a}_2,L_1\right)&=\eta_1\left(\overline{a}_2,L_1;\psi^*,\gamma_1^*,\tau_1^*\right)\\&\equiv m\left(\overline{a}_2,z;\psi^*\right)+g_1\left(\overline{a}_2,Z;\gamma_1^*\right)'\left\{t_1\left(V\right)-e_1(Z;\tau_1^*)\right\},\end{aligned} \tag{1.8}$$

$$\begin{aligned}\eta_2\left(\overline{a}_2,\overline{L}_2\right)&=\eta_2\left(\overline{a}_2,\overline{L}_2;\psi^*,\gamma_1^*,\gamma_2^*,\tau_1^*,\tau_2^*\right)\\&\equiv\eta_1\left(\overline{a}_2,L_1;\psi^*,\gamma_1^*,\tau_1^*\right)+g_2\left(\overline{a}_2,L_1;\gamma_2^*\right)'\left\{t_2\left(L_2\right)-e_2(a_1,L_1;\tau_2^*)\right\}.\end{aligned} \tag{1.9}$$

We say that (1.2), (1.8) and (1.9) are compatible models with shared parameters because, for every combination of values for $\psi^*,\gamma_1^*,\gamma_2^*,\tau_1^*$ and τ_2^*, there exists at least one distribution such that $\eta_0\left(\overline{a}_2,Z\right)=m\left(\overline{a}_2,Z;\psi^*\right)$, $\eta_1\left(\overline{a}_2,L_1\right)=\eta_1\left(\overline{a}_2,L_1;\psi^*,\gamma_1^*,\tau_1^*\right)$ and $\eta_2\left(\overline{a}_2,\overline{L}_2\right)=\eta_2\left(\overline{a}_2,\overline{L}_2;\psi^*,\gamma_1^*,\gamma_2^*,\tau_1^*,\tau_2^*\right)$. This follows from facts (1.3), (1.4) and from the fact that η_0, ρ_1, ρ_2, $f\left(V|Z\right)$ and $f\left(L_2|L_1,A_1\right)$ are variation independent.

1.6 Estimation exploiting the compatible models for the special case $K = 2$

1.6.1 Preliminaries

In this section we describe, in sequence, three estimators of ψ^*, each conferring more protection against model misspecification than the previous one.

Our estimators of ψ^* rely on preliminary method of moments fits of models (1.6) and (1.7) to estimate $\tau^* \equiv (\tau_1^*, \tau_2^*)$. Although τ^* also indexes the model $\eta_2(\overline{A}_2, \overline{L}_2; \psi^*, \gamma_1^*, \gamma_2^*, \tau^*)$ for $E\left(Y|\overline{A}_2, \overline{L}_2\right)$, we ignore this model for estimating τ^* because it carries little or no information about it when $(\psi^*, \gamma_1^*, \gamma_2^*)$ is unknown. For instance, consider the case in which A_j and L_j are binary, $j = 1, 2$, Z is nill and all working models and the MSMM are saturated. In such case, $\dim(\tau_1) = 1, \dim(\tau_2) = \dim(\gamma_1) = \dim(\psi) = 4$ and $\dim(\gamma_2) = 8$, so $\dim(\theta) = 21$ where $\theta = (\psi, \gamma_1, \gamma_2, \tau_1, \tau_2)$. However, there are only 16 means $E\left(Y|\overline{A}_2 = \overline{a}_2, \overline{L}_2 = \overline{l}_2\right)$, so θ is not identified under a model that just assumes that $E\left(Y|\overline{A}_2, \overline{L}_2\right) = \eta_2\left(\overline{A}_2, \overline{L}_2; \theta^*\right)$.

1.6.2 A regression estimator

Here, we describe an estimator $\widehat{\psi}_R$ of ψ^*, throughout referred to as regression estimator (R) that, under regularity conditions, is CAN under a model $\mathcal{R}_2$ for η_2 defined by restrictions (1.6), (1.7) and (1.9). Note that model $\mathcal{R}_2$ determines the parametric model (1.8) for η_1 since (1.7) and (1.9) imply (1.8). This is seen by taking the conditional expectation given $(A_1 = a_1, L_1)$ on both sides of (1.9). Likewise, $\mathcal{R}_2$ implies the restriction (1.2) that defines the MSMM under the identifying assumptions.

As indicated earlier, we first compute the method of moment estimators $\widehat{\tau}_1$ and $\widehat{\tau}_2$ of τ_1^* and τ_2^* under models (1.6) and (1.7). For instance, $\widehat{\tau}_1$ solves

$$\mathbb{P}_n\left[q_1(Z)\left\{t_1(V) - e_1(Z; \tau_1)\right\}\right] = 0,$$

where $q_1(z)$ is a a user-specified conformable matrix-valued function and $\widehat{\tau}_2$ solves

$$\mathbb{P}_n\left[q_2(A_1, L_1)\left\{t_2(L_2) - e_2(A_1, L_1; \tau_2)\right\}\right] = 0,$$

where $q_2(a_1, l_1)$ is a user-specified conformable matrix-valued function. In each of the preceding equations, 0 is a vector of zeros of an appropriate dimension.

Next, we compute the least squares estimator $\left(\widehat{\psi}_R, \widehat{\gamma}_{1,R}, \widehat{\gamma}_{2,R}\right)$ from the fit of the model $\eta_2\left(\overline{A}_2, \overline{L}_2; \psi^*, \gamma_1^*, \gamma_2^*, \widehat{\tau}_1, \widehat{\tau}_2\right)$ for $E\left(Y|\overline{A}_2, \overline{L}_2\right)$ where $(\psi^*, \gamma_1^*, \gamma_2^*)$ is unknown and $(\widehat{\tau}_1, \widehat{\tau}_2)$ is regarded as known. That is, we compute $\left(\widehat{\psi}_R, \widehat{\gamma}_{1,R}, \widehat{\gamma}_{2,R}\right)$ the solution of

$$\mathbb{P}_n\left[\frac{\partial \eta_2\left(\overline{A}_2, \overline{L}_2; \psi, \gamma_1, \gamma_2, \widehat{\tau}_1, \widehat{\tau}_2\right)}{\partial(\psi, \gamma_1, \gamma_2)}' \left\{Y - \eta_2\left(\overline{A}_2, \overline{L}_2; \psi, \gamma_1, \gamma_2, \widehat{\tau}_1, \widehat{\tau}_2\right)\right\}\right] = 0.$$

1.6.3 A doubly robust estimator

Here we describe an estimator $\widehat{\psi}_{DR}$ which is doubly robust in the following sense. Let $\mathcal{P}_1$ be a parametric model $\pi_1(a_1, l_1; \alpha_1^*)$ for $\pi_1(a_1, l_1)$ and let $\mathcal{P}_2$ be a parametric model $\pi_2\left(\overline{a}_2, \overline{l}_2; \alpha_2^*\right)$

for $\pi_2\left(\overline{a}_2, \overline{l}_2\right)$. Also, let $\mathcal{M}$ be the model defined by restriction (1.2), i.e. the MSMM under the identifying assumptions, for the case in which $K = 2$. Then, under regularity conditions, $\widehat{\psi}_{DR}$ is CAN under the union model that assumes that either (i) model $\mathcal{R}_2$ holds or (ii) models $\mathcal{M}$, $\mathcal{P}_1$ and $\mathcal{P}_2$ hold, but not necessarily both (i) and (ii) hold.

For any $\eta \equiv (\eta_1, \eta_2)$ and $\pi \equiv (\pi_1, \pi_2)$, not just the true ones, and any function d of $\left(\overline{A}_2, Z\right)$ define the estimating function

$$U_d(\psi, \eta, \pi) \equiv S_d^2(\eta_2, \pi_1, \pi_2) + S_d^1(\eta_1, \eta_2, \pi_1) + S_d^0(\psi, \eta_1) \tag{1.10}$$

where

$$\begin{aligned}
S_d^2(\eta_2, \pi_1, \pi_2) &\equiv \frac{d\left(\overline{A}_2, Z\right)}{\pi_1(A_1, L_1)\,\pi_2\left(\overline{A}_2, \overline{L}_2\right)}\left\{Y - \eta_2\left(\overline{A}_2, \overline{L}_2\right)\right\}, \\
S_d^1(\eta_1, \eta_2, \pi_1) &\equiv \sum_{a_2 \in \mathcal{A}_2} \frac{d(A_1, a_2, Z)}{\pi_1(A_1, L_1)}\left\{\eta_2\left(A_1, a_2, \overline{L}_2\right) - \eta_1(A_1, a_2, L_1)\right\}, \\
S_d^0(\psi, \eta_1) &\equiv \sum_{a_1 \in \mathcal{A}_1} \sum_{a_2 \in \mathcal{A}_2} d(\overline{a}_2, Z)\left\{\eta_1(\overline{a}_2, L_1) - m(\overline{a}_2, Z; \psi)\right\}.
\end{aligned}$$

To calculate $\widehat{\psi}_{DR}$, we first run the procedure in Section 1.6.2 and define

$$\widehat{\eta}_{2,R}\left(\overline{a}_2, \overline{l}_2\right) \equiv \eta_2\left(\overline{a}_2, \overline{l}_2; \widehat{\psi}_R, \widehat{\gamma}_{1,R}, \widehat{\gamma}_{2,R}, \widehat{\tau}_1, \widehat{\tau}_2\right)$$

and

$$\widehat{\eta}_{1,R}(\overline{a}_2, l_1) \equiv \eta_1\left(\overline{a}_2, l_1; \widehat{\psi}_R, \widehat{\gamma}_{1,R}, \widehat{\tau}_1\right).$$

Second, we compute $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ the MLEs of α_1^* and α_2^* under $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively and define

$$\widehat{\pi}_1(a_1, l_1) \equiv \pi_1(a_1, l_1; \widehat{\alpha}_1)$$

and

$$\widehat{\pi}_2\left(\overline{a}_2, \overline{l}_2\right) \equiv \pi_2\left(\overline{a}_2, \overline{l}_2; \widehat{\alpha}_2\right).$$

Finally, the estimator $\widehat{\psi}_{DR}$ solves

$$\mathbb{P}_n\left\{U_{\widehat{d}}(\psi, \widehat{\eta}_R, \widehat{\pi})\right\} = 0$$

where $\widehat{\eta}_R = (\widehat{\eta}_{1,R}, \widehat{\eta}_{2,R})$, $\widehat{\pi} = (\widehat{\pi}_1, \widehat{\pi}_2)$ and $\widehat{d}\left(\overline{A}_2, Z\right)$ is any, possibly data dependent, function of the same dimension as ψ^*, for instance, $\widehat{d}\left(\overline{A}_2, Z\right) = \left\{\frac{\partial m\left(\overline{A}_2, Z; \psi\right)}{\partial \psi}\right\}\Big|_{\psi = \widehat{\psi}_R}$. The estimator $\widehat{\psi}_{DR}$ is doubly robust essentially because

(I) as shown in [1], under $\mathcal{M}$,

$$E\left\{U_d(\psi^*, \eta', \pi')\right\} = 0 \tag{1.11}$$

if either (η_1', η_2') is equal to the true (η_1, η_2) or (π_1', π_2') is equal to the true (π_1, π_2),

(II) by construction, $\widehat{\eta}_{k,R}$ converges to the true $\eta_k, k = 1, 2$, under $\mathcal{R}_2$, and

(III) $\widehat{\pi}_k$ converges to the true π_k under $\mathcal{P}_k, k = 1, 2$.

Here and throughout, for any functional χ of the observed data law, any parametric model $\mathcal{G}$ assuming $\chi = \chi_{\beta^*}$ for some $\beta^* \in \Gamma$ (with Γ in a Euclidean space) and any estimator $\widehat{\beta}$ of β^*, we say that $\widehat{\chi} \equiv \chi_{\widehat{\beta}}$ "converges to" (or equivalently "is consistent for") χ under $\mathcal{G}$ if $\widehat{\beta}$ is consistent for β^* under $\mathcal{G}$.

1.6.4 A multiply robust estimator

Here, we propose an estimator $\widehat{\psi}_{MR}$ that confers even more protection against model misspecification than $\widehat{\psi}_{DR}$. Specifically, let $\mathcal{R}_1$ be the model defined by restrictions (1.6) and (1.8) . Note that $\mathcal{R}_2$ implies $\mathcal{R}_1$ and $\mathcal{R}_1$ implies $\mathcal{M}$. The estimator $\widehat{\psi}_{MR}$ is multiply robust in the sense that, under regularity conditions, it is CAN under the union model that assumes that any, but not necessarily all, of the following conditions (i), (ii) or (iii) is satisfied: (i) model $\mathcal{R}_2$ holds; (ii) models $\mathcal{R}_1$ and $\mathcal{P}_2$ hold; (iii) models $\mathcal{M}$, $\mathcal{P}_1$, and $\mathcal{P}_2$ hold. Thus, whereas $\widehat{\psi}_{DR}$ yields valid inferences if (i) or (iii) holds, $\widehat{\psi}_{MR}$ also does it if (ii) holds even when (i) and (iii) fail. The following table summarizes the definition of the models introduced in subsections 1.6.2 to 1.6.4.

Table 1

Definition of models

Model	Restrictions defining the model
$\mathcal{M}$	(1.2) for $K = 2$
$\mathcal{R}_1$	(1.6) , (1.8)
$\mathcal{R}_2$	(1.6) , (1.7) , (1.9)
$\mathcal{P}_1$	parametric model for $\pi_1(a_1, l_1)$
$\mathcal{P}_2$	parametric model for $\pi_2(\bar{a}_2, \bar{l}_2)$

The steps to construct $\widehat{\psi}_{MR}$ are:

1. Compute $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ the MLEs of α_1^* and α_2^* under $\mathcal{P}_1$ and $\mathcal{P}_2$ respectively and define $\widehat{\pi}_1(a_1, l_1) \equiv \pi_1(a_1, l_1; \widehat{\alpha}_1)$ and $\widehat{\pi}_2(\bar{a}_2, \bar{l}_2) \equiv \pi_2(\bar{a}_2, \bar{l}_2; \widehat{\alpha}_2)$.

2. Compute the estimators $\widehat{\tau}_1$ and $\widehat{\tau}_2$ of τ_1^* and τ_2^* as in Section 1.6.2.

3. Define $\eta_2(\bar{a}_2, \bar{l}_2; \psi, \gamma_1, \gamma_2) \equiv \eta_2(\bar{a}_2, \bar{l}_2; \psi, \gamma_1, \gamma_2, \widehat{\tau}_1, \widehat{\tau}_2)$ and $\overset{\bullet}{\eta}_2(\bar{a}_2, \bar{l}_2; \psi, \gamma_1, \gamma_2) \equiv \partial\eta_2(\bar{a}_2, \bar{l}_2; \psi, \gamma_1, \gamma_2)/\partial(\psi, \gamma_1, \gamma_2)$. Define $\left(\widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)}, \widehat{\gamma}_2^{(2)}\right)$ as a solution of

$$\mathbb{P}_n\left[\frac{\overset{\bullet}{\eta}_2(\overline{A}_2, \overline{L}_2; \psi, \gamma_1, \gamma_2)}{\widehat{\pi}_1(A_1, L_1)\,\widehat{\pi}_2(\overline{A}_2, \overline{L}_2)}\left\{Y - \eta_2(\overline{A}_2, \overline{L}_2; \psi, \gamma_1, \gamma_2)\right\}\right] = 0$$

if that equation has at least one solution and as an arbitrary value, in the same space where the parameters lie, otherwise.

Define $\widehat{\eta}_2(\bar{a}_2, \bar{l}_2) \equiv \eta_2\left(\bar{a}_2, \bar{l}_2; \widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)}, \widehat{\gamma}_2^{(2)}\right)$.

4. Define $\eta_1(\bar{a}_2, l_1; \psi, \gamma_1) \equiv \eta_1(\bar{a}_2, l_1; \psi, \gamma_1, \widehat{\tau}_1)$ and $\overset{\bullet}{\eta}_1(\bar{a}_2, l_1; \psi, \gamma_1) \equiv \partial\eta_1(\bar{a}_2, l_1; \psi, \gamma_1)/\partial(\psi, \gamma_1)$. Define $\left(\widehat{\psi}^{(1)}, \widehat{\gamma}_1^{(1)}\right)$ as a solution of

$$\mathbb{P}_n\left[\sum_{a_2 \in \mathcal{A}_2}\frac{\overset{\bullet}{\eta}_1\left(A_1, a_2, L_1; \widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)}\right)}{\widehat{\pi}_1(A_1, L_1)}\left\{\widehat{\eta}_2(A_1, a_2, \overline{L}_2) - \eta_1(A_1, a_2, L_1; \psi, \gamma_1)\right\}\right] = 0$$

if that equation has at least one solution and as an arbitrary value, in the same space where the parameters lie, otherwise.

Define $\widehat{\eta}_1(\bar{a}_2, l_1) \equiv \eta_1(\bar{a}_2, l_1; \widehat{\psi}^{(1)}, \widehat{\gamma}_1^{(1)})$.

5. Define $\overset{\bullet}{m}(\overline{a}_2, z; \psi) \equiv \partial m(\overline{a}_2, z; \psi)/\partial\psi$. Define $\widehat{\psi}_{MR}$ as a solution of

$$\mathbb{P}_n\left[\sum_{a_1\in\mathcal{A}_1}\sum_{a_2\in\mathcal{A}_2}\overset{\bullet}{m}\left(\overline{a}_2, Z; \widehat{\psi}^{(2)}\right)\left\{\widehat{\eta}_1(\overline{a}_2, L_1) - m(\overline{a}_2, Z; \psi)\right\}\right] = 0$$

if that equation has at least one solution and as an arbitrary value in R^p otherwise.

Note that the algorithm for computing $\widehat{\psi}_{MR}$ mimics the one for computing the ICE estimator $\widehat{\psi}_{ICE}$ except for the following modifications. First, the estimating functions in steps 3 and 4 are weighted by $(\widehat{\pi}_1\widehat{\pi}_2)^{-1}$ and $(\widehat{\pi}_1)^{-1}$ respectively. Second, the derivatives in the equations of steps 4 and 5 are evaluated at the estimators of ψ and γ_1 computed in step 3. These modifications are essential to ensure the multiple robustness of $\widehat{\psi}_{MR}$ (see Appendix A.2 and also the proof of fact (V) below). When the function g_1 of model (1.5) and m are linear in the parameters, $\overset{\bullet}{\eta}$ and $\overset{\bullet}{m}$ do not depend on ψ and γ_1. In such case, the equations in steps 4 and 5 can be implemented with standard weighted least squares software. Specifically, in step 4 one regards $\widehat{\eta}_2$ as an outcome that follows a conditional mean model $\eta_1\left(\overline{A}_2, L_1; \psi, \gamma_1\right)$. The solution of the equation in step 4 is a weighted least squares estimator of (ψ, γ_1) under such model based on a pseudo-sample in which each observation of the original sample is replicated as many times as the cardinal $\#\mathcal{A}_2$ of set $\mathcal{A}_2$, and A_2 is replaced in each replication by one distinct value of a_2 in $\mathcal{A}_2$. Likewise, in step 5 on regards $\widehat{\eta}_1$ as an outcome that follows a conditional mean model $m\left(\overline{A}_2, Z; \psi\right)$. The solution of the equation in step 5 is a weighted least squares estimator of ψ under such model based on a pseudo-sample in which each observation of the original sample is replicated as many times as $\#(\mathcal{A}_1 \times \mathcal{A}_2)$ and (A_1, A_2) is replaced in each replication by one distinct value of (a_1, a_2) in $\mathcal{A}_1 \times \mathcal{A}_2$.

Here, we provide an heuristic argument of why $\widehat{\psi}_{MR}$ should be consistent for ψ^* under a model that assumes that any, but not necessarily all, of the following conditions (i), (ii) or (iii) above is satisfied. Once consistency has been established, the convergence under regularity conditions of $\sqrt{n}\left(\widehat{\psi}_{MR} - \psi^*\right)$ to a mean zero Normal distribution follows immediately from the fact that $\widehat{\psi}_{MR}$ and all nuisance parameters ultimately solve a system of estimating equations. The precise regularity conditions under which $\widehat{\psi}_{MR}$ is consistent and asymptotically normal for ψ^* under the model that assumes that (i), (ii) or (iii) hold are given in Section 1.10 for the case of arbitrary K. Essentially, the multiple robustness of $\widehat{\psi}_{MR}$ is a consequence of the following facts:

(I) The identity (1.11) holds under $\mathcal{M}$ not only when (η_1', η_2') is equal to the true (η_1, η_2) or (π_1', π_2') is equal to the true (π_1, π_2), but also under the weaker condition that for each $j \in \{1, 2\}$, either η_j' is equal to the true η_j or π_j' is equal to the true π_j.

(II) Under regularity conditions, the estimator $\widehat{\pi}_j$ is consistent for π_j under $\mathcal{P}_j, j = 1, 2$.

(III) Under regularity conditions, the estimator $\widehat{\eta}_2$ in step 3 is consistent for the true η_2 under model $\mathcal{R}_2$.

(IV) Under regularity conditions, the estimator $\widehat{\eta}_1$ in step 4 is itself doubly robust in that it is consistent for the true η_1 under the model that assumes that $\mathcal{R}_1$ holds and that either $\mathcal{R}_2$ or $\mathcal{P}_2$ holds.

(V) The estimator $\widehat{\psi}_{MR}$ in step 5 solves $\mathbb{P}_n\left\{U_{\widehat{d}}(\psi, \widehat{\eta}, \widehat{\pi})\right\} = 0$ for $\widehat{d}\left(\overline{A}_2, Z\right) = \overset{\bullet}{m}\left(\overline{A}_2, Z; \widehat{\psi}^{(2)}\right)$, $\widehat{\eta} \equiv (\widehat{\eta}_1, \widehat{\eta}_2)$ computed in steps 3 and 4, and $\widehat{\pi} \equiv (\widehat{\pi}_1, \widehat{\pi}_2)$ computed in step 1.

Facts (I)-(V) imply that, under regularity conditions, $\widehat{\psi}_{MR}$ is CAN under the model that assumes that $\mathcal{M}$ holds and that for each $j \in \{1,2\}$ either model $\mathcal{R}_j$ or model $\mathcal{P}_j$ holds. The fact that $\mathcal{R}_2$ implies $\mathcal{R}_1$ and, $\mathcal{R}_1$ implies $\mathcal{M}$ then gives that $\widehat{\psi}_{MR}$ is CAN under the model that assumes that any, but not necessarily all, of the following conditions (i), (ii) or (iii) is satisfied: (i) model $\mathcal{R}_2$ holds; (ii) models $\mathcal{R}_1$ and $\mathcal{P}_2$ hold; (iii) models $\mathcal{M}, \mathcal{P}_1$, and $\mathcal{P}_2$ hold.

Tchetgen Tchetgen ([53]) first noted the remarkable fact (I). Later, Molina et al. ([24]) noted that this fact is a consequence of the likelihood factorization that takes place in coarsened at random models. For completeness, in Proposition 1 of Section 1.10 we give an independent proof of that fact for the case of arbitrary K.

Fact (III) is true because $\mathcal{R}_2$ is a regression model for the outcome Y on covariates $\overline{A}_2$ and $\overline{L}_2$, and in addition, under regularity conditions, $\widehat{\tau}_1$ and $\widehat{\tau}_2$ converge to τ_1^* and τ_2^* respectively under $\mathcal{R}_2$.

Fact (IV) is true because of the theoretical results in [24]. In Appendix A.2 we give an intuitive argument invoking counterfactuals.

Fact (V) is true because, when $\widehat{d}\left(\overline{A}_2, Z\right) = \overset{\bullet}{m}\left(\overline{A}_2, Z; \widehat{\psi}^{(2)}\right)$,

(i) $\widehat{\psi}_{MR}$ solves $\mathbb{P}_n\left\{S^0_{\widehat{d}}(\psi, \widehat{\eta}_1)\right\} = 0$ by step 5,

(ii) $\mathbb{P}_n\left\{S^1_{\widehat{d}}(\widehat{\eta}_1, \widehat{\eta}_2, \widehat{\pi}_1)\right\} = 0$, by step 4 and the fact that $\overset{\bullet}{m}\left(A_1, a_2, Z; \widehat{\psi}^{(2)}\right)$ is a subvector of $\overset{\bullet}{\eta}_1\left(A_1, a_2, L_1; \widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)}\right)$, and

(iii) $\mathbb{P}_n\left\{S^2_{\widehat{d}}(\widehat{\eta}_2, \widehat{\pi}_1, \widehat{\pi}_2)\right\} = 0$ by step 3 and the fact that $\overset{\bullet}{m}\left(\overline{A}_2, Z; \widehat{\psi}^{(2)}\right)$ is also a subvector of $\overset{\bullet}{\eta}_2\left(\overline{A}_2, \overline{L}_2; \widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)}, \widehat{\gamma}_2^{(2)}\right)$.

Note that evaluating $\overset{\bullet}{\eta}_1$ and $\overset{\bullet}{m}$ at $\psi = \widehat{\psi}^{(2)}$ in steps 4 and 5 is key to ensure that (ii) and (iii) hold. Likewise, as shown in Appendix A.2, evaluating $\overset{\bullet}{\eta}_1$ at $\gamma = \widehat{\gamma}_1^{(2)}$ in step 4 is key to ensure fact (IV). Therefore, evaluating $\overset{\bullet}{\eta}_1$ and $\overset{\bullet}{m}$ at $\widehat{\psi}^{(2)}$ and $\widehat{\gamma}_1^{(2)}$ in steps 4 and 5 is essential to guarantee the multiple robustness of $\widehat{\psi}_{MR}$.

1.6.5 On our proposed models for ρ_k

Our choice of model (1.5) for ρ_1 and ρ_2 is based on considerations of flexibility and ease of implementation. First, the formulation (1.5) is flexible because, as indicated earlier, it includes polynomial models of any order. Because any function of finite-valued variables can be expressed as a polynomial then, in particular, when all variables but Y are finite-valued, even the saturated models for ρ_1 and ρ_2 can be written as (1.5). Note also that if g_1 depends on all $\overline{a}_2$ and g_2 depends on a_2, then (1.5) imply models for η_1 and η_2 that allow the possibility that (1) L_1 is a modifier for the additive effect of $\overline{A}_2$ on Y, i. e., that the differences $E\left(Y_{\overline{a}_2}|L_1\right) - E\left(Y_{\overline{a}_2^*}|L_1\right)$ depend on L_1 for $\overline{a}_2 \neq \overline{a}_2^*$ and (2) A_1, L_1, L_2 are modifiers for the additive effect of A_2 on Y. Second, implementation is facilitated under models (1.5) and (1.6) – (1.7) because τ_k^* is estimated separately from γ_k^*. In principle, it is possible to implement R, DR and MR estimators under arbitrary models $\rho_1\left(\overline{a}_2, L_1; \gamma_1^*\right)$ and $\rho_2\left(\overline{a}_2, \overline{L}_2; \gamma_2^*\right)$, and models $e_1\left(Z; \tau_1^*\right)$ for $E\left\{\rho_1\left(\overline{a}_2, L_1; \gamma_1^*\right)|Z\right\}$ and $e_2\left(A_1, L_1; \tau_2^*\right)$ for $E\left\{\rho_2\left(A_1, a_2, \overline{L}_2; \gamma_2^*\right)|A_1, L_1\right\}$. Such implementation would follow the algorithms described in the preceding sections except that one will need to first compute method of moments estimates $\widehat{\tau}_1\left(\gamma_1\right)$ and $\widehat{\tau}_2\left(\gamma_2\right)$ over a fine grid of (γ_1, γ_2) values and then follow the R, DR and MR algorithms

with $\widehat{\tau}_1(\gamma_1)$ and $\widehat{\tau}_2(\gamma_2)$ instead of $\widehat{\tau}_1$ and $\widehat{\tau}_2$. While feasible, this strategy would be computationally intense.

One will typically not be free to choose arbitrary models for each of the entries of the vectors $E\{t_1(V)|Z\}$ and $E\{t_2(L_2)|A_1, L_1\}$. For instance, if $t_1(V) = \left(V, V^2\right)$, then models for $E\{t_1(V)|Z\} = E\left\{\left(V, V^2\right)|Z\right\}$ will be tied by the inequality $E\left(V^2|Z\right) \geq E(V|Z)^2$. One strategy to come up with models that do not violate any logical constraint is to derive them from fully parametric models for the densities $f(v|z)$ and $f(l_2|l_1, a_1)$. We emphasize that under this strategy the conditions for consistency of the R, DR and MR estimators will remain dependent on the validity of just the mean models (1.6) and (1.7) implied by the fully parametric models, and not on the full validity of the latter. This is because τ_1^* and τ_2^* are not estimated by maximum likelihood but rather by the method of moments.

1.7 MR estimation for arbitrary K

In this section, we generalize the construction of compatible models for the functionals η_k proposed in Section 1.5, to the case in which there are $K > 2$ time points. We also generalize the R, DR and MR estimation algorithms of Section 1.6.

1.7.1 Compatible parametric working models for the η_k

The derivation of compatible models for the $\eta_k's$ with shared parameters for the case of $K > 2$ is completely analogous to the one proposed in Section 1.5 for the case of $K = 2$. As indicated in that section, Robins et al. [42] proposed a parameterization of the likelihood that depends on $f(V|Z)$, $f\left(L_k|\overline{L}_{k-1}, \overline{A}_{k-1}\right)$, $k = 2, \ldots, K$, $\eta_0(\overline{a}_2, z)$,

$$\rho_1(\overline{a}_K, l_1) \equiv \eta_1(\overline{a}_K, l_1) - \eta_1(\overline{a}_K, z, v = v_0)$$

and

$$\rho_k\left(\overline{a}_K, \overline{l}_k\right) \equiv \eta_k\left(\overline{a}_K, \overline{l}_k\right) - \eta_k\left(\overline{a}_K, \overline{l}_{k-1}, l_k = l_{k,0}\right),$$

$k = 2, \ldots, K$, where v_0 and $l_{k,0}$ are any baseline levels of V and L_k, $k = 2, \ldots, K$, respectively and $l_1 \equiv (z, v)$. These functionals are variation independent, as shown in Appendix A.1 for the case in which $K = 2$. The proof for arbitrary K is completely analogous. As in the case of $K = 2$, straightforward calculations yield

$$\eta_1(\overline{a}_K, l_1) = \eta_0(\overline{a}_K, z) + \rho_1(\overline{a}_K, l_1) - E\{\rho_1(\overline{a}_K, L_1)|Z = z\}, \tag{1.12}$$

and

$$\begin{aligned}\eta_k\left(\overline{a}_K, \overline{l}_k\right) = {} & \eta_0(\overline{a}_K, z) + \rho_1(\overline{a}_K, l_1) - E\{\rho_1(\overline{a}_K, L_1)|Z = z\} \\ & + \textstyle\sum_{j=2}^{k}\left[\rho_j\left(\overline{a}_K, \overline{l}_j\right) - E\left\{\rho_j\left(\overline{a}_K, \overline{L}_j\right)|\overline{A}_{j-1} = \overline{a}_{j-1}, \overline{L}_{j-1} = \overline{l}_{j-1}\right\}\right],\end{aligned} \tag{1.13}$$

$k = 2, \ldots, K$. These identities imply that parametric models for the $\rho_j's$, for $E\{\rho_1(\overline{a}_K, L_1)|Z\}$ and $E\left\{\rho_j\left(\overline{a}_K, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}, j = 2, \ldots, K$, and the MSMM $\eta_0(\overline{a}_K, z) = m(\overline{a}_K, z; \psi^*)$ determine compatible parametric, shared parameter, models for the $\eta_k, k = 0, \ldots, K$. For reasons analogous to those explained in Section 1.6.5, we propose modeling the $\rho_j's$ as

$$\rho_1(\overline{a}_K, L_1) = g_1(\overline{a}_K, Z; \gamma_1^*)' t_1(V), \tag{1.14}$$

and

$$\rho_j\left(\overline{a}_K, \overline{L}_j\right) = g_j\left(\overline{a}_K, \overline{L}_{j-1}; \gamma_j^*\right)' t_j(L_j), \tag{1.15}$$

$j = 2, \ldots, K$, where the $g_j's$ are user specified vector-valued functions, the $\gamma_j^{*\prime}s$ are finite dimensional parameters, and the $t_j's$ are user-specified conformable vector-valued functions verifying $t_1(v_0) = 0$ and $t_j(l_{j,0}) = 0, j = 2, \ldots, K$, so that the definitional restrictions $\rho_1(\overline{a}_K, z, v = v_0) = 0$ and $\rho_j\left(\overline{a}_K, \overline{l}_{j-1}, l_{j,0}\right) = 0$, $j = 2, \ldots, K$, are respected. Note that, as in the case of $K = 2$, models (1.14) and (1.15) include polynomials in the variables intervening in the $\rho_j's$ with unknown coefficients.

Under models (1.14) and (1.15), in order to specify models for $E\{\rho_1(\overline{a}_K, L_1)|Z\}$ and $E\left\{\rho_j\left(\overline{a}_K, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$ it suffices to specify parametric conditional mean models

$$E\{t_1(V)|Z\} = e_1(Z; \tau_1^*), \tag{1.16}$$

and, for $j = 2, \ldots, K$,

$$E\left\{t_j(L_j)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\} = e_j\left(\overline{A}_{j-1}, \overline{L}_{j-1}; \tau_j^*\right). \tag{1.17}$$

where the $e_{'j}s$ are user specified conformable vector-valued functions and the $\tau_j^{*\prime}s$ are finite dimensional parameters.

By (1.12) and (1.13), the MSMM (1.2) and the models (1.14), (1.15), (1.16) and (1.17) imply the following compatible, shared parameter, models for the $\eta_j'^s$,

$$\begin{aligned}\eta_1\left(\overline{a}_K, L_1\right) &= \eta_1\left(\overline{a}_K, L_1; \psi^*, \gamma_1^*, \tau_1^*\right) \\ &\equiv m\left(\overline{a}_K, Z; \psi^*\right) + g_1\left(\overline{a}_K, Z; \gamma_1^*\right)'\left\{t_1\left(V\right) - e_1\left(Z; \tau_1^*\right)\right\},\end{aligned} \tag{1.18}$$

and

$$\begin{aligned}\eta_j\left(\overline{a}_K, \overline{L}_j\right) &= \eta_j\left(\overline{a}_K, \overline{L}_j; \psi^*, \overline{\gamma}_j^*, \overline{\tau}_j^*\right) \\ &\equiv \eta_{j-1}\left(\overline{a}_K, \overline{L}_{j-1}; \psi^*, \overline{\gamma}_{j-1}^*, \overline{\tau}_{j-1}^*\right) + g_j\left(\overline{a}_K, \overline{L}_{j-1}; \gamma_j^*\right)'\left\{t_j\left(L_j\right) - e_j(\overline{a}_{j-1}, \overline{L}_{j-1}; \tau_j^*)\right\},\end{aligned} \tag{1.19}$$

$j = 2, \ldots, K$. We say that (1.2), (1.18) and (1.19) for $j = 2, ..., K$, are compatible models with shared parameters because, for every combination of values for $\psi^*, \overline{\gamma}_K^*$ and $\overline{\tau}_K^*$, there exists at least one distribution such that $\eta_0\left(\overline{a}_K, Z\right) = m\left(\overline{a}_K, Z; \psi^*\right)$ and $\eta_k\left(\overline{a}_K, \overline{L}_k\right) = \eta_k\left(\overline{a}_K, \overline{L}_k; \psi^*, \overline{\gamma}_k^*, \overline{\tau}_k^*\right), k = 1, \ldots, K$. This follows from (1.12), (1.13) and from the fact that η_0,the $\rho_j'^s, f\left(V|Z\right)$ and $f\left(L_j|\overline{L}_{j-1}, \overline{A}_{j-1}\right), j = 2, \ldots, K$ are variation independent functionals.

1.7.2 Estimation exploiting the compatible models

In this section, we extend the three estimators of Section 1.6, to the case of arbitrary K exposure time points. Again, in the construction of our estimators, we will separately estimate $\overline{\tau}_K^*$ by the method of moments. We will not exploit model $\eta_K\left(\overline{A}_K, \overline{L}_K; \psi^*, \overline{\gamma}_K^*, \overline{\tau}_K^*\right)$ for $E\left(Y|\overline{A}_K, \overline{L}_K\right)$ to estimate $\overline{\tau}_K^*$ because this model carries little or no information about this parameter when the remaining parameters are unknown.

A regression estimator

To calculate the regression estimator $\widehat{\psi}_R$, as indicated earlier, we first compute $\widehat{\overline{\tau}}_K \equiv (\widehat{\tau}_1, \ldots, \widehat{\tau}_K)$ where $\widehat{\tau}_j$ is a method of moment estimator of τ_j^*, under (1.16) if $j = 1$, and (1.17) if $j > 1$. Next, we compute the least squares estimator $\left(\widehat{\psi}_R, \widehat{\overline{\gamma}}_{K,R}\right) \equiv \left(\widehat{\psi}_R, \widehat{\gamma}_{1,R}, \ldots, \widehat{\gamma}_{K,R}\right)$ from the fit of the model $\eta_K\left(\overline{A}_K, \overline{L}_K; \psi^*, \overline{\gamma}_K^*, \widehat{\overline{\tau}}_K\right)$ for $E\left(Y|\overline{A}_K, \overline{L}_K\right)$ where $(\psi^*, \overline{\gamma}_K^*)$ is unknown and $\widehat{\overline{\tau}}_K$ is regarded as known. Under regularity conditions, the estimator $\widehat{\psi}_R$ is CAN under the model $\mathcal{R}_K$ defined by restriction (1.16), restriction (1.17) for $j = 2, \ldots, K$, and restriction (1.19) for $j = K$.

Note that model $\mathcal{R}_K$ determines the parametric model for η_{K-1} defined by equation (1.19) for $j = K - 1$. This is because equations (1.17) for $j = K$ and (1.19) for $j = K$ imply equation (1.19) for $j = K - 1$, as is seen by computing the conditional expectation given $\left(\overline{A}_{K-1} = \overline{a}_{K-1}, \overline{L}_{K-1}\right)$ on both sides of (1.19) for $j = K$. Likewise, $\mathcal{R}_K$ implies the models for the $\eta_j'^s$ defined by restrictions (1.18) if $j = 1$ and (1.19) if $j = 2, \ldots, K - 2$, and also implies the restriction (1.2) that defines the MSMM under the indentifying assumptions.

A doubly robust estimator

For each $k = 1, \ldots, K$, let $\mathcal{P}_k$ be a parametric model $\pi_k\left(\overline{a}_k, \overline{l}_k; \alpha_k^*\right)$ for $\pi_k\left(\overline{a}_k, \overline{l}_k\right)$. Also, let $\mathcal{M}$ be the MSMM under the identified assumptions, defined by restriction (1.2). Our proposed estimator $\widehat{\psi}_{DR}$ is doubly robust in the sense that, under regularity conditions, it is CAN under the union model that assumes that either (i) $\mathcal{R}_K$ holds or (ii) model $\mathcal{M}$ and models $\mathcal{P}_k, k = 1, \ldots, , K$, hold, but not necessarily both (i) and (ii) hold.

For any $\eta \equiv (\eta_1, \ldots, \eta_K)$ and $\pi \equiv (\pi_1, \ldots, \pi_K)$, not just the true ones, and any function d of $\left(\overline{A}_K, Z\right)$, define the estimating function

$$U_d(\psi, \eta, \pi) \equiv S_d^K(\eta_K, \pi) + \sum_{k=1}^{K-1} S_d^k(\eta_k, \eta_{k+1}, \pi_1, \ldots, \pi_k) + S_d^0(\psi, \eta_1), \tag{1.20}$$

where

$$S_d^K(\eta_K, \pi) \equiv \frac{d\left(\overline{A}_K, Z\right)}{\prod_{j=1}^K \pi_j\left(\overline{A}_j, \overline{L}_j\right)} \left\{Y - \eta_K\left(\overline{A}_K, \overline{L}_K\right)\right\},$$

for $k = 1, \ldots, K-1$,

$$S_d^k(\eta_k, \eta_{k+1}, \pi_1, \ldots, \pi_k) \equiv \sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{d(\overline{A}_k, \underline{a}_{k+1}, Z)}{\prod_{j=1}^k \pi_j\left(\overline{A}_j, \overline{L}_j\right)} \left\{\eta_{k+1}(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1}) - \eta_k(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k)\right\},$$

and

$$S_d^0(\psi, \eta_1) \equiv \sum_{\underline{a}_1 \in \underline{\mathcal{A}}_1} d(\underline{a}_1, Z) \left\{\eta_1(\underline{a}_1, L_1) - m(\underline{a}_1, Z; \psi)\right\},$$

where recall that $\underline{\mathcal{A}}_k \equiv \mathcal{A}_k \times \cdots \times \mathcal{A}_K$, $k = 1, \ldots, K$.

To compute $\widehat{\psi}_{DR}$ we first run the preceding procedure in the regression estimator algorithm and, for each $k = 1, \ldots, K$, we define

$$\widehat{\eta}_{k,R}\left(\overline{a}_K, \overline{l}_k\right) \equiv \eta_k(\overline{a}_K, \overline{l}_k; \widehat{\psi}_R, \widehat{\overline{\gamma}}_{k,R}, \widehat{\overline{\tau}}_k).$$

Second, for each $k = 1, \ldots, K$, we compute $\widehat{\alpha}_k$ the MLE of α_k^* under $\mathcal{P}_k$ and define

$$\widehat{\pi}_k\left(\overline{a}_k, \overline{l}_k\right) \equiv \pi_k\left(\overline{a}_k, \overline{l}_k; \widehat{\alpha}_k\right).$$

Finally, the estimator $\widehat{\psi}_{DR}$ solves

$$\mathbb{P}_n\left\{U_{\widehat{d}}(\psi, \widehat{\eta}_R, \widehat{\pi})\right\} = 0$$

where $\widehat{\eta}_R = (\widehat{\eta}_{1,R}, \ldots, \widehat{\eta}_{K,R}), \widehat{\pi} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_K)$ and $\widehat{d}\left(\overline{A}_K, Z\right)$ is any, possibly data dependent, function of the same dimension as ψ^*, for instance, $\widehat{d}\left(\overline{A}_K, Z\right) = \left\{\partial m\left(\overline{A}_K, Z; \psi\right)/\partial\psi\right\}\big|_{\psi=\widehat{\psi}_R}$. The estimator $\widehat{\psi}_{DR}$ is doubly robust essentially because

(I) as shown in [1], under $\mathcal{M}$,

$$E\left\{U_d(\psi^*, \eta', \pi')\right\} = 0, \tag{1.21}$$

if either η' is equal to the true η or π' is equal to the true π,

(II) by construction, $\widehat{\eta}_{k,R}$ converges to the true η_k under $\mathcal{R}_K, k = 1, \ldots, K$, and

(III) $\widehat{\pi}_k$ converges to the true π_k under $\mathcal{P}_k$, $k = 1, \ldots, K$.

A multiply robust estimator

Here, we propose an estimator $\widehat{\psi}_{MR}$ that confers even more protection against model misspecification than $\widehat{\psi}_{DR}$. Specifically, let $\mathcal{R}_1$ be the model defined by restrictions (1.16) and (1.18) . Also, for each $k \in \{2, \ldots, K-1\}$, let $\mathcal{R}_k$ be the model defined by restriction (1.16) , restrictions (1.17) for $j = 2, \ldots, k$, and restriction (1.19) for $j = k$. That is, $\mathcal{R}_1$ is the model defined by restrictions

$$\begin{aligned} E\{t_1(V)|Z\} &= e_1(Z;\tau_1^*) \text{ and} \\ \eta_1(\overline{a}_K, L_1) &= \eta_1(\overline{a}_K, L_1; \psi^*, \gamma_1^*, \tau_1^*) \\ &\equiv m(\overline{a}_K, Z; \psi^*) + g_1(\overline{a}_K, Z; \gamma_1^*)' \{t_1(V) - e_1(Z; \tau_1^*)\} \end{aligned}$$

and, for each $k = 2, \ldots, K-1$, $\mathcal{R}_k$ is the model defined by restrictions

$$\begin{aligned} E\{t_1(V)|Z\} &= e_1(Z;\tau_1^*), \\ E\left\{t_j(L_j)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\} &= e_j\left(\overline{A}_{j-1}, \overline{L}_{j-1}; \tau_j^*\right), j = 2, \ldots, k, \text{ and} \\ \eta_k\left(\overline{a}_K, \overline{L}_k\right) &= \eta_k\left(\overline{a}_K, \overline{L}_k; \psi^*, \overline{\gamma}_k^*, \overline{\tau}_k^*\right) \\ &\equiv \eta_{k-1}\left(\overline{a}_K, \overline{L}_{k-1}; \psi^*, \overline{\gamma}_{k-1}^*, \overline{\tau}_{k-1}^*\right) \\ &+ g_k\left(\overline{a}_K, \overline{L}_{k-1}; \gamma_k^*\right)' \left\{t_k(L_k) - e_k(\overline{a}_{k-1}, \overline{L}_{k-1}; \tau_k^*)\right\}. \end{aligned}$$

The following table summarizes the definition of the models introduced in subsections 1.7.2 to 1.7.2.

Table 2

Definition of models

Model	Restrictions defining the model
$\mathcal{M}$	(1.2)
$\mathcal{R}_1$	(1.16) , (1.18)
for $k = 2, \ldots, K$	
$\mathcal{R}_k$	(1.16) , (1.17) , $j = 2, \ldots, k$, (1.19) , $j = k$
for $k = 1, \ldots, K$	
$\mathcal{P}_k$	parametric model for $\pi_k\left(\overline{a}_k, \overline{l}_k\right)$

Note that, for each $k = 1, \ldots, K-1$, $\mathcal{R}_{k+1}$ implies $\mathcal{R}_k$. Likewise, $\mathcal{R}_1$ implies the MSMM $\mathcal{M}$. The estimator $\widehat{\psi}_{MR}$ is multiply robust in the sense that, under regularity conditions, it is CAN under the model that assumes that any, but not necessarily all, of the following conditions (i), (ii) or (iii) is satisfied: (i) $\mathcal{R}_K$ holds; (ii) for some $k \in \{1, \ldots, K-1\}$, model $\mathcal{R}_k$ and models $\mathcal{P}_{k+1}, \ldots, \mathcal{P}_K$ hold; (iii) model $\mathcal{M}$ and models $\mathcal{P}_1, \ldots, \mathcal{P}_K$ hold. Thus, whereas $\widehat{\psi}_{DR}$ yields valid inferences if (i) or (iii) holds, $\widehat{\psi}_{MR}$ also does it if (ii) holds even when (i) and (iii) fail.

The following algorithm yields the proposed estimator.

Algorithm 1

1. For $k = 1, \ldots, K$ we compute $\widehat{\alpha}_k$ the MLE of α_k^* under $\mathcal{P}_k$, which we will assume it solves

$$\mathbb{P}_n\left[\frac{\partial}{\partial \alpha_k} \ln \pi_k\left(\overline{A}_k, \overline{L}_k; \alpha_k\right)\right] = 0.$$

Define $\widehat{\pi}_k\left(\overline{a}_k, \overline{l}_k\right) \equiv \pi_k\left(\overline{a}_k, \overline{l}_k; \widehat{\alpha}_k\right)$.

2. Compute a method of moments estimator $\widehat{\tau}_1$ of τ_1^* that solves the moment equations

$$\mathbb{P}_n \left[q_1 \left(Z \right) \left\{ t_1 \left(V \right) - e_1 \left(Z; \tau_1 \right) \right\} \right] = 0,$$

where $q_1 \left(z \right)$ is a a user-specified conformable matrix-valued function and, for $k = 2, \ldots, K$, compute a method of moments estimator $\widehat{\tau}_k$ of τ_k^* that solves the moment equations

$$\mathbb{P}_n \left[q_k \left(\overline{A}_{k-1}, \overline{L}_{k-1} \right) \left\{ t_k \left(L_k \right) - e_k \left(\overline{A}_{k-1}, \overline{L}_{k-1}; \tau_k \right) \right\} \right] = 0,$$

where $q_k \left(\overline{a}_{k-1}, \overline{l}_{k-1} \right)$ is a user-specified conformable matrix-valued function.

3. Define $\eta_K \left(\overline{a}_K, \overline{l}_K; \psi, \overline{\gamma}_K \right) \equiv \eta_K \left(\overline{a}_K, \overline{l}_K; \psi, \overline{\gamma}_K, \widehat{\overline{\tau}}_K \right)$ and
$\overset{\bullet}{\eta}_K \left(\overline{a}_K, \overline{l}_K; \psi, \overline{\gamma}_K \right) \equiv \partial \eta_K \left(\overline{a}_K, \overline{l}_K; \psi, \overline{\gamma}_K \right) / \partial \left(\psi, \overline{\gamma}_K \right)$.
Define $\left(\widehat{\psi}^{(K)}, \widehat{\overline{\gamma}}_K^{(K)} \right) \equiv \left(\widehat{\psi}^{(K)}, \widehat{\gamma}_1^{(K)}, \ldots, , \widehat{\gamma}_K^{(K)} \right)$ as a solution of

$$\mathbb{P}_n \left[\frac{\overset{\bullet}{\eta}_K \left(\overline{A}_K, \overline{L}_K; \psi, \overline{\gamma}_K \right)}{\prod_{j=1}^K \widehat{\pi}_j \left(\overline{A}_j, \overline{L}_j \right)} \left\{ Y - \eta_K \left(\overline{A}_K, \overline{L}_K; \psi, \overline{\gamma}_K \right) \right\} \right] = 0$$

if that equation has at least one solution and as an arbitrary value, in the same space where the parameters lie, otherwise.

Define $\widehat{\eta}_K \left(\overline{a}_K, \overline{l}_K \right) \equiv \eta_K \left(\overline{a}_K, \overline{l}_K; \widehat{\psi}^{(K)}, \widehat{\overline{\gamma}}_K^{(K)} \right)$.

4. For $k = K-1, \ldots, 1$, iteratively define $\eta_k \left(\overline{a}_K, \overline{l}_k; \psi, \overline{\gamma}_k \right) \equiv \eta_k \left(\overline{a}_K, \overline{l}_k; \psi, \overline{\gamma}_k, \widehat{\overline{\tau}}_k \right)$ and
$\overset{\bullet}{\eta}_k \left(\overline{a}_K, \overline{l}_k; \psi, \overline{\gamma}_k \right) \equiv \partial \eta_k \left(\overline{a}_K, \overline{l}_k; \psi, \overline{\gamma}_k \right) / \partial \left(\psi, \overline{\gamma}_k \right)$.
Define $\left(\widehat{\psi}^{(k)}, \widehat{\overline{\gamma}}_k^{(k)} \right) \equiv \left(\widehat{\psi}^{(k)}, \widehat{\gamma}_1^{(k)}, \ldots, \widehat{\gamma}_k^{(k)} \right)$ as a solution of

$$\mathbb{P}_n \left[\sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{\overset{\bullet}{\eta}_k \left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \widehat{\psi}^{(K)}, \widehat{\overline{\gamma}}_k^{(K)} \right)}{\prod_{j=1}^k \widehat{\pi}_j \left(\overline{A}_j, \overline{L}_j \right)} \left\{ \widehat{\eta}_{k+1} \left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1} \right) - \eta_k \left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \psi, \overline{\gamma}_k \right) \right\} \right] = 0$$

if that equation has at least one solution and as an arbitrary value, in the same space where the parameters lie, otherwise.

Define $\widehat{\eta}_k(\overline{a}_K, \overline{l}_k) \equiv \eta_k(\overline{a}_K, \overline{l}_k; \widehat{\psi}^{(k)}, \widehat{\overline{\gamma}}_k^{(k)})$.

5. Define $\overset{\bullet}{m} \left(\overline{a}_K, z; \psi \right) \equiv \partial m \left(\overline{a}_K, z; \psi \right) / \partial \psi$.
Define $\widehat{\psi}_{MR}$ as a solution of

$$\mathbb{P}_n \left[\sum_{\underline{a}_1 \in \underline{\mathcal{A}}_1} \overset{\bullet}{m} \left(\underline{a}_1, Z; \widehat{\psi}^{(K)} \right) \left\{ \widehat{\eta}_1(\underline{a}_1, L_1) - m \left(\underline{a}_1, Z; \psi \right) \right\} \right] = 0$$

if that equation has at least one solution and an as an arbitrary value in R^p otherwise.

As in the case in which $K = 2$, when $g_k, k = 1, \ldots, K-1$, and m are linear in the parameters, the derivatives $\overset{\bullet}{\eta}_k, k = 1, \ldots, K-1$, and $\overset{\bullet}{m}$ do not depend on ψ and $\overline{\gamma}_k$. In such case, the equations in steps 4 and 5 can be implemented with standard weighted least squares software.

The multiple robustness of $\widehat{\psi}_{MR}$ is a consequence of the following facts:

(I) The identity (1.21) holds not only when $(\eta'_1, \ldots, \eta'_K)$ is equal to the true $(\eta_1, \ldots, \eta_K)$ or $(\pi'_1, \ldots, \pi'_K)$ is equal to the true $(\pi_1, \ldots, \pi_K)$, but also under the weaker condition that, for each $k \in \{1, \ldots, K\}$, either η'_k is equal to the true η_k or π'_k is equal to the true π_k.

(II) Under regularity conditions, the estimator $\widehat{\pi}_k$ is consistent for π_k under $\mathcal{P}_k$, $k = 1, \ldots, K$.

(III) Under regularity conditions, the estimator $\widehat{\eta}_K$ in step 3 is consistent for η_K under model $\mathcal{R}_K$.

(IV) Under regularity conditions, for each $k = 1, \ldots, (K-1)$, the estimator $\widehat{\eta}_k$ in step 4 is itself multiply robust in that it is consistent for η_k under the model that assumes that $\mathcal{R}_k$ holds and that, for each $j \in \{k+1, \ldots, K\}$, either $\mathcal{R}_j$ or $\mathcal{P}_j$ holds.

(V) The estimator $\widehat{\psi}_{MR}$ in step 5 actually solves the equation $\mathbb{P}_n \left\{ U_{\widehat{d}}(\psi, \widehat{\eta}, \widehat{\pi}) \right\} = 0$ for $\widehat{d}\left(\overline{A}_K, Z\right) = \overset{\bullet}{m}\left(\overline{A}_K, Z; \widehat{\psi}^{(K)}\right)$, $\widehat{\eta} \equiv (\widehat{\eta}_1, \ldots, \widehat{\eta}_K)$ computed in steps 3 and 4 and $\widehat{\pi} \equiv (\widehat{\pi}_1, \ldots, \widehat{\pi}_K)$ computed in step 1.

Facts (I)-(V) imply that, under regularity conditions, $\widehat{\psi}_{MR}$ is CAN for ψ^* under the model that assume that $\mathcal{M}$ holds and that, for each $j \in \{1, \ldots, K\}$, either $\mathcal{R}_j$ or $\mathcal{P}_j$ holds. The fact that, for each $k = 1, \ldots, K-1$, $\mathcal{R}_{k+1}$ implies $\mathcal{R}_k$ and that $\mathcal{R}_1$ implies $\mathcal{M}$, then gives that $\widehat{\psi}_{MR}$ is CAN for ψ^* under the model that assumes that any, but not necessarily all, of the following conditions (i), (ii) or (iii) is satisfied: (i) $\mathcal{R}_K$ holds; (ii) for some $k \in \{1, \ldots., K-1\}$, model $\mathcal{R}_k$ and models $\mathcal{P}_{k+1}, \ldots, \mathcal{P}_K$ hold; (iii) model $\mathcal{M}$ and models $\mathcal{P}_1, \ldots, \mathcal{P}_K$ hold.

As indicated in Section 1.6.4, the remarkable fact (I) was first noticed in [53], and it was later shown to be a consequence of the likelihood factorization that takes place in coarsened at random models in [24]. For completeness, in Proposition 1 of Section 1.10 we give an independent proof of that fact.

Fact (III) follows from the fact that $\mathcal{R}_K$ is a regression model for the outcome Y on covariates $\overline{A}_K$ and $\overline{L}_K$, and in addition, for each $j \in \{1, \ldots, K\}$, under regularity conditions, $\widehat{\tau}_j$ converges to τ_j^* under $\mathcal{R}_K$.

As in the case of $K = 2$, fact (IV) follows applying results in [24]. Also, an heuristic argument can be given using counterfactuals and backward induction, generalizing the arguments used in Appendix A.2 to prove fact (IV) in the case in which $K = 2$.

Fact (V) follows from the facts that, when $\widehat{d}\left(\overline{A}_K, Z\right) = \overset{\bullet}{m}\left(\overline{A}_K, Z; \widehat{\psi}^{(K)}\right)$,

(i) $\widehat{\psi}_{MR}$ solves

$$\mathbb{P}_n \left\{ S^0_{\widehat{d}}(\psi, \widehat{\eta}_1) \right\} = 0$$

by step 5,

(ii) for each $k \in \{1, \ldots, K-1\}$,

$$\mathbb{P}_n \left\{ S^k_{\widehat{d}}(\widehat{\eta}_k, \widehat{\eta}_{k+1}, \widehat{\pi}_1, \ldots, \widehat{\pi}_k) \right\} = 0$$

by step 4 and the fact that $\overset{\bullet}{m}$ is a subvector of $\overset{\bullet}{\eta}_k$, and

(iii) $\mathbb{P}_n \left\{ S_{\widehat{d}}^K \left(\widehat{\eta}_K, \widehat{\pi} \right) \right\} = 0$ by step 3 and the fact that $\overset{\bullet}{m}$ is also a subvector of $\overset{\bullet}{\eta}_K$.

1.8 Example

We applied our methods to analyze data from the National Heart Lung and Blood Institute Growth and Health Study. The study sought to investigate racial differences in dietary, physical activity, family, and psychosocial factors associated with obesity from pre-adolescence through maturation between African-American and Caucasian girls. During 1987 and 1988, girls aged 9 and 10 were recruited from Richmond, CA and Cincinnati, OH and also from families enrolled in one HMO in the Washington, D.C. area. The follow-up period was 9 years with annual examinations. The recorded data included: anthropometric measurements, dietary information, physical activity and family socioeconomic status.

We analyzed data from the first three cycles of the study. We are interested in the effect of diet (A_k) on cycles $k = 1, 2, 3$ on the logarithm of BMI (Y) measured at cycle 4. To illustrate our methods we chose to dichotomize diet as $A_k = 1$ if the daily percentage of energy from saturated fatty acids, computed from data reported in the questionnaire of cycle k, was less than 10, and $A_k = 0$ otherwise. A thorough analysis would need to carefully consider the appropriate scale for diet. Our analysis used a baseline covariate $L_1 = (Z, V)$ where Z is race (1= Caucasian, 0= African-American) and $V = (V_1, V_2)$ with V_1 =household income at visit 1 (categorical variable with categories: 1-less than $ 10,000, 2-between $ 10,000 and $ 20,000, 3-between $ 20,000 and $ 40,000, 4-more than $ 40,000) and V_2 =logarithm of a physical activity score computed from responses to the questionnaire administered in cycle 1, a covariate L_2 =log(BMI) on cycle 2 and a covariate $L_3 = (L_{3,1}, L_{3,2})$ with $L_{3,1}$ =log(BMI) and $L_{3,2}$ =logarithm of physical activity score, both on cycle 3. The covariate L_2 did not include physical activity on cycle 2 because this variable was not recorded at that cycle.

Our analysis was based on 1303 children without missing data in any of the variables. We estimated the parameter ψ of the following MSMM

$$E(Y_{\overline{a}_3}|Z) = \psi_0 + Z\psi_1 + (\psi_2 + Z\psi_3)\, a_1 + (\psi_4 + Z\psi_5)\, a_2 + (\psi_6 + Z\psi_7)\, a_3.$$

This model assumes that, within strata $Z = z$ of race, the direct effect of diet at cycle k controlling by intervention the previous cycle diets is

$$\alpha_{k,z} \equiv \psi_{2k} + z\psi_{2k+1},$$

$k = 1, 2, 3$, and hence it does not change with the previous cycle diets.

To compute our MR estimator of ψ we postulated a working model

$$\rho_1(\overline{a}_3, l_1) = g_1(\overline{a}_3, z; \gamma_1)' \, t_1(v)$$

for

$$\rho_1(\overline{a}_3, l_1) \equiv E(Y_{\overline{a}_3}|A_1 = a_1, Z = z, V = v) - E(Y_{\overline{a}_3}|A_1 = a_1, Z = z, V = v_0)$$

where $l_1 = (z, v)$ and $v_0 = (v_{10}, v_{20}) = (4, 0)$ is a baseline level for V. We also postulated working models

$$\rho_k\left(\overline{a}_3, \overline{l}_k\right) = g_k\left(\overline{a}_3, \overline{l}_{k-1}; \gamma_k\right)' t_k(l_k)$$

for

$$\rho_k\left(\overline{a}_3, \overline{l}_k\right) \equiv E\left(Y_{\overline{a}_3}|\overline{A}_k = \overline{a}_k, \overline{L}_{k-1} = \overline{l}_{k-1}, L_k = l_k\right) - E\left(Y_{\overline{a}_3}|\overline{A}_k = \overline{a}_k, \overline{L}_{k-1} = \overline{l}_{k-1}, L_k = l_{k0}\right),$$

for $k = 2, 3$, where $l_{20} = 0$ and $l_{30} = (0, 0)$. In our models

$$t_1(V) = [I(V_1 = 1), I(V_1 = 2), I(V_1 = 3), V_2]'$$

and

$$t_k(L_k) = L_k$$

for $k = 2, 3$. The functions $g_1(\overline{a}_3, z; \gamma_1)$ and $g_k\left(\overline{a}_3, \overline{l}_{k-1}; \gamma_k\right), k = 2, 3$, were vector-valued functions of conformable dimension with j^{th} entry being of the form $[1, \overline{a}_3, z]\gamma_{1,j}$ and $\left[1, \overline{a}_3, \overline{l}_{k-1}\right]\gamma_{k-1,j}$ respectively. Also, our working models for $E\{t_1(V)|Z\}$ and $E\left\{t_k(L_k)|\overline{A}_{k-1}, \overline{L}_{k-1}\right\}$ were distinct linear models in all the conditioning variables and were estimated by ordinary least squares. In addition, our treatment models assumed logistic regressions for each A_k with linear terms in race, household income, the last prior diet, the last prior log(BMI) and the closest prior available logarithm of physical activity score.

As shown in Section 1.10 above for the case of the MR estimator, usual empirical sandwich variance techniques [50] can be used to derive estimators that are consistent for the asymptotic variances of $\widehat{\psi}_R$, $\widehat{\psi}_{DR}$ and $\widehat{\psi}_{MR}$ under the respective models in which they are CAN. This is because, ultimately these estimators solve estimating equations of the form $\sum_{i=1}^{n} \Psi_i(\theta) = 0$ with θ a parameter vector that includes ψ and several, finite dimensional, nuisance parameters. However, because handling the analytic expression for $\Psi_i(\theta)$ is cumbersome, in this section and in Section 1.9, we use instead the non-parametric bootstrap variance estimator which is consistent since $\widehat{\psi}_R$, $\widehat{\psi}_{DR}$ and $\widehat{\psi}_{MR}$ are regular and asymptotically linear estimators [13].

Table 3 reports the MR, DR, R and IPTW estimators of $\alpha_{k,z} \equiv \psi_{2k} + z\psi_{2k+1}, k = 1, 2, 3$ and $z = 0, 1$, their estimated standard errors (SE) and 95% Wald type confidence intervals using the bootstrap variance estimator from 1000 bootstrap samples, of which 16 were discarded due to lack of convergence. All but one of the estimated $\alpha_{k,z}$ are non-significant. Furthermore, all estimated $\alpha_{k,z}$ are negative, except for all estimators of the effect of diet at cycle 2 for Caucasians, i.e. of $\alpha_{2,1}$. Also, note that even though, for Caucasians $(Z = 1)$, the estimated effects of A_1 are greater than those of A_3, for all the estimators but the IPTW, the SEs are also higher. As predicted by theory, the IPTW estimator is the one with highest SE and the R estimator is the one with the lowest SE. Interestingly, the SEs of the MR and the DR estimators are similar.

Table 3

Estimators of $\alpha_{k,z} \equiv \psi_{2k} + z\psi_{2k+1}$ [bootstrap SE] (normal theory bootstrap 95% CI).

	MR	DR	R	IPTW
α_{10}	0[0.038] (-0.074,0.074)	-0.003[0.039] (-0.079,0.073)	-0.019[0.022] (-0.062,0.023)	-0.041[0.049] (-0.138,0.056)
α_{20}	-0.014[0.019] (-0.051,0.023)	-0.006[0.023] (-0.05,0.039)	0.005[0.012] (-0.019,0.029)	-0.047[0.048] (-0.14,0.047)
α_{30}	-0.009[0.014] (-0.036,0.017)	-0.007[0.017] (-0.041,0.027)	-0.008[0.008] (-0.023,0.008)	-0.028[0.05] (-0.125,0.069)
α_{11}	-0.043[0.027] (-0.096,0.01)	-0.035[0.026] (-0.085,0.015)	-0.018[0.024] (-0.065,0.029)	-0.032[0.026] (-0.084,0.019)
α_{21}	0.006[0.014] (-0.021,0.033)	0.011[0.016] (-0.02,0.043)	0.015[0.009] (0.002,0.033)	0.007[0.026] (-0.044,0.057)
α_{31}	-0.008[0.011] (-0.03,0.014)	-0.007[0.013] (-0.032,0.019)	-0.006[0.008] (-0.021,0.008)	-0.045[0.025] (-0.094,0.003)

1.9 A simulation study

We conducted a simulation study under a scenario that roughly mimics the data structure in the study of Section 1.8, excluding the baseline variable Z. We generated 1000 samples, each of size 1000 according to the data generating process in Table 4 with parameter values given in Table 5. Under this process and the identifying assumptions of Section 1.3,

$$E(Y_{\overline{a}_3}) = m\left(\overline{a}_3;\psi\right) \equiv \psi_0 + \psi_1 a_1 + \psi_2 a_2 + \psi_3 a_3.$$

Also, the following holds

$$E\left(Y_{\overline{a}_3}|A_1 = a_1, L_1\right) = m\left(\overline{a}_3;\psi\right) + g_1\left\{t_1\left(L_1\right) - e_1\right\} \tag{1.22}$$

$$E\left(Y_{\overline{a}_3}|\overline{A}_2 = \overline{a}_2, \overline{L}_2\right) = m\left(\overline{a}_3;\psi\right) + \textstyle\sum_{j=1}^{2} g_j\left\{t_j\left(L_j\right) - e_j\right\} \tag{1.23}$$

$$E\left(Y_{\overline{a}_3}|\overline{A}_3 = \overline{a}_3, \overline{L}_3\right) = m\left(\overline{a}_3;\psi\right) + \textstyle\sum_{j=1}^{3} g_j\left\{t_j\left(L_j\right) - e_j\right\} \tag{1.24}$$

$$E\left\{t_1\left(L_1\right)\right\} = e_1\left(\tau_1\right) \tag{1.25}$$

$$E\left\{t_2\left(L_2\right)|A_1, L_1\right\} = e_2\left(A_1, L_1; \tau_2\right) \tag{1.26}$$

$$E\left\{t_3\left(L_3\right)|\overline{A}_2, \overline{L}_2\right\} = e_3\left(\overline{A}_2, \overline{L}_2; \tau_3\right) \tag{1.27}$$

with $L_1 = V \equiv (V_1, V_2)$ and with g_j, t_j and $e_j, j = 1, 2, 3$ defined in Table 4. In particular, the following models hold. Model $\mathcal{R}_1$ defined by restrictions (1.22) and (1.25), $\mathcal{R}_2$ defined by restrictions (1.23), (1.25) and (1.26), and $\mathcal{R}_3$ defined by restrictions (1.24), (1.25), (1.26) and (1.27). Finally, the data also follows models $\mathcal{P}_k$ given by $\text{logit}\left\{\Pr\left(A_k = 1|\overline{A}_{k-1}, \overline{L}_k\right)\right\} = c_k\left(\overline{A}_{k-1}, \overline{L}_k\right)'\alpha_k$ with $c_k(\cdot,\cdot)$ given in Table 4, $1 \leq k \leq 3$.

We considered the following six scenarios: (1) $\mathcal{R}_k$ and $\mathcal{P}_k$ correct, $k = 1, 2, 3$, (2) only $\mathcal{R}_k$ correct, $k = 1, 2, 3$, (3) only $\mathcal{R}_1, \mathcal{R}_2$ and $\mathcal{P}_3$ correct, (4) only $\mathcal{R}_1, \mathcal{P}_2$ and $\mathcal{P}_3$ correct, (5) only $\mathcal{P}_k$ correct, $k = 1, 2, 3$ (6) $\mathcal{R}_k$ and $\mathcal{P}_k$ incorrect, $k = 1, 2, 3$.
The incorrect $\mathcal{P}_1, \mathcal{P}_2$ and $\mathcal{P}_3$ models were logistic regression models with covariates $\widetilde{c}_1\left(L_1\right) \equiv [1, V_1]$, $\widetilde{c}_2\left(A_1, \overline{L}_2\right) \equiv \left[1, \overline{L}_2\right]$, $\widetilde{c}_3\left(\overline{A}_2, \overline{L}_3\right) \equiv [1, V_1, A_2, L_{3,2}]$ respectively. Different choices were used to build the incorrect $\mathcal{R}_k$ models, depending on the scenario. Specifically, in scenario (3) g_3, t_3, and e_3 were incorrect, in scenario (4) g_2, e_2, g_3, t_3, and e_3 were incorrect and in scenarios (5) and (6) the functions g_1, g_2, e_2, g_3, t_3, and e_3 were incorrect. When g_j was misspecified the function $\widetilde{g}_j\left(\overline{a}_3, \overline{L}_{j-1}; \widetilde{\gamma}_j\right) \equiv \widetilde{\gamma}_j$ with $\widetilde{\gamma}_j$ of conformable dimension was used, $j = 1, 2, 3$. For the remaining functions the following incorrect choices were used: $\widetilde{t}_3\left(L_3\right) \equiv L_3, \widetilde{e}_2\left(A_1, L_1; \widetilde{\tau}_2\right) \equiv [1, L_1]\,\widetilde{\tau}_2$ and $\widetilde{e}_3\left(\overline{A}_2, \overline{L}_2; \widetilde{\tau}_3\right) \equiv [\widetilde{e}_{3,1}, \widetilde{e}_{3,2}]'$ with $\widetilde{e}_{3,1} \equiv \left[1, L_1, \overline{A}_2\right]\widetilde{\tau}_{3,1}$, $\widetilde{e}_{3,2} \equiv \left[1, \overline{L}_2\right]\widetilde{\tau}_{3,2}$.

We computed Wald type confidence intervals using the bootstrap variance estimator with 1000 bootsptrap replications. Figure 1 reports results for $\widehat{\psi}_{MR}, \widehat{\psi}_{DR}, \widehat{\psi}_{R}$ and $\widehat{\psi}_{IPTW}$. Observe that, as predicted by theory, $\widehat{\psi}_{IPTW}$ was virtually unbiased if every $\mathcal{P}_k$ was correct, but considerably biased otherwise. Similarly, $\widehat{\psi}_{R}$ was unbiased when every $\mathcal{R}_k$ was correct but badly biased otherwise. Also, $\widehat{\psi}_{DR}$ was virtually unbiased when the sequence of $\mathcal{R}_k$ models or the sequence of $\mathcal{P}_k$ models was correct. Interestingly, it was also virtually unbiased under scenario (4), but it was considerably biased otherwise. In contrast, $\widehat{\psi}_{MR}$ was unbiased in all scenarios except when all the models were incorrect. Even in this unfavorable scenario its bias was smaller than the bias of the other estimators. The graphs reporting the SEs indicate that, as predicted by theory, $\widehat{\psi}_{IPTW}$ had larger SEs. Interestingly the SEs of $\widehat{\psi}_{MR}$ and $\widehat{\psi}_{DR}$ were always very similar only slightly larger than those

of $\widehat{\psi}_R$ under scenarios in which the three estimators are consistent. As for confidence intervals, those centered at the MR estimators, had actual coverage probability very close to the nominal 95% in all scenarios including, surprisingly, the case with all models incorrect. Intervals centered at the remaining estimators had actual coverage probability smaller than the nominal 95% in at least one scenario.

Figure 1: *simulation results*

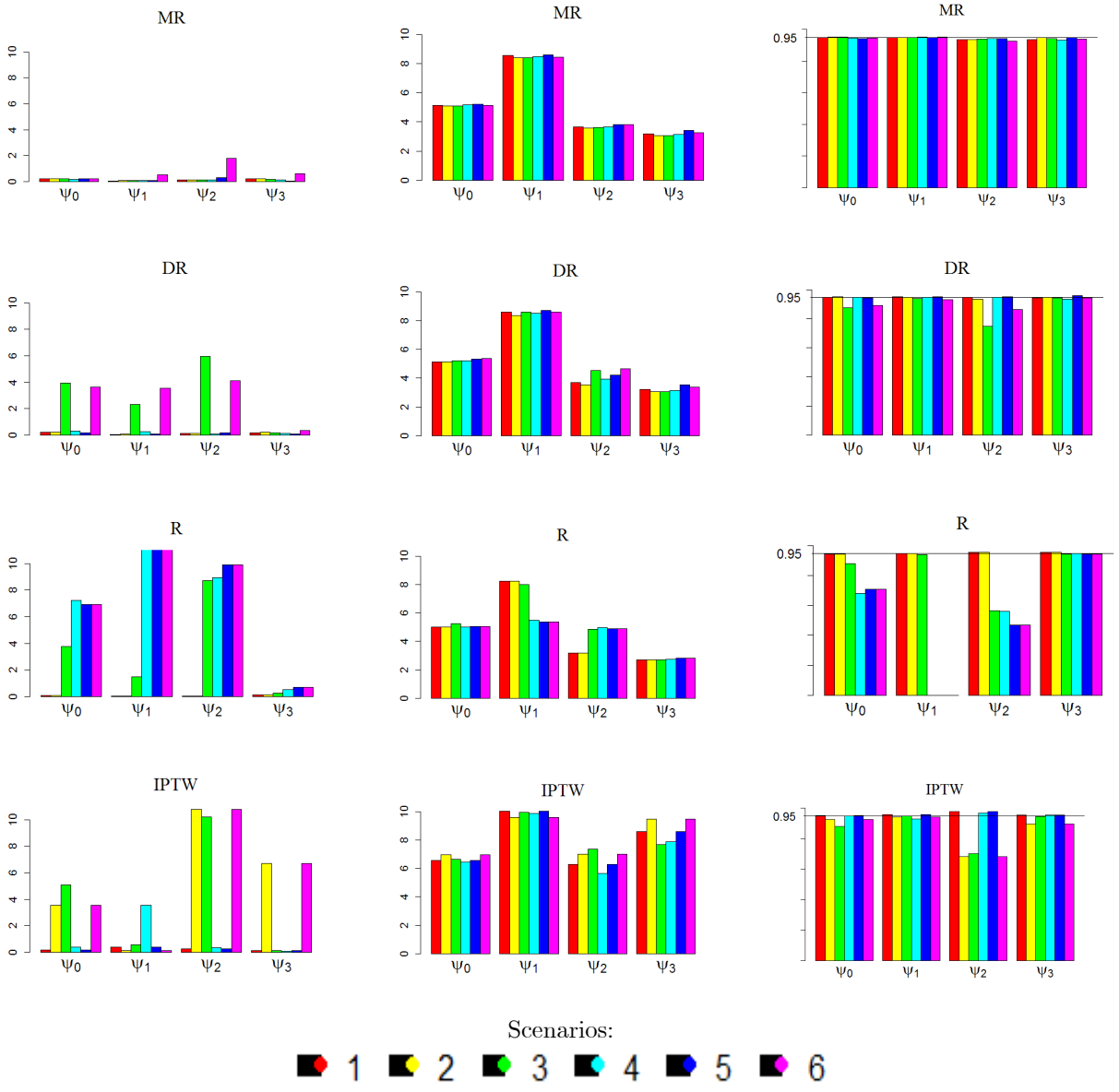

Table 4

Data generating process in the simulation study

$L_1 = [V_1, V_2]'$; $V_1 = \sum_{j=1}^{4} jI_j$ where $I \equiv [I_1, I_2, I_3, I_4] \sim Mult([p_1, p_2, p_3, p_4], 1)$

$V_2|V_1 \sim N\left(\mu_1(V_1), \sigma_1^2(V_1)\right)$, $L_2|(A_1, L_1) \sim N\left(\mu_2, \sigma_2^2\right)$; $\mu_2 \equiv [1, L_1, A_1, V_1V_2, L_1A_1]\,\tau_2$

$L_3 = [L_{3,1}, L_{3,2}]'$; $L_3|\left(\overline{A}_2, \overline{L}_2\right) \sim N\left(\left[\begin{array}{c} \mu_{3,1} \\ \mu_{3,2} \end{array}\right], \Sigma\right)$; $\left[\begin{array}{c} \mu_{3,1} \\ \mu_{3,2} \end{array}\right] \equiv \left[\begin{array}{c} [1, \overline{L}_2, \overline{A}_2, L_2A_2, V_1L_2, V_2A_1]\,\tau_{3,1} \\ [I, IV_2, IL_2, IA_2]\,\tau_{3,2} \end{array}\right]$

$Y|\left(\overline{A}_3, \overline{L}_3\right) \sim N\left(\mu_4, \sigma_4^2\right)$; $\mu_4 \equiv m\left(\overline{A}_3; \psi\right) + \sum_{j=1}^{3} g_j\left\{t_j(L_j) - e_j\right\}$; $m\left(\overline{A}_3; \psi\right) = \left[1, \overline{A}_3\right]\psi$

$g_1 \equiv [g_{1,1}, g_{1,2}, g_{1,3}, g_{1,4}]$; $g_{1,j} \equiv \left[1, \overline{A}_3\right]\gamma_{1,j}$, $g_2 \equiv \left[1, L_1, \overline{A}_3\right]\gamma_2$

$g_3 \equiv [g_{3,1}, g_{3,2}, g_{3,3}]$; $g_{3,j} = \left[1, V_1, L_2, \overline{A}_3\right]\gamma_{3,j}$; $t_1(L_1) = [I_1, I_2, I_3, V_2]'$; $t_2(L_2) = L_2$

$t_3(L_3) = [L_{3,1}, L_{3,2}, L_{3,1}L_{3,2}]'$; $e_1 \equiv \tau_1 \equiv \left[p_1, p_2, p_3, \sum_{j=1}^{4} p_j\mu_1(j)\right]'$; $e_2 \equiv \mu_2$

$e_3\left(\overline{A}_2, \overline{L}_2; \tau_3\right) \equiv [\mu_{3,1}, \mu_{3,2}, \Sigma_{1,2} + \mu_{3,1}\mu_{3,2}]'$ with $\tau_3 \equiv [\mu_{3,1}, \mu_{3,2}, \Sigma_{1,2}]'$

A_k binary; $\Pr(A_1 = 1|L_1) = \text{expit}\left\{c_1(L_1)'\alpha_1\right\}$

$\Pr\left(A_k = 1|\overline{A}_{k-1}, \overline{L}_k\right) = \text{expit}\left\{c_k\left(\overline{A}_{k-1}, \overline{L}_k\right)'\alpha_k\right\}, k = 2, 3$

$c_1(L_1) \equiv [I, IV_2]'$; $c_2\left(A_1, \overline{L}_2\right) \equiv [1, L_1, A_1, L_2, V_1V_2, L_1A_1, L_1L_2, A_1L_2]'$

$c_3\left(\overline{A}_2, \overline{L}_3\right) = [1, V_1, A_2, L_3, V_1A_2, V_1L_3, A_2L_3, L_{3,1}L_{3,2}]'$

Table 5

Parameter values for the data generating process of the simulation study

$[p_1, p_2, p_3, p_4] = [.15, .2, .3, .35]$; $[\mu_1(1), \mu_1(2), \mu_1(3), \mu_1(4)] = [3.35, 3.37, 3.38, 3.42]$

$[\sigma_1(1), \sigma_1(2), \sigma_1(3), \sigma_1(4)] = [0.48, 0.45, 0.43, 0.44]$; $\tau_2 = [2.98, -.01, -.02, -.05, 0, -.01, -.01]'$

$\sigma_2 = 0.12$; $\tau_{3,1} = [.03, .45, -.02, .01, 1.17, .1, -.06, -.16, -.01]'$

$\tau_{3,2} = [3.18, 3.32, 3.47, 3.61, .3, .22, .15, .07, -.55, -.47, -.38, -.29, .45, .38, .31, .24]'$

$\Sigma = \left[\begin{array}{cc} 1.26 \times 10^{-3} & -5.26 \times 10^{-4} \\ -5.26 \times 10^{-4} & 0.28 \end{array}\right]$; $\sigma_4 = .04$; $\psi = [3.22, -.05, -.075, -.1]'$

$\gamma_{1,1} = [.12, .05, -.04, -.02]'$; $\gamma_{1,2} = [.07, .04, -.02, -.01]'$; $\gamma_{1,3} = [.04, -.01, -.02, -.01]'$

$\gamma_{1,4} = [-.03, -.01, .01, .01]'$; $\gamma_2 = [1.15, -.12, .01, -.05, -.06, .1]'$

$\gamma_{3,1} = [1.66, -.06, .04, .09, -.34, -.37]'$; $\gamma_{3,2} = [.54, -.04, -.2, -.05, -.24, -.1]'$

$\gamma_{3,3} = [-.12, .01, .04, .02, .08, .04]'$; $\alpha_1 = [-3.23, -2.39, -1.85, -1.49, .50, .39, .34, .32]'$

$\alpha_2 = [4.5, -1.2, -.69, 3.26, -1.35, .2, .3, .24, -.58, .08, -.48]'$

$\alpha_3 = [7, -1.04, 5.81, -3.04, -1.2, -.21, .55, -.15, -1.17, -.33, .5]'$

1.10 Consistency and asymptotic normality of the MR estimator

In this section, we formally prove the multiple robustness of the MR estimator $\widehat{\psi}_{MR}$ proposed in Section 1.7.2. That is, we provide regularity conditions ensuring its consistency and asymptotic normality under the model that assumes that any, but not necessarily all, of the following conditions (i), (ii) or (iii) is satisfied: (i) $\mathcal{R}_K$ holds; (ii) for some $k \in \{1, \ldots, K-1\}$, model $\mathcal{R}_k$ and models $\mathcal{P}_{k+1}, \ldots, \mathcal{P}_K$ hold; (iii) model $\mathcal{M}$ and models $\mathcal{P}_1, \ldots, \mathcal{P}_K$ hold. Throughout this section, P denotes the distribution of the observed data $O = (L_1, A_1, \ldots, L_K, A_K, Y)$. Also, throughout this section, $\pi_k, k = 1, \ldots, K, \eta_k, k = 0, \ldots, K$, and $\rho_k, k = 1, \ldots, K$, are the functionals defined in Sections 1.4.1, 1.4.2 and 1.7.1 respectively. Notice that, although all these functionals depend on P, for simplicity we omit that subscript. We start by providing rigorous definitions of model $\mathcal{M}$ and models $\mathcal{R}_k$ and $\mathcal{P}_k, k = 1, \ldots, K$.

Let

$$\mathcal{M} \equiv \{P : \exists! \psi(P) \in \Xi \subseteq R^p \text{ such that } \eta_0(\overline{a}_K, z) = m(\overline{a}_K, z; \psi(P)) \ \}$$

where $m(\overline{a}_K, z; \cdot)$ is a user-specified real-valued function of $\psi \in R^p$.

For $k = 1, \ldots, K$, let

$$\mathcal{P}_k \equiv \left\{P : \exists! \alpha_k(P) \in \Delta_k \subseteq R^{a_k} \text{ such that } \pi_k\left(\overline{a}_k, \overline{l}_k\right) = \pi_k\left(\overline{a}_k, \overline{l}_k; \alpha_k(P)\right)\right\}$$

where $\pi_k\left(\overline{a}_k, \overline{l}_k; \cdot\right)$ is a user-specified real-valued function of $\alpha_k \in R^{a_k}$.

Let

$$\mathcal{S}_1 \equiv \left\{P : \exists! \gamma_1(P) \in \Gamma_1 \subseteq R^{s_1} \text{ such that } \rho_1(\overline{a}_K, l_1) = g_1(\overline{a}_K, z; \gamma_1(P))' t_1(v)\right\}$$

where $g_1(\overline{a}_K, z; \cdot)$ is a user-specified vector-valued function of $\gamma_1 \in R^{s_1}$ and $t_1(\cdot)$ is a user-specified conformable vector-valued function verifying $t_1(v_0) = 0$ with v_0 any baseline level of V.

For $k = 2, \ldots, K$, let

$$\mathcal{S}_k \equiv \left\{P : \exists! \gamma_k(P) \in \Gamma_k \subseteq R^{s_k} \text{ such that } \rho_k\left(\overline{a}_K, \overline{l}_k\right) = g_k\left(\overline{a}_K, \overline{l}_{k-1}; \gamma_k(P)\right)' t_k(l_k)\right\}$$

where $g_k\left(\overline{a}_K, \overline{l}_{k-1}; \cdot\right)$ is a user-specified conformable vector-valued function of $\gamma_k \in R^{s_k}$ and $t_k(\cdot)$ is a user-specified conformable vector-valued function verifying $t_k(l_{k,0}) = 0$ with $l_{k,0}$ any baseline level of L_k.

Let

$$\mathcal{E}_1 \equiv \{P : \exists! \tau_1(P) \in \Upsilon_1 \subseteq R^{r_1} \text{ such that } E_P\{t_1(V) | Z = z\} = e_1(z; \tau_1(P))\}$$

where $e_1(z; \cdot)$ is a user-specified conformable vector-valued function of $\tau_1 \in R^{r_1}$.

For $k = 2, \ldots, K$, let

$$\mathcal{E}_k \equiv \left\{P : \exists! \tau_k(P) \in \Upsilon_k \subseteq R^{r_k} \text{ such that } E_P\left\{t_k(L_k) | \overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_{k-1} = \overline{l}_{k-1}\right\} = e_k\left(\overline{a}_{k-1}, \overline{l}_{k-1}; \tau_k(P)\right)\right\}$$

where $e_k\left(\overline{a}_{k-1}, \overline{l}_{k-1}; \cdot\right)$ is a user-specified conformable vector-valued function of $\tau_k \in R^{r_k}$.

Finally, for $k = 1, \ldots, K$, let

$$\mathcal{R}_k \equiv \mathcal{M} \cap \left(\bigcap_{j=1}^{k} \mathcal{S}_j\right) \cap \left(\bigcap_{j=1}^{k} \mathcal{E}_j\right).$$

We will now provide regularity conditions under which $\sqrt{n}\left(\widehat{\psi}_{MR}-\psi\left(P\right)\right)$ converges to a mean-zero normal distribution for every $P\in\mathcal{F}$ with

$$\mathcal{F}\equiv\mathcal{R}_{K}\cup\bigcup_{k=1}^{K-1}\left(\mathcal{R}_{k}\cap\bigcap_{j=k+1}^{K}\mathcal{P}_{j}\right)\cup\left(\mathcal{M}\cap\bigcap_{j=1}^{K}\mathcal{P}_{j}\right).$$

Later, we will derive the asymptotic variance $AVar\left(\widehat{\psi}_{MR}\right)$ of $\widehat{\psi}_{MR}$, i.e. the variance of the limiting normal distribution of $\sqrt{n}\left(\widehat{\psi}_{MR}-\psi\left(P\right)\right)$ and we will also provide a consistent -under $\mathcal{F}$- estimator of it.

Our derivation of the consistency and asymptotic normality of $\widehat{\psi}_{MR}$ under $\mathcal{F}$ relies on the following facts.

(I) The elements of $\widehat{\psi}_{MR}$ are the last p components of a vector $\widehat{\theta}$ defined in (1.28) below that, under the assumptions of Lemma 1 below, solves the equation $\mathbb{P}_{n}\left(\phi_{\theta}\right)=0$, with ϕ_{θ} defined in (1.29) below.

(II) For every $P\in\mathcal{F}$ verifying Condition SPob below, $E_{P}\left(\phi_{\theta}\right)=0$ has a unique solution, denoted in (1.37) by $\theta^{\dagger}\left(P\right)$, whose last p components are the elements of $\psi\left(P\right)$.

(III) Under the assumptions of Lemma 3 below, $\widehat{\theta}$ is CAN for $\theta^{\dagger}\left(P\right)$ under model $\mathcal{F}$.

The results presented in this section are proved in Appendix A.3.

We start with statement (I). Define

$$\widehat{\theta}\equiv\left(\widehat{\overline{\alpha}}_{K},\widehat{\overline{\tau}}_{K},\widehat{\psi}^{(K)},\widehat{\overline{\gamma}}_{K}^{(K)},\widehat{\psi}^{(K-1)},\widehat{\overline{\gamma}}_{K-1}^{(K-1)},\ldots,\widehat{\psi}^{(1)},\widehat{\gamma}_{1}^{(1)},\widehat{\psi}_{MR}\right). \tag{1.28}$$

with $\widehat{\overline{\alpha}}_{K},\widehat{\overline{\tau}}_{K},\left(\widehat{\psi}_{k}^{(k)},\widehat{\overline{\gamma}}_{k}^{(k)}\right),k=1,\ldots,K$, and $\widehat{\psi}_{MR}$ the estimators defined in steps 1, 2, 3-4 and 5 respectively of Algorithm 1 of Section 1.7.2. Also define

$$\theta\equiv\left(\overline{\alpha}_{K},\overline{\tau}_{K},\psi^{(K)},\overline{\gamma}_{K}^{(K)},\psi^{(K-1)},\overline{\gamma}_{K-1}^{(K-1)},\ldots,\psi^{(1)},\gamma_{1}^{(1)},\psi\right)$$

as a free parameter vector with each component having the same dimension as the corresponding component in $\widehat{\theta}$. For instance, for $k=1,\ldots,K$, $\overline{\gamma}_{k}^{(k)}\equiv\left(\gamma_{1}^{(k)},\ldots,\gamma_{k}^{(k)}\right)$ with each $\gamma_{j}^{(k)},j=1,\ldots,k$, having the same dimension as $\widehat{\gamma}_{j}^{(k)}$. Note that, for every $k,l\geq j$, $\gamma_{j}^{(k)}$ and $\gamma_{j}^{(l)}$ have the same dimension. We also define

$$\overline{\gamma}_{k}\equiv\left(\gamma_{1},\ldots,\gamma_{k}\right)$$

with each $\gamma_{j},j=1,\ldots,k$, a free parameter vector with the same dimension as $\gamma_{j}^{(k)}$. We let

$$q\equiv\sum_{k=1}^{K}a_{k}+\sum_{k=1}^{K}r_{k}+\sum_{k=1}^{K}\left(K-k+1\right)s_{k}+\left(K+1\right)p$$

be the dimension of θ.

Note that, if each of the equations in steps 3 to 5 of Algorithm 1 has at least one solution, then $\widehat{\theta}$ solves a joint system of estimating equations given by $\mathbb{P}_n\left(\Psi_\theta\right)=0$ where

$$\Psi_\theta \equiv \left(\Psi_\theta^1,\ldots,\Psi_\theta^K,\Psi_\theta^{K+1},\ldots,\Psi_\theta^{2K},\Psi_\theta^{2K+1},\ldots,\Psi_\theta^{3K+1}\right)',$$

with each $\Psi_\theta^j, j=1,\ldots,3K+1$ defined next. For $k=1,\ldots,K$,

$$\Psi_\theta^k\left(o\right) \equiv \frac{\partial}{\partial\alpha_k}\ln\pi_k\left(\overline{a}_k,\overline{l}_k;\alpha_k\right),$$

$$\Psi_\theta^{K+1}\left(o\right) \equiv q_1\left(z\right)\left\{t_1\left(v\right)-e_1\left(z;\tau_1\right)\right\}$$

and, for $k=2,\ldots,K$,

$$\Psi_\theta^{K+k}\left(o\right) \equiv q_k\left(\overline{a}_{k-1},\overline{l}_{k-1}\right)\left\{t_k\left(l_k\right)-e_k\left(\overline{a}_{k-1},\overline{l}_{k-1};\tau_k\right)\right\}.$$

For notational convenience, we index the estimating functions $\Psi_\theta^{2K+1},\ldots,\Psi_\theta^{3K+1}$ by Ψ_θ^{3K+1-k}, $k=K,\ldots,0$. We do so because, as will become clear next, for $k=K,\ldots,1,\Psi_\theta^{3K+1-k}$ is the estimating function used to compute $\left(\widehat{\psi}_k^{(k)},\widehat{\overline{\gamma}}_k^{(k)}\right)$ and Ψ_θ^{3K+1} is the estimating function used to compute $\widehat{\psi}_{MR}$. We denote $\overset{\bullet}{\eta}_k\left(\overline{a}_K,\overline{l}_k;\psi,\overline{\gamma}_k,\overline{\tau}_k\right) \equiv \partial\eta_k\left(\overline{a}_K,\overline{l}_k;\psi,\overline{\gamma}_k,\overline{\tau}_k\right)/\partial\left(\psi,\overline{\gamma}_k\right), k=1,\ldots,K$, and $\overset{\bullet}{m}\left(\overline{a}_K,z;\psi\right) \equiv \partial m\left(\overline{a}_K,z;\psi\right)/\partial\psi$. Now, when $k=K,\Psi_\theta^{3K+1-k}=\Psi_\theta^{2K+1}$ with

$$\Psi_\theta^{2K+1}\left(o\right) \equiv \frac{\overset{\bullet}{\eta}_K\left(\overline{a}_K,\overline{l}_K;\psi^{(K)},\overline{\gamma}_K^{(K)},\overline{\tau}_K\right)}{\prod_{s=1}^K\pi_s\left(\overline{a}_s,\overline{l}_s;\alpha_s\right)}\left\{y-\eta_K\left(\overline{a}_K,\overline{l}_K;\psi^{(K)},\overline{\gamma}_K^{(K)},\overline{\tau}_K\right)\right\},$$

for $k=K-1,\ldots,1$,

$$\Psi_\theta^{3K+1-k}\left(o\right) \equiv \sum_{\underline{a}'_{k+1}\in\underline{\mathcal{A}}_{k+1}}\frac{\overset{\bullet}{\eta}_k\left(\overline{a}_k,\underline{a}'_{k+1},\overline{l}_k;\psi^{(K)},\overline{\gamma}_k^{(K)},\overline{\tau}_k\right)}{\prod_{s=1}^k\pi_s\left(\overline{a}_s,\overline{l}_s;\alpha_s\right)}\left\{\eta_{k+1}\left(\overline{a}_k,\underline{a}'_{k+1},\overline{l}_{k+1};\psi^{(k+1)},\overline{\gamma}_{k+1}^{(k+1)},\overline{\tau}_{k+1}\right)\right.$$
$$\left.-\eta_k\left(\overline{a}_k,\underline{a}'_{k+1},\overline{l}_k;\psi^{(k)},\overline{\gamma}_k^{(k)},\overline{\tau}_k\right)\right\},$$

and when $k=0,\Psi_\theta^{3K+1-k}=\Psi_\theta^{3K+1}$ with

$$\Psi_\theta^{3K+1}\left(o\right) \equiv \sum_{\underline{a}'_1\in\underline{\mathcal{A}}_1}\overset{\bullet}{m}\left(\underline{a}'_1,z;\psi^{(K)}\right)\left\{\eta_1\left(\underline{a}'_1,l_1;\psi^{(1)},\gamma_1^{(1)},\tau_1\right)-m\left(\underline{a}'_1,z;\psi\right)\right\}.$$

In Lemma 1, generalizing the argument used to prove facts (IV) and (V) of Subsection 1.6.4, we show that if each one of the equations in steps 3 to 5 of Algorithm 1 has at least one solution, then $\widehat{\theta}$ actually solves another system of estimating equations given by $\mathbb{P}_n\left(\phi_\theta\right)=0$ with

$$\phi_\theta \equiv \left(\phi_\theta^1,\ldots,\phi_\theta^K,\phi_\theta^{K+1},\ldots,\phi_\theta^{2K},\phi_\theta^{2K+1},\ldots,\phi_\theta^{3K+1}\right)', \tag{1.29}$$

where, for $k=1,\ldots,K$,

$$\phi_\theta^k \equiv \Psi_\theta^k$$

and

$$\phi_\theta^{K+k} \equiv \Psi_\theta^{K+k},$$
$$\phi_\theta^{2K+1} \equiv \Psi_\theta^{2K+1}$$

and, for $k = 0, \ldots, K-1$,

$$\phi_\theta^{3K+1-k} \equiv \sum_{j=k}^{K} \varphi_\theta^{k,j} \tag{1.30}$$

where, for $k = 1, \ldots, K-1$

$$\varphi_\theta^{k,j}(o) \equiv \sum_{\underline{a}'_{j+1} \in \underline{\mathcal{A}}_{j+1}} \frac{\overset{\bullet}{\eta}_k\left(\overline{a}_j, \underline{a}'_{j+1}, \overline{l}_k; \psi^{(K)}, \overline{\gamma}_k^{(K)}, \overline{\tau}_k\right)}{\prod_{s=1}^{j} \pi_s\left(\overline{a}_s, \overline{l}_s; \alpha_s\right)} \left\{\eta_{j+1}\left(\overline{a}_j, \underline{a}'_{j+1}, \overline{l}_{j+1}; \psi^{(j+1)}, \overline{\gamma}_{j+1}^{(j+1)}, \overline{\tau}_{j+1}\right)\right. \tag{1.31}$$
$$\left. -\eta_j\left(\overline{a}_j, \underline{a}'_{j+1}, \overline{l}_j; \psi^{(j)}, \overline{\gamma}_j^{(j)}, \overline{\tau}_j\right)\right\},$$

for $j = k, \ldots, K-1$, and

$$\varphi_\theta^{k,K}(o) \equiv \frac{\overset{\bullet}{\eta}_k\left(\overline{a}_K, \overline{l}_k; \psi^{(K)}, \overline{\gamma}_k^{(K)}, \overline{\tau}_k\right)}{\prod_{s=1}^{K} \pi_s\left(\overline{a}_s, \overline{l}_s; \alpha_s\right)} \left\{y - \eta_K\left(\overline{a}_K, \overline{l}_K; \psi^{(K)}, \overline{\gamma}_K^{(K)}, \overline{\tau}_K\right)\right\}. \tag{1.32}$$

Also, for $k = 0$,

$$\varphi_\theta^{0,0}(o) \equiv \sum_{\underline{a}'_1 \in \underline{\mathcal{A}}_1} \overset{\bullet}{m}\left(\underline{a}'_1, z; \psi^{(K)}\right) \left\{\eta_1\left(\underline{a}'_1, l_1; \psi^{(1)}, \gamma_1^{(1)}, \tau_1\right) - m\left(\underline{a}'_1, z; \psi\right)\right\}, \tag{1.33}$$

$$\varphi_\theta^{0,j}(o) \equiv \sum_{\underline{a}'_{j+1} \in \underline{\mathcal{A}}_{j+1}} \frac{\overset{\bullet}{m}\left(\overline{a}_j, \underline{a}'_{j+1}, z; \psi^{(K)}\right)}{\prod_{s=1}^{j} \pi_s\left(\overline{a}_s, \overline{l}_s; \alpha_s\right)} \left\{\eta_{j+1}\left(\overline{a}_j, \underline{a}'_{j+1}, \overline{l}_{j+1}; \psi^{(j+1)}, \overline{\gamma}_{j+1}^{(j+1)}, \overline{\tau}_{j+1}\right)\right. \tag{1.34}$$
$$\left. -\eta_j\left(\overline{a}_j, \underline{a}'_{j+1}, \overline{l}_j; \psi^{(j)}, \overline{\gamma}_j^{(j)}, \overline{\tau}_j\right)\right\},$$

for $j = 1, \ldots, K-1$, and

$$\varphi_\theta^{0,K}(o) \equiv \frac{\overset{\bullet}{m}\left(\overline{a}_K, z; \psi^{(K)}\right)}{\prod_{s=1}^{K} \pi_s\left(\overline{a}_s, \overline{l}_s; \alpha_s\right)} \left\{y - \eta_K\left(\overline{a}_K, \overline{l}_K; \psi^{(K)}, \overline{\gamma}_K^{(K)}, \overline{\tau}_K\right)\right\}. \tag{1.35}$$

Note that, for $k = 0, \ldots, K-1$, $\varphi_\theta^{k,k} = \Psi_\theta^{3K+1-k}$.

Lemma 1 *If each one of the equations in steps 3 to 5 of Algorithm 1 has at least one solution, then $\widehat{\theta}$ solves $\mathbb{P}_n(\phi_\theta) = 0$.*

To see facts (II) and (III) we need to introduce the following notation. For $k = 1, \ldots, K$, define the following subvectors of θ :

$$\begin{aligned} \theta_k &\equiv \alpha_k, \\ \theta_{K+k} &\equiv \tau_k, \\ \theta_{3K+1-k} &\equiv \left(\psi^{(k)}, \overline{\gamma}_k^{(k)}\right) \end{aligned}$$

and also define

$$\theta_{3K+1} \equiv \psi.$$

Note that, with these definitions, θ can be written as

$$\theta = \left(\theta_1, \ldots, \theta_K, \theta_{K+1}, \ldots, \theta_{2K}, \theta_{2K+1}, \ldots, \theta_{3K+1}\right).$$

Also, for $j = 1, \ldots, 3K+1$, let $\overline{\theta}_j \equiv (\theta_1, \ldots, \theta_j)$. Note that, for $k = 1, \ldots, K, \overline{\theta}_{3K+1-k} \equiv (\theta_1, \ldots, \theta_{3K+1-k}) \equiv \left(\overline{\alpha}_K, \overline{\tau}_K, \psi^{(K)}, \overline{\gamma}_K^{(K)}, \ldots, \psi^{(k)}, \overline{\gamma}_k^{(k)}\right)$. Analogously, define the subvectors of $\widehat{\theta}$, $\widehat{\theta}_k, \widehat{\theta}_{K+k}, \widehat{\theta}_{3K+1-k}, k = 1, \ldots, K, \widehat{\theta}_{3K+1}$ and $\widehat{\overline{\theta}}_{3K+1-k}, k = 1, \ldots, K$. Note that, for each $k = 1, \ldots, K$, ϕ_θ^k, ϕ_θ^{K+k} and ϕ_θ^{3K+1-k} depend on θ only through α_k, τ_k and $\left(\overline{\alpha}_K, \overline{\tau}_K, \psi^{(K)}, \overline{\gamma}_K^{(K)}, \ldots, \psi^{(k)}, \overline{\gamma}_k^{(k)}\right)$ respectively. Likewise, for $k = 0, \ldots, K-1$ and $j = k, \ldots, K$, $\varphi_\theta^{k,j}$ depends on θ only through a subvector of $\left(\overline{\alpha}_K, \overline{\tau}_K, \psi^{(K)}, \overline{\gamma}_K^{(K)}, \ldots, \psi^{(j)}, \overline{\gamma}_k^{(j)}\right)$. Hence, for $k = 1, \ldots, K$, we will write indistinctly

$$\phi_\theta^k, \phi_{\alpha_k}^k \text{ or } \phi_{\theta_k}^k$$

$$\phi_\theta^{K+k}, \phi_{\tau_k}^{K+k} \text{ or } \phi_{\theta_{K+k}}^{K+k}$$

and

$$\phi_\theta^{3K+1-k}, \phi_{\left(\overline{\alpha}_K, \overline{\tau}_K, \psi^{(K)}, \overline{\gamma}_K^{(K)}, \ldots, \psi^{(k)}, \overline{\gamma}_k^{(k)}\right)}^{3K+1-k} \text{ or } \phi_{\overline{\theta}_{3K+1-k}}^{3K+1-k}.$$

Also, for $k = 0, \ldots, K-1$ and $j = k, \ldots, K$, we will use indistinctly the notation

$$\varphi_\theta^{k,j}, \varphi_{\left(\overline{\alpha}_K, \overline{\tau}_K, \psi^{(K)}, \overline{\gamma}_K^{(K)}, \ldots, \psi^{(k)}, \overline{\gamma}_k^{(k)}\right)}^{k,j} \text{ or } \varphi_{\overline{\theta}_{3K+1-j}}^{k,j}.$$

We now focus on fact (II). Lemma 2 below establishes that, for every $P \in \mathcal{F}$ verifying Condition SPob below, the equation in θ, $E_P(\phi_\theta) = 0$, has a unique solution whose last p components are the elements of the vector $\psi(P)$. The proof of that lemma relies on the following Proposition 1. Recall that, as defined in (1.20), for any $\eta \equiv (\eta_1, \ldots, \eta_K)$ and $\pi \equiv (\pi_1, \ldots, \pi_K)$, not just the true ones, and any function d of $\left(\overline{A}_K, Z\right)$

$$U_d(\psi, \eta, \pi) \equiv S_d^K(\eta_K, \pi) + \sum_{k=1}^{K-1} S_d^k(\eta_k, \eta_{k+1}, \pi_1, \ldots, \pi_k) + S_d^0(\psi, \eta_1),$$

where

$$S_d^K(\eta_K, \pi) \equiv \frac{d\left(\overline{A}_K, Z\right)}{\prod_{j=1}^K \pi_j\left(\overline{A}_j, \overline{L}_j\right)} \left\{Y - \eta_K\left(\overline{A}_K, \overline{L}_K\right)\right\},$$

for $k = 1, \ldots, K-1$,

$$S_d^k\left(\eta_k, \eta_{k+1}, \pi_1, \ldots, \pi_k\right) \equiv \sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{d(\overline{A}_k, \underline{a}_{k+1}, Z)}{\prod_{j=1}^k \pi_j\left(\overline{A}_j, \overline{L}_j\right)} \left\{\eta_{k+1}(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1}) - \eta_k(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k)\right\},$$

and

$$S_d^0\left(\psi, \eta_1\right) \equiv \sum_{\underline{a}_1 \in \underline{\mathcal{A}}_1} d\left(\underline{a}_1, Z\right)\left\{\eta_1(\underline{a}_1, L_1) - m(\underline{a}_1, Z; \psi)\right\}.$$

Now, for $k = 1, \ldots, K-1$, for arbitrary η and π, and any function d_k of $\left(\overline{A}_K, \overline{L}_k\right)$, we define

$$\begin{aligned} U_{d_k}^k\left\{\left(\psi, \overline{\gamma}_k, \overline{\tau}_k\right), \underline{\eta}_{k+1}, \widetilde{\pi}_{k+1}\right\} \equiv S_{d_k}^{k,K}\left(\eta_K, \pi_{k+1}, \ldots, \pi_K\right) + \sum_{r=k+1}^{K-1} S_{d_k}^{k,r}\left(\eta_r, \eta_{r+1}, \pi_{k+1}, \ldots, \pi_r\right) \\ + S_{d_k}^{k,k}\left\{\left(\psi, \overline{\gamma}_k, \overline{\tau}_k\right), \eta_{k+1}\right\}, \end{aligned} \tag{1.36}$$

where

$$S_{d_k}^{k,K}\left(\eta_K, \pi_{k+1}, \ldots, \pi_K\right) \equiv \frac{d_k\left(\overline{A}_K, \overline{L}_k\right)}{\prod_{j=k+1}^K \pi_j\left(\overline{A}_j, \overline{L}_j\right)}\left\{Y - \eta_K\left(\overline{A}_K, \overline{L}_K\right)\right\},$$

for $r = k+1, \ldots, K-1$,

$$S_{d_k}^{k,r}\left(\eta_r, \eta_{r+1}, \pi_{k+1}, \ldots, \pi_r\right) \equiv \sum_{\underline{a}_{r+1} \in \underline{\mathbb{A}}_{r+1}} \frac{d_k\left(\overline{A}_r, \underline{a}_{r+1}, \overline{L}_k\right)}{\prod_{j=k+1}^r \pi_j\left(\overline{A}_j, \overline{L}_j\right)}\left\{\eta_{r+1}\left(\overline{A}_r, \underline{a}_{r+1}, \overline{L}_{r+1}\right) - \eta_r\left(\overline{A}_r, \underline{a}_{r+1}, \overline{L}_r\right)\right\},$$

and

$$\begin{aligned} & S_{d_k}^{k,k}\left\{\left(\psi, \overline{\gamma}_k, \overline{\tau}_k\right), \eta_{k+1}\right\} \equiv \\ & \equiv \sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} d_k\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k\right)\left\{\eta_{k+1}\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1}\right) - \eta_k\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \psi, \overline{\gamma}_k, \overline{\tau}_k\right)\right\}, \end{aligned}$$

Here and throughout, $\sum_{r=K}^{K-1} \cdot \equiv 0$.

Proposition 1 states a property of the functions U_d and $U_{d_k}^k, k = 1, \ldots, K$ that is central to understand the multiple robustness of $\widehat{\psi}_{MR}$. Note that, for any $\widetilde{\theta}$ and $\theta \in R^q$,

$$\phi^{3K+1}_{\left(\widetilde{\theta}_{3K}, \theta_{3K+1}\right)} = U_d\left(\psi, \widetilde{\eta}, \widetilde{\pi}\right)$$

with $d\left(\overline{a}_K, z\right) \equiv \dot{m}\left(\overline{a}_K, z; \widetilde{\psi}^{(K)}\right)$, $\widetilde{\eta}_k\left(\overline{a}_K, \overline{l}_k\right) \equiv \eta_k\left(\overline{a}_K, \overline{l}_k; \widetilde{\psi}^{(k)}, \widetilde{\overline{\gamma}}_k^{(k)}, \widetilde{\overline{\tau}}_k\right)$ and $\widetilde{\pi}_k\left(\overline{a}_k, \overline{l}_k\right) \equiv \pi_k\left(\overline{a}_k, \overline{l}_k; \widetilde{\alpha}_k\right)$, $k = 1, \ldots, K$. Likewise, for $k = 1, \ldots, K$,

$$\phi^{3K+1-k}_{\left(\widetilde{\theta}_{3K+1-(k+1)}, \theta_{3K+1-k}\right)} = U_{d_k}^k\left\{\left(\psi^{(k)}, \overline{\gamma}_k^{(k)}, \widetilde{\overline{\tau}}_k\right), \underline{\widetilde{\eta}}_{k+1}, \underline{\widetilde{\pi}}_{k+1}\right\}$$

with $d_k\left(\overline{a}_K, \overline{l}_k\right) \equiv \frac{\dot{\eta}_k\left(\overline{a}_K, \overline{l}_k; \widetilde{\psi}^{(K)}, \widetilde{\overline{\gamma}}_k^{(K)}, \widetilde{\overline{\tau}}_k\right)}{\prod_{s=1}^k \pi_s\left(\overline{A}_s, \overline{L}_s; \widetilde{\alpha}_s\right)}$, $\underline{\widetilde{\eta}}_{k+1} \equiv \left(\widetilde{\eta}_{k+1}, \ldots, \widetilde{\eta}_K\right)$ and $\underline{\widetilde{\pi}}_{k+1} \equiv \left(\widetilde{\pi}_{k+1}, \ldots, \widetilde{\pi}_K\right)$. For this reason, the result in the following proposition is key to prove the Lemma 2 below. Although the theory of Molina et al. [24] implies the following proposition, for completeness, we provide an independent proof of it.

Proposition 1 *Let U_d and $U_{d_k}^k, k = 1, \ldots, K-1$, be the functions defined in (1.20) and (1.36) respectively. Let $\widetilde{\eta} \equiv (\widetilde{\eta}_1, \ldots, \widetilde{\eta}_K)$ and $\widetilde{\pi} \equiv (\widetilde{\pi}_1, \ldots, \widetilde{\pi}_K)$ where, for $k = 1, \ldots, K$, $\widetilde{\eta}_k$ is an arbitrary real-valued function with domain in the sample space of $\left(\overline{A}_K, \overline{L}_k\right)$ and $\widetilde{\pi}_k$ is an arbitrary function with range in $(0,1)$ and domain in the sample space of $\left(\overline{A}_k, \overline{L}_k\right)$. The following holds:*

(a) If $P \in \mathcal{M}$ then

$$E_P\left\{U_d\left(\psi\left(P\right), \widetilde{\eta}, \widetilde{\pi}\right)\right\} = 0$$

for any function d, whenever, for each $k \in \{1, \ldots, K\}$, either $\widetilde{\eta}_k = \eta_k$ or $\widetilde{\pi}_k = \pi_k$.

(b) For $k = 1, \ldots, K-1$, if $P \in \mathcal{R}_k$ then

$$E_P\left[U_{d_k}^k\left\{\left(\psi\left(P\right), \overline{\gamma}_k\left(P\right), \overline{\tau}_k\left(P\right)\right), \underline{\widetilde{\eta}}_{k+1}, \underline{\widetilde{\pi}}_{k+1}\right\}\right] = 0$$

for any function d_k, provided that, for each $j \in \{k+1, \ldots, K\}$, either $\widetilde{\eta}_j = \eta_j$ or $\widetilde{\pi}_j = \pi_j$.

To prove facts (II) and (III) we make the following assumptions.

Condition D For $k = 1, \ldots, K$, $\pi_k\left(\overline{a}_k, \overline{l}_k; \alpha_k\right)$ is differentiable with respect to α_k and there exists $\sigma > 0$ such that $\pi_k\left(\overline{a}_k, \overline{l}_k; \alpha_k\right) > \sigma$ for every $\left(\overline{a}_k, \overline{l}_k\right) \in \overline{\mathcal{A}}_k \times \overline{\mathcal{L}}_k$ and $\alpha_k \in \Delta_k$.

As an example, Condition D is satisfied when, for each $k = 1, \ldots, K$, the support $\mathcal{L}_k$ of L_k and the parameter space Δ_k of α_k are compact, and $\pi_k\left(\cdot, \cdot; \alpha_k\right)$ is a logistic regression model

$$\pi_k\left(\overline{a}_k, \overline{l}_k; \alpha_k\right) \equiv \frac{\exp\left\{\alpha_k' c_k\left(\overline{a}_k, \overline{l}_k\right)\right\}}{1 + \exp\left\{\alpha_k' c_k\left(\overline{a}_k, \overline{l}_k\right)\right\}}$$

with $c_k\left(\cdot, \cdot\right)$ a continuous function.

Condition SPob The following holds:

(i) for $k = 1, \ldots, K$, the equation in $\theta_k\left(= \alpha_k\right)$,

$$E_P\left(\phi_{\theta_k}^k\right) = 0,$$

has a unique solution which we denote indistinctly $\theta_k^\dagger\left(P\right)$ or $\alpha_k^\dagger\left(P\right)$,

(ii) for $k = 1, \ldots, K$, the equation in $\theta_{K+k}\left(= \tau_k\right)$,

$$E_P\left(\phi_{\theta_{K+k}}^{K+k}\right) = 0,$$

has a unique solution which we denote indistinctly $\theta_{K+k}^\dagger\left(P\right)$ or $\tau_k^\dagger\left(P\right)$,

(iii) for $k = 1, \ldots, K$, the equation in $\theta_{3K+1-k}\left(= \left(\psi^{(k)}, \overline{\gamma}_k^{(k)}\right)\right)$,

$$E_P\left\{\phi^{3K+1-k}_{\left(\overline{\theta}_{3K-k}^\dagger(P), \theta_{3K+1-k}\right)}\right\} = 0,$$

has a unique solution which we denote indistinctly $\theta_{3K+1-k}^\dagger\left(P\right)$ or $\left(\psi^{\dagger(k)}\left(P\right), \overline{\gamma}_k^{\dagger(k)}\left(P\right)\right)$ and

(iv) the equation in $\theta_{3K+1}\,(=\psi)$,

$$E_P\left\{\phi^{3K+1}_{\left(\overline{\theta}^{\dagger}_{3K}(P),\theta_{3K+1}\right)}\right\}=0,$$

has a unique solution which we denote indistinctly $\theta^{\dagger}_{3K+1}(P)$ or $\psi^{\dagger}(P)$.

For any P verifying Condition SPob, we define

$$\begin{aligned}\theta^{\dagger}(P) &\equiv \left(\theta^{\dagger}_1(P),\ldots,\theta^{\dagger}_K(P),\theta^{\dagger}_{K+1}(P),\ldots,\theta^{\dagger}_{2K}(P),\theta^{\dagger}_{2K+1}(P),\ldots,\theta^{\dagger}_{3K+1}(P)\right) \\ &= \left(\overline{\alpha}^{\dagger}_K(P),\overline{\tau}^{\dagger}_K(P),\psi^{\dagger(K)}(P),\overline{\gamma}^{\dagger}_K{}^{(K)}(P),\psi^{\dagger(K-1)}(P),\overline{\gamma}^{\dagger}_{K-1}{}^{(K-1)}(P),\ldots\right. \\ &\left.,\psi^{\dagger(1)}(P),\gamma^{\dagger(1)}_1(P),\psi^{\dagger}(P)\right)\end{aligned} \tag{1.37}$$

Lemma 2 *Suppose that $P\in\mathcal{F}$ satisfies Condition SPob and Condition D holds. Then*

(a) the equation in θ, $E_P(\phi_\theta)=0$, has a unique solution at $\theta^{\dagger}(P)$ and

(b) $\theta^{\dagger}_{3K+1}(P)\left(=\psi^{\dagger}(P)\right)$ coincides with $\psi(P)$.

Now, turn to fact (III). By definition, $\widehat{\psi}_{MR}$ is the vector of the last p components of the vector $\widehat{\theta}$ defined in (1.28) and, under the assumptions of Lemma 2, $\psi(P)$ is the vector of the last p components of $\theta^{\dagger}(P)$. Hence, to prove that $\sqrt{n}\left\{\widehat{\psi}_{MR}-\psi(P)\right\}$ converges to a mean zero Normal distribution under $\mathcal{F}$ it suffices to show that $\sqrt{n}\left\{\widehat{\theta}-\theta^{\dagger}(P)\right\}$ does. Now, notice that, under the assumptions of Lemma 1, $\widehat{\theta}$ is a solution of $\mathbb{P}_n(\phi_\theta)=0$ and, under the assumptions of Lemma 2, $\theta^{\dagger}(P)$ is the only solution of $E_P(\phi_\theta)=0$. Thus, under the assumptions of Lemmas 1 and 2, $\widehat{\theta}$ is a Z-estimator of $\theta^{\dagger}(P)$ (see Chapter 5 of [58]). We can now apply Theorems 5.41 and 5.42 of [58] on consistency and asymptotic normality of Z-estimators to our problem. For completeness, in Proposition 3 of Appendix A.3, we present a corollary of those theorems, that contains the results that we will use to derive the consistency and asymptotic normality of $\widehat{\theta}$ under $\mathcal{F}$.

Next, we introduce some notation that will be used in the rest of this section. Consider an open subset $\mathcal{B}$ of an Euclidean space and a random vector X with range in some subset $\mathcal{X}$ of an Euclidean space. Given a collection of functions $\left\{f_\beta(\cdot):\mathcal{X}\to R^N;\beta\in\mathcal{B}\right\}$, for $\beta\in\mathcal{B}$ and $x\in\mathcal{X}$, we denote

$$\dot{f}_\beta(x)\equiv\frac{\partial}{\partial\beta}f_\beta(x)$$

whenever such derivative exists. Also, given $\mathcal{N}\subseteq\mathcal{B}$, we say that $f_\beta(\cdot)$ is dominated by a fixed integrable function in $\mathcal{N}$ iff

$$\|f_\beta(x)\|\leq f(x)\ \text{ for all } \beta\in\mathcal{N} \text{ and } x\in\mathcal{X}$$

for some measurable function $f(\cdot):\mathcal{X}\to R^N$ such that $E\{f(X)\}<\infty$.

Throughout, we make the following assumptions.

Condition S The event $C_n \equiv$"for $k = 1, \ldots, K$, the equations $\mathbb{P}_n\left(\phi^k_{\theta_k}\right) = 0$ and $\mathbb{P}_n\left(\phi^{K+k}_{\theta_{K+k}}\right) = 0$ have at most one solution and, for $k = 0, \ldots, K$, if $\mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widehat{\theta}_{3K+1-(k+1)}}\right) = 0$, then the equation in $\theta_{3K+1-k}, \mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\theta}_{3K+1-(k+1)}, \theta_{3K+1-k}\right)}\right) = 0$, has at most one solution" occurs with probability tending to one under P.

Condition D(ϕ_θ) $\phi_\theta(o)$ is twice continuously differentiable w.r.t. θ for every o.

Also, for P verifying Condition SPob, we assume the following conditions.

Condition Moment2 $E_P\left(\left\|\phi_{\theta^\dagger(P)}\right\|^2\right) < \infty$.

Condition NonSing the matrix $E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)$ exists and is nonsingular.

Condition Domination the second-order partial derivatives of $\phi_\theta(o)$ w.r.t. θ are dominated by a fixed integrable function in a neighborhood of $\theta^\dagger(P)$.

Note that Condition Moment2 is a standard condition on the boundedness on the second order moment of the estimating function. The requirement in Condition D that $\pi_k\left(\overline{a}_k, \overline{l}_k; \alpha_k\right)$ is bounded away form zero is made so as to help ensure that this condition is satisfied.

Lemma 3 *Let $O_1, \ldots, O_n$ be i.i.d. copies of $O \sim P \in \mathcal{F}$ and assume that Conditions D, SPob, S, D(ϕ_θ), Moment2, NonSing and Domination hold. Then,*

$$\sqrt{n}\left\{\widehat{\theta} - \theta^\dagger(P)\right\} = -E_P\left\{\dot{\phi}_{\theta^\dagger(P)}\right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{\theta^\dagger(P)}(O_i) + o_P(1),$$

i.e., $\widehat{\theta}$ is an asymptotically linear estimator of $\theta^\dagger(P)$ with influence function

$$\xi(o) \equiv -E_P\left\{\dot{\phi}_{\theta^\dagger(P)}\right\}^{-1} \phi_{\theta^\dagger(P)}(o).$$

In particular, the sequence $\sqrt{n}\left\{\widehat{\theta} - \theta^\dagger(P)\right\}$ converges to a mean zero Normal distribution with variance $E_P\left\{\dot{\phi}_{\theta^\dagger(P)}\right\}^{-1} E_P\left\{\phi_{\theta^\dagger(P)}\phi'_{\theta^\dagger(P)}\right\}\left[E_P\left\{\dot{\phi}_{\theta^\dagger(P)}\right\}^{-1}\right]'$.

From now on, throughout this section, we index the components of θ indistinctly with $\theta_k, \theta_{K+k}, k = 1, \ldots, K$, and $\theta_{3K+1-k}, k = 0, \ldots, K$, or with θ_s, $s = 1, \ldots, 3K+1$. Likewise, we index the components of ϕ_θ indistinctly with $\phi^k_{\theta_k}, \phi^{K+k}_{\theta_{K+k}}, k = 1, \ldots, K, \phi^{3K+1-k}_{\theta_{3K+1-k}}, k = 0, \ldots, K$, or with $\phi^s_{\theta_s}, s = 1, \ldots, 3K+1$. Also, in the second indexing, for $s = 1, \ldots, 2K$, we write indistinctly $\phi^s_{\overline{\theta}_s}$ or $\phi^s_{\theta_s}$, since $\phi^s_{\overline{\theta}_s}$ depends on $\overline{\theta}_s$ only through θ_s.

Having derived the consistency and asymptotic normality of $\widehat{\psi}_{MR}$ under $\mathcal{F}$, we will now (1) derive its asymptotic variance and (2) find an estimator of it, which is consistent under $\mathcal{F}$. To

derive a formula for the asymptotic variance of $\widehat{\psi}_{MR}$, it suffices to find its influence function. Since $\widehat{\psi}_{MR}$ is the vector of the last p components of $\widehat{\theta}$, its influence function is, under the assumptions of Lemma 3, the vector-valued function of the last p components of the influence function of $\widehat{\theta}$,

$$\xi(o) \equiv -E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)^{-1}\phi_{\theta^\dagger(P)}(o).$$

Therefore, the influence function of $\widehat{\psi}_{MR}$ is equal to $-M\phi_{\theta^\dagger(P)}(o)$, where M is the submatrix composed by the last p rows of $E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)^{-1}$. Since, for $s=1,\ldots,3K+1$, ϕ_θ^s depends on θ at most on $\overline{\theta}_s$, $E_P\left(\dot{\phi}_\theta\right)$ admits the following representation:

$$E_P\left(\dot{\phi}_\theta\right) = \begin{bmatrix} E_P\left(\frac{\partial}{\partial\theta_1}\phi^1_{\overline{\theta}_1}\right) & 0_{d_1\times d_2} & 0_{d_1\times d_3} & \cdots & 0_{d_1\times d_{3K+1}} \\ E_P\left(\frac{\partial}{\partial\theta_1}\phi^2_{\overline{\theta}_2}\right) & E_P\left(\frac{\partial}{\partial\theta_2}\phi^2_{\overline{\theta}_2}\right) & 0_{d_2\times d_3} & \cdots & 0_{d_2\times d_{3K+1}} \\ E_P\left(\frac{\partial}{\partial\theta_1}\phi^3_{\overline{\theta}_3}\right) & E_P\left(\frac{\partial}{\partial\theta_2}\phi^3_{\overline{\theta}_3}\right) & E_P\left(\frac{\partial}{\partial\theta_3}\phi^3_{\overline{\theta}_3}\right) & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0_{d_{3K}\times d_{3K+1}} \\ E_P\left(\frac{\partial}{\partial\theta_1}\phi^{3K+1}_{\overline{\theta}_{3K+1}}\right) & E_P\left(\frac{\partial}{\partial\theta_2}\phi^{3K+1}_{\overline{\theta}_{3K+1}}\right) & E_P\left(\frac{\partial}{\partial\theta_3}\phi^{3K+1}_{\overline{\theta}_{3K+1}}\right) & \cdots & E_P\left(\frac{\partial}{\partial\theta_{3K+1}}\phi^{3K+1}_{\overline{\theta}_{3K+1}}\right) \end{bmatrix},$$

where d_s is the dimension of $\theta_s, s=1,\ldots,3K+1$. That is, $E_P\left(\dot{\phi}_\theta\right)$ is a *lower-triangular-block-matrix* as defined in Subsection A.3.5 of Appendix A.3. Exploiting this structure, in Theorem 1 we recursively derive the influence function of each $\widehat{\theta}_s$ and in particular, we derive the influence of $\widehat{\theta}_{3K+1} = \widehat{\psi}_{MR}$ as the last step of this recursion. Also, as shown in Lemma 12 of that subsection, we can relax the Condition NonSing by making the following assumption.

Condition NonSing2 the matrix $E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)$ exists and, for $s=1,\ldots,3K+1$, $E_P\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\overline{\theta}^\dagger_s(P)}\right\}$ is nonsingular.

To find consistent estimators for the asymptotic variances of the $\widehat{\theta}'_s s$, we also make the following assumption.

Condition M The function $\phi_\theta(o)$ and its first-order partial derivatives w.r.t. θ are measurable w.r.t. o for every θ in a neighborhood of $\theta^\dagger(P)$.

The following result gives, as Lemma 3, sufficient conditions for the consistency and asymptotic normality of $\widehat{\theta}$ for $\theta^\dagger(P)$ under $\mathcal{F}$, and also provides a recursive formula to compute the influence function of each $\widehat{\theta}_s, s=1,\ldots,3K+1$. Moreover, it provides consistent estimators for the asymptotic variances of the $\widehat{\theta}'_s s$, which are computed in a recursive way.

Theorem 1 *Let $O_1,\ldots,O_n$ be i.i.d. copies of $O\sim P\in\mathcal{F}$ and assume that Conditions D, SPob, S, D(ϕ_θ), Moment2, Domination, NonSing2 and M hold. Then, for $s=1,\ldots,3K+1$,*

(a) $\widehat{\theta}_s$ *is an asymptotically linear estimator of* $\theta_s^\dagger(P)$ *with influence function* $\xi_s(o)$ *where, for* $s = 1, \ldots, 2K$,

$$\xi_s(o) \equiv -\left[E_P\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\theta s}\right)\Big|_{\theta_s=\theta_s^\dagger(P)}\right\}\right]^{-1}\phi^s_{\theta_s^\dagger(P)}(o),$$

and, for $s = 2K+1, \ldots, 3K+1$,

$$\xi_s(o) \equiv -\left[E_P\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\overline{\theta}_s^\dagger(P)}\right\}\right]^{-1}\left[\phi^s_{\overline{\theta}_s^\dagger(P)}(o) + \sum_{j=1}^{s-1} E_P\left\{\left(\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\overline{\theta}_s^\dagger(P)}\right\}\xi_j(o)\right].$$

(b) $\widehat{AVar}\left(\widehat{\theta}_s\right) \equiv \mathbb{P}_n\left(\widehat{\xi}_s\widehat{\xi}_s^T\right)$ *converges in probability, under* P, *to* $AVar\left(\widehat{\theta}_s\right) \equiv E_P\left(\xi_s\xi_s^T\right)$ *where, for* $s = 1, \ldots, 2K$,

$$\widehat{\xi}_s(o) \equiv -\left[\mathbb{P}_n\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\theta_s}\right)\Big|_{\theta_s=\widehat{\theta}_s}\right\}\right]^{-1}\phi^s_{\widehat{\theta}_s}(o) \tag{1.38}$$

which is well defined, i.e. $\mathbb{P}_n\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\theta_s}\right)\Big|_{\theta_s=\widehat{\theta}_s}\right\}$ *is non-singular, with probability going to 1 and, for* $s = 2K+1, \ldots, 3K+1$,

$$\widehat{\xi}_s(o) \equiv -\left[\mathbb{P}_n\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\widehat{\overline{\theta}}_s}\right\}\right]^{-1}\left[\phi^s_{\widehat{\overline{\theta}}_s}(o) + \sum_{j=1}^{s-1}\mathbb{P}_n\left\{\left(\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\widehat{\overline{\theta}}_s}\right\}\widehat{\xi}_j(o)\right] \tag{1.39}$$

which is well defined, i.e. $\mathbb{P}_n\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\theta_s}\right)\Big|_{\theta_s=\widehat{\theta}_s}\right\}$ *is non-singular, with probability going to 1.*

In particular, $\widehat{\psi}_{MR}$ *is an asymptotically linear estimator of* $\psi(P)$ *with influence function* $\xi_{3K+1}(o)$ *and* $\widehat{AVar}\left(\widehat{\psi}_{MR}\right) \equiv \mathbb{P}_n\left(\widehat{\xi}_{3K+1}\widehat{\xi}_{3K+1}^T\right)$ *converges in probability, under* P, *to* $AVar\left(\widehat{\psi}_{MR}\right) \equiv E_P\left(\xi_{3K+1}\xi_{3K+1}^T\right)$.

The following lemma establishes primitive conditions on the functions m, g_k, t_k, e_k, π_k and q_k under which the estimating function ϕ_θ satisfies Conditions D(ϕ_θ) and M. This lemma follows straightforwardly by the analytic expressions of $\phi_\theta(o)$ and its first and second-order partial derivatives w.r.t. θ.

Lemma 4 *Assume that*

(i) $m(\overline{a}_K, z; \psi)$ *is three times continuously differentiable w.r.t.* ψ *for every* $(\overline{a}_K, z)$ *. Also,* $m(\overline{a}_K, z; \psi)$ *and its first and second-order partial derivatives w.r.t.* ψ *are measurable w.r.t.* $(\overline{a}_K, z)$,

(ii) $g_k\left(\overline{a}_K, \overline{l}_{k-1}; \gamma_k\right)$ *is three times continuously differentiable w.r.t.* γ_k *for every* $\left(\overline{a}_K, \overline{l}_{k-1}\right)$, $k = 1, \ldots, K$. *Also,* $g_k\left(\overline{a}_K, \overline{l}_{k-1}; \gamma_k\right)$ *and its first and second-order partial derivatives w.r.t.* γ_k *are measurable w.r.t.* $\left(\overline{a}_K, \overline{l}_{k-1}\right), k = 1, \ldots, K$.

(iii) $e_1(z;\tau_1)$ *is twice continuously differentiable w.r.t.* τ_1 *for every* z *and, for* $k=2,\ldots,K$, $e_k\left(\overline{a}_{k-1},\overline{l}_{k-1};\tau_k\right)$ *is twice continuously differentiable w.r.t.* τ_k *for every* $\left(\overline{a}_{k-1},\overline{l}_{k-1}\right)$. *Also,* $e_1(z;\tau_1)$ *and its first-order partial derivatives w.r.t.* τ_1 *are measurable w.r.t.* z *and, for* $k=2,\ldots,K$, $e_k\left(\overline{a}_{k-1},\overline{l}_{k-1};\tau_k\right)$ *and its first-order partial derivatives w.r.t.* τ_k *are measurable w.r.t.* $\left(\overline{a}_{k-1},\overline{l}_{k-1}\right)$.

(iv) $\pi_k\left(\overline{a}_k,\overline{l}_k;\alpha_k\right)$ *are positive for every* $\left(\overline{a}_k,\overline{l}_k;\alpha_k\right)$ *and three times continuously differentiable w.r.t.* α_k *for every* $\left(\overline{a}_k,\overline{l}_k\right), k=1,\ldots,K$. *Also,* $\pi_k\left(\overline{a}_k,\overline{l}_k;\alpha_k\right)$ *and its first and second-order partial derivatives w.r.t.* α_k *are measurable w.r.t.* $\left(\overline{a}_k,\overline{l}_k\right), k=1,\ldots,K$.

(v) $t_1(v)$ *and* $t_k(l_k)$ *are measurable w.r.t.* v *and* l_k *respectively,* $k=2,\ldots,K$.

(vi) $q_1(z)$ *and* $q_k\left(\overline{a}_{k-1},\overline{l}_{k-1}\right)$ *are measurable w.r.t.* z *and* $\left(\overline{a}_{k-1},\overline{l}_{k-1}\right)$ *respectively,* $k=2,\ldots,K$.

Then,

(a) $\phi_\theta(o)$ *is twice continuously differentiable w.r.t.* θ *for every* o *(i.e. Condition* $D(\phi_\theta)$ *holds) and*

(b) $\phi_\theta(o)$ *and its first-order partial derivatives w.r.t.* θ *are measurable w.r.t.* o *for every* $\theta\in R^q$ *and, hence, Condition M holds.*

1.10.1 Consistency and asymptotic normality under linearity

When $m\left(\overline{a}_K,z;\psi\right)$, $g_1\left(\overline{a}_K,z;\gamma_1\right)$ or any $g_k\left(\overline{a}_K,\overline{l}_{k-1};\gamma_k\right), k=2,\ldots,K$, is a non linear function of ψ,γ_1 or γ_k respectively, parts (iii) and (iv) of Condition SPob are conditions for the existence of a unique solution of a system of non-linear equations. As such the condition will need to be verified on a case by case basis. A similar situation occurs with Condition S. On the other hand, when m and the $g_k's$ are linear in the parameters, the moment equations defining $\overline{\theta}_{2K+1}^{3K+1,\dagger}(P)\equiv\left(\psi^{\dagger(K)}(P),\overline{\gamma}_K^{\dagger(K)}(P),\psi^{\dagger(K-1)}(P),\overline{\gamma}_{K-1}^{\dagger(K-1)}(P),\ldots,\psi^{\dagger(1)}(P),\gamma_1^{\dagger(1)}(P),\psi^{\dagger}(P)\right)$ make up a system of linear equations with equal number of equations as number of unknowns. In such case, condition SPob reduces to the condition that the matrix defining the system is non-singular. Since the matrix for this system is a lower-triangular-block-matrix, the requirement of non-singularity reduces to the requirement of the non-singularity of the diagonal block matrices. Condition SPobLin describes the diagonal block matrices. A similar situation occurs for the conditions on the uniqueness of the solutions to the empirical version of the population moment equations defining $\overline{\theta}_{2K+1}^{3K+1,\dagger}(P)$ when the remaining unknown parameters are replaced by their estimators. In Lemma 6 we provide conditions that ensure that the matrix of the linear system is invertible with probability going to 1. Notice that because we have assumed that the outcome Y is unconstrained, it is reasonable to postulate models for the $\eta_k's$ in which the functions m and $g_k's$ are linear in the parameters. In what follows we formalize this argument by giving rigorous conditions for the validity of Conditions SPob and S when m and the $g_k's$ are linear in the parameters, i.e.

$$m\left(\overline{a}_K,z;\psi\right)=\mathbf{m}\left(\overline{a}_K,z\right)'\psi, \tag{1.40}$$

$$g_1\left(\overline{a}_K,z;\gamma_1\right)=\mathbf{g}_1\left(\overline{a}_K,z\right)'\gamma_1, \tag{1.41}$$

and, for $k=2,\ldots,K$,

$$g_k\left(\overline{a}_K,\overline{l}_{k-1};\gamma_k\right)=\mathbf{g}_k\left(\overline{a}_K,\overline{l}_{k-1}\right)'\gamma_k, \tag{1.42}$$

for some conformable vector-valued function $\mathbf{m}(\cdot,\cdot)$ and some conformable matrix-valued functions $\mathbf{g}_1(\cdot,\cdot)$ and $\mathbf{g}_k(\cdot,\cdot), k=2,\ldots,K$.

The fact that, under (1.40),(1.41) and (1.42), the moment equations defining $\overline{\theta}_{2K+1}^{3K+1,\dagger}(P)$ make up a system of linear equations, follows from the fact, under this setting, each $\eta_k\left(\overline{a}_K,\overline{l}_k;\psi,\overline{\gamma}_k,\overline{\tau}_k\right)$ is linear in $(\psi,\overline{\gamma}_k), k=1,\ldots,K,$. To see this, for $k=1,\ldots,K$, define

$$H_k\left(\overline{a}_K,\overline{l}_k;\overline{\tau}_k\right)\equiv\begin{pmatrix}\mathbf{m}\left(\overline{a}_K,z\right)\\ \mathbf{g}_1\left(\overline{a}_K,z\right)\Delta_1\left(l_1;\tau_1\right)\\ \vdots\\ \mathbf{g}_k\left(\overline{a}_K,\overline{l}_{k-1}\right)\Delta_k\left(\overline{a}_{k-1},\overline{l}_k;\tau_k\right)\end{pmatrix} \tag{1.43}$$

with

$$\Delta_1\left(l_1;\tau_1\right)\equiv t_1\left(v\right)-e_1\left(z;\tau_1\right)$$

and

$$\Delta_k\left(\overline{a}_{k-1},\overline{l}_k;\tau_k\right)\equiv t_k\left(l_k\right)-e_k(\overline{a}_{k-1},\overline{l}_{k-1};\tau_k),$$

$k=2,\ldots,K$. Then,

$$\eta_k\left(\overline{a}_K,\overline{l}_k;\psi,\overline{\gamma}_k,\overline{\tau}_k\right)=H_k\left(\overline{a}_K,\overline{l}_k;\overline{\tau}_k\right)'\begin{pmatrix}\psi\\ \gamma_1\\ \vdots\\ \gamma_k\end{pmatrix}. \tag{1.44}$$

Throughout this section, we will write indistinctly $\phi_{\theta_k}^k$, $\phi_{\theta_{K+k}}^{K+k}$, $\phi_{\overline{\theta}_{3K+1-k}}^{3K+1-k}, k=1,\ldots,K$, and $\phi_{\overline{\theta}_{3K+1}}^{3K+1}$ or $\phi_{\alpha_k}^k$, $\phi_{\tau_k}^{K+k}$, $\phi_{\left(\overline{\alpha}_K,\overline{\tau}_K,\psi^{(K)},\overline{\gamma}_K^{(K)},\ldots,\psi^{(k+1)},\overline{\gamma}_{k+1}^{(k+1)},\psi^{(k)},\overline{\gamma}_k^{(k)}\right)}^{3K+1-k}, k=1,\ldots,K$, and $\phi_{\left(\overline{\alpha}_K,\overline{\tau}_K,\psi^{(K)},\overline{\gamma}_K^{(K)},\ldots,\psi^{(1)},\gamma_1^{(1)},\psi\right)}^{3K+1}$.

The following lemma gives primitive conditions, under the linearity of m and the $g_k's$ for parts (iii) and (iv) of Condition SPob, assuming that parts (i) and (ii) of that condition hold. Part (i) of Condition SPob is satisfied when, for each $k=1,\ldots,K$, the expectation of the estimating function used to estimate α_k has a unique solution. In the case where the model for π_k is correct, this is verified when α_k is identified. For badly specified π_k models, this condition must be verified on a case-by-case basis. A similar situation occurs for $\tau_k, k=1,\ldots,K$, in part (ii) of Condition SPob.

In what follows, $\mathbf{m}\left(\overline{a}_K,z\right)$, $\mathbf{g}_1\left(\overline{a}_K,z\right)$ and $\mathbf{g}_k\left(\overline{a}_K,\overline{l}_{k-1}\right), k=2,\ldots,K$, are the functions involved in equations (1.40), (1.41) and (1.42) respectively. Likewise, $H_k\left(\overline{a}_K,\overline{l}_k;\overline{\tau}_k\right)$ are the functions defined in (1.43), $k=1,\ldots,K$.

Throughout, we make the following assumption.

Condition SPobLin P verifies parts (i) and (ii) of Condition SPob and the matrices
$E_P\left\{\frac{H_K\left(\overline{A}_K,\overline{L}_K;\overline{\tau}_K^{\dagger}(P)\right)H_K\left(\overline{A}_K,\overline{L}_K;\overline{\tau}_K^{\dagger}(P)\right)'}{\prod_{s=1}^K\pi_s\left(\overline{A}_s,\overline{L}_s;\alpha_s^{\dagger}(P)\right)}\right\}$,
$E_P\left\{\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}}\frac{H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k^{\dagger}(P)\right)H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k^{\dagger}(P)\right)'}{\prod_{s=1}^k\pi_s\left(\overline{A}_s,\overline{L}_s;\alpha_s^{\dagger}(P)\right)}\right\}, k=1,\ldots,K-1$, and
$E_P\left\{\sum_{\underline{a}_1\in\underline{\mathcal{A}}_1}\mathbf{m}\left(\underline{a}_1,Z\right)\mathbf{m}\left(\underline{a}_1,Z\right)'\right\}$ are nonsingular.

Lemma 5 *Suppose that the functions $m(\cdot,\cdot;\psi)$ and $g_k(\cdot,\cdot;\gamma_k)$, $k=1,\ldots,K$, verify equations (1.40)-(1.42) and that P verifies Condition SPobLin. Then, P verifies Condition SPob.*

The following lemma establishes primitive conditions for Condition S. These conditions include the following assumptions.

Condition SLin The event $B_n \equiv$"for $k=1,\ldots,K$, the equations $\mathbb{P}_n\left(\phi^k_{\theta_k}\right)=0$ and $\mathbb{P}_n\left(\phi^{K+k}_{\theta_{K+k}}\right)=0$ have at most one solution" occurs with probability tending to one under P.

Condition R The function $\frac{H_K\left(\overline{a}_K,\overline{l}_K;\overline{\tau}_K\right)H_K\left(\overline{a}_K,\overline{l}_K;\overline{\tau}_K\right)'}{\prod_{s=1}^K \pi_s\left(\overline{a}_s,\overline{l}_s;\alpha_s\right)}$ is regular according to Definition 2 of Appendix A.4 where $\left(\overline{a}_K,\overline{l}_K\right)$ plays the roll of x and $\left(\overline{\alpha}_K,\overline{\tau}_K\right)$ plays the roll of β and, for $k=1,\ldots,K-1$, the function $\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}} \frac{H_k\left(\overline{a}_k,\underline{a}_{k+1},\overline{l}_k;\overline{\tau}_k\right)H_k\left(\overline{a}_k,\underline{a}_{k+1},\overline{l}_k;\overline{\tau}_k\right)'}{\prod_{s=1}^k \pi_s\left(\overline{a}_s,\overline{l}_s;\alpha_s\right)}$ is regular according to Definition 2 of Appendix A.4 where $\left(\overline{a}_k,\overline{l}_k\right)$ plays the roll of x and $\left(\overline{\alpha}_k,\overline{\tau}_k\right)$ plays the roll of β.

Lemma 6 *Suppose that the functions $m(\cdot,\cdot;\psi)$ and $g_k(\cdot,\cdot;\gamma_k)$, $k=1,\ldots,K$, verify equations (1.40)-(1.42). Let $O_1,\ldots,O_n$ be i.i.d. copies of $O\sim P$. Assume that Conditions Moment2, Domination, NonSing2, SPobLin, SLin and R hold. Also assume that*

(i) $\pi_k\left(\overline{a}_k,\overline{l}_k;\alpha_k\right)$ is three times continuously differentiable w.r.t. α_k for every $\left(\overline{a}_k,\overline{l}_k\right)$, $k=1,\ldots,K$, and

(ii) $e_1(z;\tau_1)$ is twice continuously differentiable w.r.t. τ_1 for every z and, for $k=2,\ldots,K$, $e_k\left(\overline{a}_{k-1},\overline{l}_{k-1};\tau_k\right)$ is twice continuously differentiable w.r.t. τ_k for every $\left(\overline{a}_{k-1},\overline{l}_{k-1}\right)$.

Then, Condition S holds.

Finally, note that if the functions $m(\cdot,\cdot;\psi)$ and $g_k(\cdot,\cdot;\gamma_k)'s$ verify equations (1.40) – (1.42) with $\mathbf{m}$ and $\mathbf{g}_k, k=1,\ldots,K$, measurable functions, then assumptions (i) and (ii) of Lemma 4 are satisfied.

1.11 MR estimation for repeated outcomes

1.11.1 Marginal structural mean model for repeated and unconstrained outcomes

In this section, we provide a discussion of what it takes to generalize the proposal of R, DR and MR estimation of Section 1.7, to the case of a MSMM for repeated and unconstrained outcomes; that is, when rather than being interested in a single outcome measured at the end of a longitudinal study, we are also interested in outcomes which correspond to a specific component of the vector L_k measured at each occasion t_k^-. To formalize the inferential problem, suppose as in earlier sections that the observed data are n i.i.d. copies of

$$O \equiv (L_1, A_1, \ldots, L_K, A_K, Y_{K+1}),$$

where Y_{K+1} is an outcome of interest at time t_{K+1}, which is unconstrained. Also, for $k = 1, \ldots, K$, A_k is the treatment given at time t_k taking values in a finite set $\mathcal{A}_k$ ($t_{k-1} < t_k$). As in Section 1.7, L_1 is a vector of covariates, measured at time t_1^-, that we write as $L_1 = (Z, V_1)$. For each $k = 2, \ldots, K$, we now decompose

$$L_k = (Y_k, V_k)$$

where Y_k is an outcome of interest and V_k is a vector of covariates, both measured at time t_k^-, i.e. an instant prior to t_k .

Analogously to the case of a single outcome, for each $k = 1, \ldots, K$, we define the counterfactual variable $Y_{k+1,\overline{a}_k}$ to be the subject's response at time t_k^- if, possibly contrary to fact, treatment regime $\overline{a}_k$ is followed up to that time point. We make the identifying assumptions of

(1) consistency:

$$Y_{k+1,\overline{a}_k} = Y_{k+1} \text{ if } \overline{A}_k = \overline{a}_k,$$

$k = 1, \ldots, K$,

(2) no unmeasured confounding (NUC): for all $\overline{a}_k$,

$$Y_{k+1,\overline{a}_k} \amalg A_j | \overline{L}_j, \overline{A}_{j-1} = \overline{a}_{j-1},$$

$1 \leq j \leq k, 1 \leq k \leq K$, and

(3) positivity: for all k and $\overline{a}_k$, if $f(\overline{a}_{k-1}, \overline{l}_k) > 0$ then $f(a_k | \overline{a}_{k-1}, \overline{l}_k) > 0$.

A straightforward extension of the argument used for a single outcome implies that under the identifying assumptions, the parameter $\psi^* \equiv \left(\psi^{2,*}, \ldots, \psi^{K+1,*}\right) \in R^{p_2} \times \cdots \times R^{p_{K+1}}$ of the MSMM defined by the restrictions

$$E\left(Y_{k+1,\overline{a}_k} | Z\right) = m^{k+1}\left(\overline{a}_k, Z; \psi^{k+1,*}\right) \text{ for all } \overline{a}_k \tag{1.45}$$

where $m^{k+1}(\cdot, \cdot; \cdot)$ is specified for all $k = 1, \ldots, K$.

In this section we will discuss a number of non-trivial subtle issues that arise in extending the methods described in earlier sections for estimating the parameters of MSMM for a single outcome to the case of parameters of MSMM for repeated outcomes. At first sight, the inferential problem appears to be a trivial extension. However, this is not the case for the following reason. Suppose we

were to estimate each $\psi^{k+1,*}$ separately, regarding each time t_{k+1} as if it were the end of the study, i.e. disregarding the data measured after time t_{k+1} and regarding Y_{k+1} as the sole outcome of interest. For estimating a single $\psi^{k+1,*}$ we would specify and estimate working nested compatible models following the steps described in earlier sections. However, if we wish our models to be compatible across all times $t_k, k = 1, ..., K$, we will not be free to specify the components of the nested models for each t_k freely, essentially because the outcome Y_k at a given time t_k becomes a component of the covariate vector L_k when we estimate the parameters of the marginal structural mean model for a future outcome Y_{k+j} for $j \geq 1$. This implies that we need to take extra care in the formulation of our nested models and in the procedure we use to estimate the model parameters. In the next subsections we elaborate on these points.

1.11.2 Compatible parametric working models for the $\eta_j^{k+1}s$

For each $k = 1, \ldots, K$, let

$$\eta_k^{k+1}\left(\overline{a}_k, \overline{l}_k\right) \equiv E\left(Y_{k+1} | \overline{A}_k = \overline{a}_k, \overline{L}_k = \overline{l}_k\right)$$

and, for $j = 1, \ldots, k-1$, let

$$\eta_j^{k+1}\left(\overline{a}_k, \overline{l}_j\right) \equiv E\left\{\eta_{j+1}^{k+1}\left(\overline{a}_k, \overline{L}_{j+1}\right) | \overline{A}_j = \overline{a}_j, \overline{L}_j = \overline{l}_j\right\}.$$

Also let

$$\eta_0^{k+1}\left(\overline{a}_k, z\right) = E\left\{\eta_1^{k+1}\left(\overline{a}_k, L_1\right) | Z = z\right\},$$

$k = 1, \ldots, K$. Since for each $k \in \{1, \ldots, K\}$, (1.45) is a MSMM for a single outcome, then, under the identifying assumptions,

$$\eta_j^{k+1}\left(\overline{a}_k, \overline{L}_j\right) = E\left(Y_{k+1,\overline{a}_k} | \overline{A}_j = \overline{a}_j, \overline{L}_j\right),$$

$j = 1, \ldots, k$, and

$$\eta_0^{k+1}\left(\overline{a}_k, Z\right) = E\left(Y_{k+1,\overline{a}_k} | Z\right).$$

Hence, under these assumptions, given $k \in \{1, \ldots, K\}$, (1.45) is equivalent to a model for the observed data O_k defined by the sole restriction

$$\eta_0^{k+1}\left(\overline{a}_k, Z\right) = m^{k+1}\left(\overline{a}_k, Z; \psi^{k+1,*}\right) \text{ for all } \overline{a}_k. \tag{1.46}$$

To arrive at compatible, shared parameter, models for all the $\eta_j^{k+1\prime}s$, we cannot merely generalize the proposal of Section 1.7.1. To see this, we first define the functionals $\rho_j^{k+1\prime}s$ analogous to the $\rho_j's$ of that section. That is, for $k = 1, \ldots, K$, let

$$\rho_1^{k+1}\left(\overline{a}_k, l_1\right) \equiv \eta_1^{k+1}\left(\overline{a}_k, l_1\right) - \eta_1^{k+1}\left(\overline{a}_k, z, v_1 = v_{1,0}\right)$$

and, for $k = 2, \ldots, K$ and $j = 2, \ldots, k$, let

$$\rho_j^{k+1}\left(\overline{a}_k, \overline{l}_j\right) \equiv \eta_j^{k+1}\left(\overline{a}_k, \overline{l}_j\right) - \eta_j^{k+1}\left(\overline{a}_k, \overline{l}_{j-1}, l_j = l_{j,0}\right).$$

where $v_{1,0}$ and $l_{j,0}$ are any baseline levels of V_1 and $L_j, j = 2, \ldots, K$, respectively and $l_1 \equiv (z, v_1)$. For each $k = 1, \ldots, K$, the same arguments as those of Section 1.7.1 yield

$$\eta_1^{k+1}(\bar{a}_k, l_1) = \eta_0^{k+1}(\bar{a}_k, z) + \rho_1^{k+1}(\bar{a}_k, l_1) - E\left\{\rho_1^{k+1}(\bar{a}_k, L_1)\,|Z = z\right\} \tag{1.47}$$

and, for $s = 2, \ldots, K$, and $k = s, \ldots, K$,

$$\begin{aligned}\eta_s^{k+1}\left(\bar{a}_k, \bar{l}_s\right) = {} & \eta_0^{k+1}(\bar{a}_k, z) + \rho_1^{k+1}(\bar{a}_k, l_1) - E\left\{\rho_1^{k+1}(\bar{a}_k, L_1)\,|Z = z\right\} \\ & + \textstyle\sum_{j=2}^{s}\left[\rho_j^{k+1}\left(\bar{a}_k, \bar{l}_j\right) - E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1} = \bar{a}_{j-1}, \overline{L}_{j-1} = \bar{l}_{j-1}\right\}\right].\end{aligned} \tag{1.48}$$

As in Section 1.7.1, these identities imply that parametric models for the $\rho_j^{k+1\prime}s$, for $E\left\{\rho_1^{k+1}(\bar{a}_k, L_1)\,|Z\right\}$ and for the $E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}'s$, together with the MSMM for the observed data defined by the restrictions $\eta_0^{k+1}(\bar{a}_k, Z) = m^{k+1}\left(\bar{a}_k, Z; \psi^{k+1,*}\right), k = 1, \ldots, K+1$, determine parametric models for the $\eta_j^{k+1\prime}s$. However, now one is not free to choose arbitrary models for the $\rho_j^{k+1\prime}s$, for $E\left\{\rho_1^{k+1}(\bar{a}_k, L_1)\,|Z\right\}$ and for the $E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}'s$. This is because the $\eta_0^{k+1\prime}s$, the $\rho_j^{k+1\prime}s$, $f(V_1|Z)$ and the $f\left(L_j|\overline{L}_{j-1}, \overline{A}_{j-1}\right)'s$ are no longer variation independent. To see why, notice that, for each $k = 2, \ldots, K$,

1. η_k^{k+1} is determined by η_0^{k+1}, ρ_j^{k+1}, $j = 1, \ldots, k$, $E\left\{\rho_1^{k+1}(\bar{a}_k, L_1)\,|Z\right\}$ and $E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$, $j = 2, \ldots, k$,

2. for each $j = 2, \ldots, k, \eta_{j-1}^{j}\left(\overline{A}_{j-1}, \overline{L}_{j-1}\right) = E\left(Y_j|\overline{A}_{j-1}, \overline{L}_{j-1}\right)$, which is not variation independent with $E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$ since both depend on $f\left(L_j|\overline{L}_{j-1}, \overline{A}_{j-1}\right)$, and

3. for each $j = 2, \ldots, k, \eta_{j-1}^{j}$ is determined by η_0^j, $\rho_s^j, s = 1, \ldots, j-1$, $E\left\{\rho_1^j(\bar{a}_{j-1}, L_1)\,|Z\right\}$ and $E\left\{\rho_s^j\left(\bar{a}_{j-1}, \overline{L}_s\right)|\overline{A}_{s-1}, \overline{L}_{s-1}\right\}$, $s = 2, \ldots, j-1$.

Thus, for $k = 2, \ldots, K$ and $j = 2, \ldots, k, \eta_0^j$, $\rho_s^j, s = 1, \ldots, j-1$, $f(V_1|Z)$ and $f\left(L_s|\overline{A}_{s-1}, \overline{L}_{s-1}\right)$, $s = 2, \ldots, j-1$ are not variation independent with $\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)$ and $f\left(L_j|\overline{A}_{j-1}, \overline{L}_{j-1}\right)$ since, both $\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)$ and $f\left(L_j|\overline{A}_{j-1}, \overline{L}_{j-1}\right)$ determine $E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$.

We can nevertheless derive compatible models for the $\eta_s^{k+1\prime}s$ from parametric models for the $\rho_j^{k+1\prime}s$, $E\left\{\rho_1^{k+1}(\bar{a}_k, L_1)\,|Z\right\}$, the $E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}'s$ and the MSMMs for the observed data (1.46), exploiting the relationships (1.47) and (1.48), as in Section 1.7.1. The distinction is that now, for each $k = 2, \ldots, K$ and $j = 2, \ldots, k$, we must carefully choose the working model for $E\left\{\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$ so as to be compatible with the model for η_{j-1}^j implied by the assumed models for η_0^j, $\rho_s^j, s = 1, \ldots, j-1$, $E\left\{\rho_1^j(\bar{a}_{j-1}, L_1)\,|Z\right\}$ and $E\left\{\rho_s^j\left(\bar{a}_{j-1}, \overline{L}_s\right)|\overline{A}_{s-1}, \overline{L}_{s-1}\right\}$, $s = 2, \ldots, j-1$.

Analogously to the case of one outcome, we propose the following models the $\rho_j^{k+1\prime}s$. For $k = 1, \ldots, K$, we assume

$$\rho_1^{k+1}(\bar{a}_k, L_1) = g_1^{k+1}(\bar{a}_k, Z; \gamma_1^{k+1,*})' t_1(V_1), \tag{1.49}$$

and, for $k = 2, \ldots, K$ and $j = 2, \ldots, k$ we assume

$$\rho_j^{k+1}\left(\bar{a}_k, \overline{L}_j\right) = g_j^{k+1}(\bar{a}_k, \overline{L}_{j-1}; \gamma_j^{k+1,*})' t_j(L_j), \tag{1.50}$$

where, for each $k = 1, \ldots, K$ and $j = 1, \ldots, k$, g_j^{k+1} is a user specified vector-valued function and $\gamma_j^{k+1,*}$ is a finite dimensional parameter. Also, the $t_j's$ are user-specified conformable vector-valued functions verifying $t_1(v_{1,0}) = 0$ and $t_j(l_{j,0}) = 0, j = 2, \ldots, K$, so that the definitional restrictions $\rho_1^{k+1}(\overline{a}_k, z, v_1 = v_{1,0}) = 0, k = 1, \ldots, K$, and $\rho_j^{k+1}(\overline{a}_k, \overline{l}_{j-1}, l_j = l_{j,0}) = 0, k = 2, \ldots, K$, $j = 2, \ldots, k$, are respected.

Under models (1.49) and (1.50), in order to specify models for $\left\{\rho_1^{k+1}(\overline{a}_k, L_1)|Z\right\}$ and $E\left\{\rho_j^{k+1}(\overline{a}_k, \overline{L}_j)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$ it suffices to specify parametric models for $E\{t_1(V_1)|Z\}$ and $E\left\{t_j(L_j)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$. We assume

$$E\{t_1(V_1)|Z\} = e_1(Z; \tau_1^*), \tag{1.51}$$

where e_1 is any user-specified conformable vector-valued function and τ_1^* is a finite dimensional parameter.

Since, for each $j = 2, \ldots, K$, the model for $E\left\{t_j(L_j)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$ must be compatible with the one proposed for η_{j-1}^j and, at the same time, the models for the $\eta_j^{k+1}, k = j, \ldots, K$, are derived from the postulated models for $E\left\{t_s(L_s)|\overline{A}_{s-1}, \overline{L}_{s-1}\right\}, s = 1, \ldots, j$, we propose specifying the models for $E\left\{t_j(L_j)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\}$ in ascending order, together with the models for the $\eta_j^{k+1\prime}s$ as follows.

1. For $k = 1, \ldots, K$, let

$$\begin{aligned}\eta_1^{k+1}(\overline{a}_k, L_1) &= \eta_1^{k+1}\left(\overline{a}_k, L_1; \psi^{k+1,*}, \gamma_1^{k+1,*}, \tau_1^*\right) \\ &\equiv m^{k+1}(\overline{a}_k, Z; \psi^{k+1}) + g_1^{k+1}(\overline{a}_k, Z; \gamma_1^{k+1,*})'\{t_1(V_1) - e_1(Z; \tau_1^*)\}.\end{aligned} \tag{1.52}$$

2. For $j = 2, \ldots, K$,

   (i) let

$$E\left\{t_j(L_j)|\overline{A}_{j-1}, \overline{L}_{j-1}\right\} = e_j\left(\overline{A}_{j-1}, \overline{L}_{j-1}; \overline{\psi}^{j,*}, \overline{\gamma}^{j,*}, \overline{\tau}_j^*\right) \tag{1.53}$$

   where e_j is a user-specified conformable vector-valued function satisfying certain conditions described below and $\overline{\tau}_j^* \equiv (\tau_1^*, \ldots, \tau_j^*)$ is a finite dimensional parameter. Here, in an abuse of notation, for $j = 2, \ldots, K$,we write

$$\overline{\psi}^{j,*} \equiv (\psi^{2,*}, \ldots, \psi^{j,*})$$

   and

$$\overline{\gamma}^{j,*} \equiv (\gamma^{2,*}, \ldots, \gamma^{j,*})$$

   where

$$\gamma^j \equiv \left(\gamma_1^j, \ldots, \gamma_{j-1}^j\right).$$

   (ii) for $k = j, \ldots, K$, let

$$\begin{aligned}\eta_j^{k+1}(\overline{a}_k, \overline{L}_j) &= \eta_j^{k+1}\left(\overline{a}_k, \overline{L}_j; \psi^{k+1,*}, \overline{\gamma}_j^{k+1,*}, \overline{\psi}^{j,*}, \overline{\gamma}^{j,*}, \overline{\tau}_j^*\right) \\ &\equiv \eta_{j-1}^{k+1}\left(\overline{a}_k, \overline{L}_{j-1}; \psi^{k+1,*}, \overline{\gamma}_{j-1}^{k+1,*}, \overline{\psi}^{j-1,*}, \overline{\gamma}^{j-1,*}, \overline{\tau}_{j-1}^*\right) \\ &+ g_j^{k+1}(\overline{a}_k, \overline{L}_{j-1}; \gamma_j^{k+1})'\left\{t_j(L_j) - e_j\left(\overline{a}_{j-1}, \overline{L}_{j-1}; \overline{\psi}^{j,*}, \overline{\gamma}^{j,*}, \overline{\tau}_j^*\right)\right\}.\end{aligned} \tag{1.54}$$

Here, for each $k=1,\ldots,K$ and $j=1,\ldots,k$,

$$\overline{\gamma}_j^{k+1,*} \equiv \left(\gamma_1^{k+1,*},\ldots,\gamma_j^{k+1,*}\right).$$

For each $s=2,\ldots,K$, e_s must satisfy that the model determined by (1.53) for $j=s$ is compatible with the model for η_{s-1}^s determined by restriction (1.52) for $k=1$, if $s=2$, and by restriction (1.54) for $k=j=s-1$, if $s\geq 2$. That is, the functions $e_s's$ must satisfy that, for any $\left(\overline{\psi}^{s,*},\overline{\gamma}^{s,*},\overline{\tau}_s^*\right)$, there is at least one distribution for the observed data O that satisfies $E\left\{t_s\left(L_s\right)|\overline{A}_{s-1},\overline{L}_{s-1}\right\} = e_s\left(\overline{A}_{s-1},\overline{L}_{s-1};\overline{\psi}^{s,*},\overline{\gamma}^{s,*},\overline{\tau}_s^*\right)$ and $\eta_{s-1}^s\left(\overline{a}_{s-1},\overline{L}_{s-1}\right) = \eta_{s-1}^s\left(\overline{a}_{s-1},\overline{L}_{s-1};\psi^{s,*},\gamma^{s,*},\overline{\psi}^{s-1,*},\overline{\gamma}^{s-1,*},\overline{\tau}_{s-1}^*\right)$. For example, if one of the components of $t_s\left(L_s\right)$ is Y_s, then the corresponding component of $e_s\left(\overline{A}_{s-1},\overline{L}_{s-1};\overline{\psi}^{s,*},\overline{\gamma}^{s,*},\overline{\tau}_s^*\right)$ must be equal to $\eta_{s-1}^s\left(\overline{a}_{s-1},\overline{L}_{s-1};\psi^{s,*},\gamma^{s,*},\overline{\psi}^{s-1,*},\overline{\gamma}^{s-1,*},\overline{\tau}_{s-1}^*\right)$, where $\overline{\psi}^{1,*}\equiv\overline{\gamma}^{1,*}\equiv$nill.

As in the case of one outcome, we propose to derive e_1 from a fully parametric model for the density $f\left(V_1|Z\right)$. For $s=2,\ldots,K$, we propose to derive e_s from a fully parametric model for the density $f\left(L_s|\overline{A}_{s-1},\overline{L}_{s-1}\right)$ that is compatible with the model for $\eta_{s-1}^s\left(\overline{A}_{s-1},\overline{L}_{s-1}\right)$ determined by restriction (1.52) for $k=1$, if $s=2$, and by restriction (1.54) for $k=j=s-1$, if $s\geq 3$. Note that, in Section 1.7.1, the only parameter indexing each e_s is τ_s^*. Now, each e_s is also allowed to depend on $\left(\overline{\psi}^{s,*},\overline{\gamma}^{s,*},\overline{\tau}_{s-1}^*\right)$ because the proposed model for η_{s-1}^s depends on those parameters.

1.11.3 Estimation exploiting the compatible models

In this section, we extend the three estimators of Section 1.7.2 to the case of multiple outcomes. Since, for each $k=1,\ldots,K$, (1.45) is a MSMM for a single outcome, the same procedures of Section 1.7.2 can be used to compute R, DR and MR estimators of each $\psi^{k+1,*}$ and, hence, R, DR and MR estimators of ψ^*. However, the algorithms yielding the different estimators in the case of repeated outcomes differ slightly from those in the case of a single outcome. Specifically, they differ in the estimation of $\overline{\tau}_K^*$. This is because, although we also estimate $\overline{\tau}_K^*$ by method of moments fits of models (1.51) and (1.53), now the estimation of $\overline{\tau}_K^*$ can not be made separately from the estimation of the other parameters. To see why, note that, as indicated in the previous subsection, when $j\geq 2$, model (1.53) is indexed not only by τ_j^* but also by $\left(\overline{\psi}^{j,*},\overline{\gamma}^{j,*},\overline{\tau}_{j-1}^*\right)$. Hence, the estimation of τ_j^* requires a preliminary estimation of $\left(\overline{\psi}^{j,*},\overline{\gamma}^{j,*},\overline{\tau}_{j-1}^*\right)$.

Before we describe the estimation algorithms, we introduce the following notational conventions. We use (t_1) to denote restriction (1.51) and we use (t_j) to denote restriction (1.53) for each $j=2,\ldots,K$. Also, for each $k=1,\ldots,K$, we we use $\left(\eta_0^{k+1}\right)$ to denote equation (1.46) and $\left(\eta_1^{k+1}\right)$ to denote equation (1.52). Finally, we use $\left(\eta_j^{k+1}\right)$ to denote equation (1.54) for each $k=2,\ldots,K$ and $j=2,\ldots,k$. Note that, for $j\geq 2$, if one of the components of $t_j\left(L_j\right)$ is Y_j, then restriction (t_j) implies restriction $\left(\eta_{j-1}^j\right)$.

Now, for $k=1,\ldots,K$, let $\mathcal{R}_1^{k+1}$ be the model defined by restrictions (t_1) and $\left(\eta_1^{k+1}\right)$. For $k=2,\ldots,K, j=2,\ldots,k$, we define $\mathcal{R}_j^{k+1}$ as the model determined by restrictions $(t_1),(t_2),\left(\eta_1^2\right),\ldots,(t_j),\left(\eta_{j-1}^j\right),\left(\eta_j^{k+1}\right)$. That is, $\mathcal{R}_j^{k+1}$ is the model determined by restrictions (t_1), restrictions (t_s) and $\left(\eta_{s-1}^s\right)$ for $s=2,\ldots,j$ and restriction $\left(\eta_j^{k+1}\right)$. Also, for each $k=1,\ldots,K$,

let $\mathcal{M}^{k+1}$ be the model defined by restriction $\left(\eta_0^{k+1}\right)$. Finally, let $\mathcal{M}$ be the model determined by all the $\left(\eta_0^{k+1}\right), k = 1, \ldots, K$, i.e., the MSMM under the identifying assumptions.

A regression estimator

For each $k = 1, \ldots, K$, we now describe a regression estimator $\widehat{\psi}_R^{k+1}$ which, under regularity conditions, is CAN for $\psi^{k+1,*}$ under the model $\mathcal{R}_k^{k+1}$. Note that model $\mathcal{R}_k^{k+1}$ determines the parametric model for η_{k-1}^{k+1} defined by restriction $\left(\eta_{k-1}^{k+1}\right)$. This is because $\left(\eta_k^{k+1}\right)$ and (t_k) imply $\left(\eta_{k-1}^{k+1}\right)$. Likewise, for each $j = 1, \ldots, k-2, \mathcal{R}_k^{k+1}$ determines the parametric model for the η_j^{k+1} defined by restriction $\left(\eta_j^{k+1}\right)$ and also implies the parametric models for η_0^{k+1} determined by $\mathcal{M}^{k+1}$. Also notice that, for $k = 2, \ldots, K$, $\mathcal{R}_k^{k+1}$ implies $\mathcal{R}_{k-1}^k$, hence $\widehat{\psi}_R = \left(\widehat{\psi}_R^2, \ldots, \widehat{\psi}_R^{K+1}\right)$ will be CAN for ψ^* under $\mathcal{R}_K^{K+1}$. Finally note that, since $\mathcal{R}_K^{K+1}$ implies all the $\mathcal{R}_k^{k+1\prime}s$ and each $\mathcal{R}_k^{k+1}$ implies $\mathcal{M}^{k+1}$, then $\mathcal{R}_K^{K+1}$ implies the MSMM $\mathcal{M}$ under the identifying assumptions.

The following steps yield the proposed estimator:

1. Compute a method of moment estimator $\widehat{\tau}_{1,R}$ from the fit of the model $e_1\left(Z; \tau_1^*\right)$ for $E\left\{t_1\left(V_1\right)|Z\right\}$.

2. Compute the least squares estimator $\left(\widehat{\psi}_R^2, \widehat{\gamma}_R^2\right)$ from the fit of the model $\eta_1^2\left(A_1, L_1; \psi^{2,*}, \gamma^{2,*}, \widehat{\tau}_{1,R}\right)$ for $E\left(Y_2|A_1, L_1\right)$ where $\left(\psi^{2,*}, \gamma^{2,*}\right)$ is unknown and $\widehat{\tau}_{1,R}$ is regarded as known.

3. For $k = 2, \ldots, K$,

   (i) compute a method of moment estimator $\widehat{\tau}_{k,R}$ from the fit of the model $e_k\left(\overline{A}_{k-1}, \overline{L}_{k-1}; \widehat{\overline{\psi}}_R^k, \widehat{\overline{\gamma}}_R^k, \widehat{\overline{\tau}}_{k-1,R}, \tau_k^*\right)$ for $E\left\{t_k(L_k)|\overline{A}_{k-1}, \overline{L}_{k-1}\right\}$ where τ_k^* is unknown and $\left(\widehat{\overline{\psi}}_R^k, \widehat{\overline{\gamma}}_R^k, \widehat{\overline{\tau}}_{k-1,R}\right)$ is regarded as known, and

   (ii) compute the least squares estimator $\left(\widehat{\psi}_R^{k+1}, \widehat{\gamma}_R^{k+1}\right)$ from the fit of the model $\eta_k^{k+1}\left(\overline{A}_k, \overline{L}_k; \psi^{k+1,*}, \gamma^{k+1,*}, \widehat{\overline{\psi}}_R^k, \widehat{\overline{\gamma}}_R^k, \widehat{\overline{\tau}}_{k,R}\right)$ for $E\left(Y_{k+1}|\overline{A}_k, \overline{L}_k\right)$ where $\left(\psi^{k+1,*}, \gamma^{k+1,*}\right)$ is unknown and $\left(\widehat{\overline{\psi}}_R^k, \widehat{\overline{\gamma}}_R^k, \widehat{\overline{\tau}}_{k,R}\right)$ is regarded as known.

We now give an inductive heuristic argument of why $\left(\widehat{\overline{\psi}}_R^{k+1}, \widehat{\overline{\gamma}}_R^{k+1}, \widehat{\overline{\tau}}_{k,R}\right)$ should be consistent for $\left(\overline{\psi}^{k+1,*}, \overline{\gamma}^{k+1,*}, \overline{\tau}_k^*\right)$ under $\mathcal{R}_k^{k+1}$, $k = 1, \ldots, K$. Our argument is heuristic because we will omit indicating the regularity conditions under which a method of moment estimator that depends on consistent estimators of nuisance parameters, is itself consistent. Also, once consistency has been established, the convergence under regularity conditions of $\sqrt{n}\left\{\left(\widehat{\overline{\psi}}_R^{k+1}, \widehat{\overline{\gamma}}_R^{k+1}, \widehat{\overline{\tau}}_{k,R}\right) - \left(\overline{\psi}^{k+1,*}, \overline{\gamma}^{k+1,*}, \overline{\tau}_k^*\right)\right\}$ to a mean zero Normal distribution, $k = 1, \ldots, K$, follows immediately from the fact that $\left(\widehat{\overline{\psi}}_R^{k+1}, \widehat{\overline{\gamma}}_R^{k+1}, \widehat{\overline{\tau}}_{k,R}\right)$, $k = 1, \ldots, K$, ultimately solve a system of estimating equations.

The estimator $\widehat{\tau}_{1,R}$ is consistent for τ_1^* under $\mathcal{R}_1^2$ because $\mathcal{R}_1^2$ implies (t_1), which is a regression model for the outcome $t_1(V_1)$ on Z. Then, since $\mathcal{R}_1^2$ also implies $\left(\eta_1^2\right)$, which is a regression model for the outcome Y_2 on covariates A_1 and L_1, $\left(\widehat{\psi}_R^2, \widehat{\gamma}_R^2\right)$ is consistent for $\left(\psi^{2,*}, \gamma^{2,*}\right)$ under that model $\mathcal{R}_1^2$. Now, for $k \geq 2$, assume that $\left(\widehat{\overline{\psi}}_R^k, \widehat{\overline{\gamma}}_R^k, \widehat{\overline{\tau}}_{k-1,R}\right)$ is consistent for $\left(\overline{\psi}^{k,*}, \overline{\gamma}^{k,*}, \overline{\tau}_{k-1}^*\right)$ under $\mathcal{R}_{k-1}^k$. Since $\mathcal{R}_k^{k+1}$ implies $\mathcal{R}_{k-1}^k$, to prove that $\left(\widehat{\overline{\psi}}_R^{k+1}, \widehat{\overline{\gamma}}_R^{k+1}, \widehat{\overline{\tau}}_{k,R}\right)$ is consistent for $\left(\overline{\psi}^{k+1,*}, \overline{\gamma}^{k+1,*}, \overline{\tau}_k^*\right)$ under $\mathcal{R}_k^{k+1}$, it suffices to prove that $\left(\widehat{\psi}_R^{k+1}, \widehat{\gamma}_R^{k+1}, \widehat{\tau}_{k,R}\right)$ is consistent for $\left(\psi^{k+1,*}, \gamma^{k+1,*}, \tau_k^*\right)$ under $\mathcal{R}_k^{k+1}$. The fact that $\widehat{\tau}_{k,R}$ is consistent for τ_k^* under $\mathcal{R}_k^{k+1}$ follows from the facts that (1) (t_k) is a conditional mean model for the outcome $t_k(L_k)$ on covariates $\overline{A}_k$ and $\overline{L}_k$, (2) $\left(\widehat{\overline{\psi}}_R^k, \widehat{\overline{\gamma}}_R^k, \widehat{\overline{\tau}}_{k-1,R}\right)$ is consistent for $\left(\overline{\psi}^{k,*}, \overline{\gamma}^{k,*}, \overline{\tau}_{k-1}^*\right)$ under $\mathcal{R}_{k-1}^k$ by inductive hypothesis, and (3) $\mathcal{R}_k^{k+1}$ implies restriction (t_k) and model $\mathcal{R}_{k-1}^k$. Finally, $\left(\widehat{\psi}_R^{k+1}, \widehat{\gamma}_R^{k+1}\right)$ is consistent for $\left(\psi^{k+1,*}, \gamma^{k+1,*}\right)$ under $\mathcal{R}_k^{k+1}$ because (1) $\left(\eta_k^{k+1}\right)$ is a regression model for the outcome Y_{k+1} on covariates $\overline{A}_k$ and $\overline{L}_k$, (2) $\left(\widehat{\overline{\psi}}_R^k, \widehat{\overline{\gamma}}_R^k, \widehat{\overline{\tau}}_{k-1,R}\right)$ is consistent for $\left(\overline{\psi}^{k,*}, \overline{\gamma}^{k,*}, \overline{\tau}_{k-1}^*\right)$ under $\mathcal{R}_{k-1}^k$ by inductive hypothesis, (3) $\widehat{\tau}_{k,R}$ is consistent for τ_k^* under $\mathcal{R}_k^{k+1}$, and (4) $\mathcal{R}_k^{k+1}$ implies $\left(\eta_k^{k+1}\right)$ and $\mathcal{R}_{k-1}^k$.

A doubly robust estimator

For each $k = 1, \ldots, K$, let $\mathcal{P}_k$ be a parametric model $\pi_k\left(\overline{a}_k, \overline{l}_k; \alpha_k^*\right)$ for $\pi_k\left(\overline{a}_k, \overline{l}_k\right)$ as in the case of a single outcome. For each $k = 1, \ldots, K$, we will now describe an estimator $\widehat{\psi}_{DR}^{k+1}$ which is doubly robust in the sense that, under regularity conditions, it is consistent and asymptotically normal under the union model that assumes that either (i) $\mathcal{R}_k^{k+1}$ holds or (ii) $\mathcal{M}^{k+1}$ and $\mathcal{P}_1, \ldots, \mathcal{P}_k$ hold, but not necessarily both (i) and (ii) hold. The fact that, for each k, $\mathcal{R}_K^{K+1}$ determines $\mathcal{R}_k^{k+1}$ and $\mathcal{M}$ determines $\mathcal{M}^{k+1}$, implies that $\widehat{\psi}_{DR} = \left(\widehat{\psi}_{DR}^2, \ldots, \widehat{\psi}_{DR}^{K+1}\right)$ is CAN under the union model that assumes that either (i) $\mathcal{R}_K^{K+1}$ holds or (ii) $\mathcal{M}$ and $\mathcal{P}_1, \ldots, \mathcal{P}_K$ hold.

For each $k = 1, \ldots, K$, and for any $\eta^{k+1} \equiv \left(\eta_1^{k+1}, \ldots, \eta_k^{k+1}\right)$ and $\overline{\pi}_k \equiv (\pi_1, \ldots, \pi_k)$, not just the true ones, and any function d_k of $\left(\overline{A}_k, Z\right)$, define the estimating function

$$U_{d_k}^{k+1}\left(\psi^{k+1}, \eta^{k+1}, \overline{\pi}_k\right) \equiv S_{d_k}^{k+1,k}\left(\eta_k^{k+1}, \overline{\pi}_k\right) + \sum_{j=1}^{k-1} S_{d_k}^{k+1,j}\left(\eta_j^{k+1}, \eta_{j+1}^{k+1}, \pi_1, \ldots, \pi_j\right) + S_{d_k}^{k+1,0}\left(\psi^{k+1}, \eta_1^{k+1}\right), \tag{1.55}$$

where

$$S_{d_k}^{k+1,k}\left(\eta_k^{k+1}, \overline{\pi}_k\right) \equiv \frac{d_k\left(\overline{A}_k, Z\right)}{\prod_{s=1}^k \pi_s\left(\overline{A}_s, \overline{L}_s\right)} \left\{Y_{k+1} - \eta_k^{k+1}\left(\overline{A}_k, \overline{L}_k\right)\right\},$$

for $j = 1, \ldots, k-1$,

$$S_{d_k}^{k+1,j}\left(\eta_j^{k+1}, \eta_{j+1}^{k+1}, \pi_1, \ldots, \pi_j\right) \equiv \sum_{\overline{a}_{j+1}^k \in \overline{\mathcal{A}}_{j+1}^k} \frac{d_k(\overline{A}_j, \overline{a}_{j+1}^k, Z)}{\prod_{s=1}^j \pi_s\left(\overline{A}_s, \overline{L}_s\right)} \left\{\eta_{j+1}^{k+1}(\overline{A}_j, \overline{a}_{j+1}^k, \overline{L}_{j+1}) - \eta_j^{k+1}(\overline{A}_j, \overline{a}_{j+1}^k, \overline{L}_j)\right\},$$

and

$$S_{d_k}^{k+1,0}\left(\psi^{k+1},\eta_1^{k+1}\right)\equiv\sum_{\overline{a}_1^k\in\overline{\mathcal{A}}_1^k}d_k\left(\overline{a}_1^k,Z\right)\left\{\eta_1^{k+1}\left(\overline{a}_1^k,L_1\right)-m^{k+1}\left(\overline{a}_1^k,Z;\psi^{k+1}\right)\right\}.$$

Here, recall that $\overline{\mathcal{A}}_{j+1}^k\equiv\mathcal{A}_{j+1}\times\cdots\times\mathcal{A}_k$ and $\overline{a}_{j+1}^k\equiv(a_{j+1},\ldots,a_k)$, $k=2,\ldots,K, j=0,\ldots,k-1$. Also, we define $\sum_{j=1}^0\cdot\equiv 0$.

To compute $\widehat{\psi}_{DR}^{k+1}, k=1,\ldots,K$, we first run the procedure in the previous section and, for each $j=1,\ldots,k$, we define

$$\widehat{\eta}_{j,R}^{k+1}\left(\overline{a}_k,\bar{l}_j\right)\equiv\eta_j^{k+1}\left(\overline{a}_k,\bar{l}_j;\widehat{\psi}_R^{k+1},\widehat{\overline{\gamma}}_{j,R}^{k+1},\widehat{\overline{\psi}}_R^{j},\widehat{\overline{\gamma}}_R^{j},\widehat{\overline{\tau}}_{j,R}\right).$$

Second, for each $k=1,\ldots,K$, we compute $\widehat{\alpha}_k$ the MLE of α_k^* under $\mathcal{P}_k$ and define

$$\widehat{\pi}_k\left(\overline{a}_k,\bar{l}_k\right)\equiv\pi_k\left(\overline{a}_k,\bar{l}_k;\widehat{\alpha}_k\right).$$

Finally, the estimator $\widehat{\psi}_{DR}^{k+1}$ solves

$$\mathbb{P}_n\left\{U_{\widehat{d}_{k+1}}^{k+1}\left(\psi^{k+1},\widehat{\eta}_R^{k+1},\widehat{\overline{\pi}}_k\right)\right\}=0$$

where $\widehat{d}_k\left(\overline{A}_k,Z\right)$ is any, possibly data dependent, function of the same dimension as $\psi^{k+1,*}$, for instance, $\widehat{d}_k\left(\overline{A}_k,Z\right)=\left\{\partial m^{k+1}\left(\overline{A}_k,Z;\psi^{k+1}\right)/\partial\psi^{k+1}\right\}\big|_{\psi^{k+1}=\widehat{\psi}_R^{k+1}}$. The estimator $\widehat{\psi}_{DR}^{k+1}$ is doubly robust essentially because (I) as shown in [1], under $\mathcal{M}^{k+1}$,

$$E\left\{U_{d_k}^{k+1}\left(\psi^{k+1,*},\eta^{k+1\prime},\overline{\pi}_k'\right)\right\}=0,\tag{1.56}$$

if either $\eta^{k+1\prime}$ is equal to the true η^{k+1} or $\overline{\pi}_k'$ is equal to the true $\overline{\pi}_k$, (II) by construction, $\widehat{\eta}_{j,R}^{k+1}$ converges to the true η_j^{k+1} under $\mathcal{R}_k^{k+1}, j=1,\ldots,k$, and (III) $\widehat{\pi}_j$ converges to the true π_j under $\mathcal{P}_j$, $j=1,\ldots,k$.

Fact (II) holds because, for each $j=1,\ldots,k$, (1) as shown in previous section, $\left(\widehat{\psi}_R^{k+1},\widehat{\overline{\gamma}}_{j,R}^{k+1}\right)$ is consistent for $\left(\psi_R^{k+1,*},\overline{\gamma}_{j,R}^{k+1,*}\right)$ under $\mathcal{R}_k^{k+1}$, $\left(\widehat{\overline{\psi}}_R^{j},\widehat{\overline{\gamma}}_R^{j},\widehat{\overline{\tau}}_{j-1,R}\right)$ is consistent for $\left(\overline{\psi}^{j,*},\overline{\gamma}^{j,*},\overline{\tau}_{j-1}^*\right)$ under $\mathcal{R}_{j-1}^j$, and $\widehat{\tau}_{j,R}$ is consistent for τ_j^* under $\mathcal{R}_j^{j+1}$, and (2) $\mathcal{R}_k^{k+1}$ implies $\mathcal{R}_{j-1}^j$ and $\mathcal{R}_j^{j+1}$.

Once consistency has been established, the convergence under regularity conditions of $\sqrt{n}\left(\widehat{\psi}_{DR}^{k+1}-\psi^{k+1,*}\right)$ to a mean zero Normal distribution, $k=1,\ldots,K$, follows immediately from the fact that $\widehat{\psi}_{DR}^{k+1}, k=1,\ldots,K$, and all nuisance parameter estimators ultimately solve a system of estimating equations.

A multiply robust estimator

We will next construct an estimator $\widehat{\psi}_{MR}^{k+1}$, for each $k=1,\ldots,K$, that is multiply robust in the sense that, under regularity conditions, it is consistent and asymptotically normal for ψ^{k+1*} under the model that assumes that $\mathcal{M}^{k+1}$ holds and that, for each $j\in\{1,\ldots,k\}$, either $\mathcal{R}_j^{k+1}$ or $\mathcal{P}_j$ holds. If we define $\mathcal{R}_j$ as the model that assumes that $\mathcal{R}_j^{k+1}$ holds for every $k=j,\ldots,K$, then

each $\widehat{\psi}_{MR}^{k+1}$ is also consistent for ψ^{k+1*} under the more restrictive model that assumes that $\mathcal{M}$ holds and that, for each $j \in \{1,\ldots,K\}$, either $\mathcal{R}_j$ or $\mathcal{P}_j$ holds. Hence, $\widehat{\psi}_{MR} \equiv \left(\widehat{\psi}_{MR}^{2},\ldots,\widehat{\psi}_{MR}^{K+1}\right)$ is consistent for ψ^* under that model.

Note that $\mathcal{R}_1$ is determined by restrictions $(t_1), \left(\eta_1^2\right), \left(\eta_1^3\right), \ldots, \left(\eta_1^{K+1}\right)$. That is, $\mathcal{R}_1$ is determined by restriction (t_1) and by restrictions $\left(\eta_1^{k+1}\right)$ for $k = 1,\ldots,K$. Likewise, for $j = 2,\ldots,K$, $\mathcal{R}_j$ is determined by restrictions $(t_1), (t_2), \left(\eta_1^2\right), \ldots, (t_j), \left(\eta_{j-1}^{j}\right), \left(\eta_j^{j+1}\right), \ldots, \left(\eta_j^{K+1}\right)$. That is, $\mathcal{R}_j$ is determined by restriction (t_1), by restrictions (t_s) and $\left(\eta_{s-1}^{s}\right)$ for $s = 2,\ldots,j$ and by restrictions $\left(\eta_j^{s+1}\right)$ for $s = j,\ldots,K$. Hence, the fact that, for $j = 1,\ldots,K-1$ and $k = j,\ldots,K$, restrictions (t_j) and $\left(\eta_j^{k+1}\right)$ imply restriction $\left(\eta_{j-1}^{k+1}\right)$, then gives that $\mathcal{R}_j$ implies $\mathcal{R}_{j-1}, j = 2,\ldots,K$, and $\mathcal{R}_1$ implies $\mathcal{M}$. Therefore, $\widehat{\psi}_{MR}$ is consistent under the model that assumes that any but, not necessarily all, of the following assumptions (i), (ii) or (iii) is satisfied: (i) $\mathcal{R}_K = \mathcal{R}_K^{K+1}$ holds; (ii) for some $j \in \{1,\ldots,K-1\}$ models $\mathcal{R}_j, \mathcal{P}_{j+1}, \ldots, \mathcal{P}_K$ hold; (iii) model $\mathcal{M}$ and models $\mathcal{P}_1,\ldots,\mathcal{P}_K$ hold. Thus, whereas $\widehat{\psi}_{DR}$ yields valid inferences if (i) or (iii) holds, $\widehat{\psi}_{MR}$ also does it if (ii) holds even when (i) and (iii) fail.

The following steps yield the proposed estimator:

1. For $k = 1,\ldots,K$ we compute $\widehat{\alpha}_k$ the MLE of α_k^* under $\mathcal{P}_k$ and define

$$\widehat{\pi}_k\left(\overline{a}_k, \overline{l}_k\right) \equiv \pi_k\left(\overline{a}_k, \overline{l}_k; \widehat{\alpha}_k\right).$$

2. Compute a method of moments estimator $\widehat{\tau}_{1,MR}$ from the fit of the model $e_1\left(Z;\tau_1^*\right)$ for $E\left\{t_1(V_1)|Z\right\}$.

3. Define $\eta_1^2\left(a_1,l_1;\psi^2,\gamma^2\right) \equiv \eta_1^2\left(a_1,l_1;\psi^2,\gamma^2,\widehat{\tau}_{1,MR}\right)$ and $\overset{\bullet}{\eta}{}_1^2\left(a_1,l_1;\psi^2,\gamma^2\right) \equiv \partial\eta_1^2\left(a_1,l_1;\psi^2,\gamma^2\right)/\partial\left(\psi^2,\gamma^2\right)$.
   Compute $\left(\widehat{\psi}^{2,(1)},\widehat{\gamma}^{2,(1)}\right)$ solving

$$\mathbb{P}_n\left[\frac{\overset{\bullet}{\eta}{}_1^2\left(A_1,L_1;\psi^2,\gamma^2\right)}{\widehat{\pi}_1\left(A_1,L_1\right)}\left\{Y_2 - \eta_1^2\left(A_1,L_1;\psi^2,\gamma^2\right)\right\}\right] = 0.$$

   Define $\widehat{\eta}_1^2\left(a_1,l_1\right) \equiv \eta_1^2\left(a_1,l_1;\widehat{\psi}^{2,(1)},\widehat{\gamma}^{2,(1)}\right)$.

4. Define $\overset{\bullet}{m}{}^2\left(a_1,z;\psi^2\right) \equiv \partial m^2\left(a_1,z;\psi^2\right)/\partial\psi^2$.
   Compute $\widehat{\psi}_{MR}^2$ solving

$$\mathbb{P}_n\left[\sum_{a_1\in\mathcal{A}_1}\overset{\bullet}{m}{}^2\left(a_1,Z;\widehat{\psi}^{2,(1)}\right)\left\{\widehat{\eta}_1^2\left(a_1,L_1\right) - m^2\left(a_1,Z;\psi^2\right)\right\}\right] = 0.$$

5. For $k = 2,\ldots,K$,

(i) define $\widehat{\overline{\psi}}^{k,(k-1)} \equiv \left(\widehat{\psi}^{2,(1)}, \ldots, \widehat{\psi}^{k,(k-1)}\right)$ and $\widehat{\overline{\gamma}}^{k,(k-1)} \equiv \left(\widehat{\gamma}^{2,(1)}, \ldots, \widehat{\gamma}^{k,(k-1)}\right)$ and compute a method of moments estimator $\widehat{\tau}_{k,MR}$ from the fit of the model $e_k\left(\overline{A}_{k-1}, \overline{L}_{k-1}; \widehat{\overline{\psi}}^{k,(k-1)}, \widehat{\overline{\gamma}}^{k,(k-1)}, \widehat{\overline{\tau}}_{k-1,MR}, \tau_k^*\right)$ for $E\left\{t_k(L_k)|\overline{A}_{k-1}, \overline{L}_{k-1}\right\}$ where τ_k^* is unknown and $\left(\widehat{\overline{\psi}}^{k,(k-1)}, \widehat{\overline{\gamma}}^{k,(k-1)}, \widehat{\overline{\tau}}_{k-1,MR}\right)$ is regarded as known,

(ii) define $\eta_k^{k+1}\left(\overline{a}_k, \overline{l}_k; \psi^{k+1}, \gamma^{k+1}\right) \equiv \eta_k^{k+1}\left(\overline{a}_k, \overline{l}_k; \psi^{k+1}, \gamma^{k+1}, \widehat{\overline{\psi}}^{k,(k-1)}, \widehat{\overline{\gamma}}^{k,(k-1)}, \widehat{\overline{\tau}}_{k,MR}\right)$ and $\overset{\bullet}{\eta}_k^{k+1}\left(\overline{a}_k, \overline{l}_k; \psi^{k+1}, \gamma^{k+1}\right) \equiv \partial \eta_k^{k+1}\left(\overline{a}_K, \overline{l}_K; \psi^{k+1}, \gamma^{k+1}\right) / \partial\left(\psi^{k+1}, \gamma^{k+1}\right)$.
Compute $\left(\widehat{\psi}^{k+1,(k)}, \widehat{\gamma}^{k+1,(k)}\right)$ solving

$$\mathbb{P}_n\left[\frac{\overset{\bullet}{\eta}_k^{k+1}\left(\overline{A}_k, \overline{L}_k; \psi^{k+1}, \gamma^{k+1}\right)}{\prod_{j=1}^k \widehat{\pi}_j\left(\overline{A}_j, \overline{L}_j\right)}\left\{Y_{k+1} - \eta_k^{k+1}\left(\overline{A}_k, \overline{L}_k; \psi^{k+1}, \gamma^{k+1}\right)\right\}\right] = 0.$$

Define $\widehat{\eta}_k^{k+1}\left(\overline{a}_k, \overline{l}_k\right) \equiv \eta_k^{k+1}\left(\overline{a}_k, \overline{l}_k; \widehat{\psi}^{k+1,(k)}, \widehat{\gamma}^{k+1,(k)}\right)$.

(iii) For $s = k-1, \ldots, 2$, iteratively define
$\eta_s^{k+1}\left(\overline{a}_k, \overline{l}_s; \psi^{k+1}, \overline{\gamma}_s^{k+1}\right) \equiv \eta_s^{k+1}\left(\overline{a}_k, \overline{l}_s; \psi^{k+1}, \overline{\gamma}_s^{k+1}, \widehat{\overline{\psi}}^{s,(s-1)}, \widehat{\overline{\gamma}}^{s,(s-1)}, \widehat{\overline{\tau}}_{s,MR}\right)$
and $\overset{\bullet}{\eta}_s^{k+1}\left(\overline{a}_k, \overline{l}_s; \psi^{k+1}, \overline{\gamma}_s^{k+1}\right) \equiv \partial \eta_s^{k+1}\left(\overline{a}_k, \overline{l}_s; \psi^{k+1}, \overline{\gamma}_s^{k+1}\right) / \partial\left(\psi^{k+1}, \overline{\gamma}_s^{k+1}\right)$.
Compute $\left(\widehat{\psi}^{k+1,(s)}, \widehat{\overline{\gamma}}_s^{k+1,(s)}\right) \equiv \left(\widehat{\psi}^{k+1,(s)}, \widehat{\gamma}_1^{k+1,(s)}, \ldots, \widehat{\gamma}_s^{k+1,(s)}\right)$ solving

$$\mathbb{P}_n\left[\sum_{\overline{a}_{s+1}^k \in \overline{\mathcal{A}}_{s+1}^k} \frac{\overset{\bullet}{\eta}_s^{k+1}\left(\overline{A}_s, \overline{a}_{s+1}^k, \overline{L}_s; \widehat{\psi}^{k+1,(k)}, \widehat{\overline{\gamma}}_s^{k+1,(k)}\right)}{\prod_{j=1}^s \widehat{\pi}_j\left(\overline{A}_j, \overline{L}_j\right)}\left\{\widehat{\eta}_{s+1}^{k+1}\left(\overline{A}_s, \overline{a}_{s+1}^k, \overline{L}_{s+1}\right) - \eta_s^{k+1}\left(\overline{A}_s, \overline{a}_{s+1}^k, \overline{L}_s; \psi^{k+1}, \overline{\gamma}_s^{k+1}\right)\right\}\right] = 0.$$

Define $\widehat{\eta}_s^{k+1}(\overline{a}_k, \overline{l}_s) \equiv \eta_s^{k+1}\left(\overline{a}_k, \overline{l}_s; \widehat{\psi}^{k+1,(s)}, \widehat{\overline{\gamma}}_s^{k+1,(s)}\right)$.

(iv) Define $\eta_1^{k+1}\left(\overline{a}_k, l_1; \psi^{k+1}, \gamma_1^{k+1}\right) \equiv \eta_1^{k+1}\left(\overline{a}_k, l_1; \psi^{k+1}, \gamma_1^{k+1}, \widehat{\tau}_{1,MR}\right)$ and
$\overset{\bullet}{\eta}_1^{k+1}\left(\overline{a}_k, \overline{l}_s; \psi^{k+1}, \gamma_1^{k+1}\right) \equiv \partial \eta_1^{k+1}\left(\overline{a}_k, l_1; \psi^{k+1}, \gamma_1^{k+1}\right) / \partial\left(\psi^{k+1}, \gamma_1^{k+1}\right)$.
Compute $\left(\widehat{\psi}^{k+1,(1)}, \widehat{\gamma}_1^{k+1,(1)}\right)$ solving

$$\mathbb{P}_n\left[\sum_{\overline{a}_2^k \in \overline{\mathcal{A}}_2^k} \frac{\overset{\bullet}{\eta}_1^{k+1}\left(A_1, \overline{a}_2^k, L_1; \widehat{\psi}^{k+1,(k)}, \widehat{\gamma}_1^{k+1,(k)}\right)}{\widehat{\pi}_1\left(A_1, L_1\right)}\left\{\widehat{\eta}_2^{k+1}\left(A_1, \overline{a}_2^k, \overline{L}_2\right) - \eta_1^{k+1}\left(A_1, \overline{a}_2^k, L_1; \psi^{k+1}, \gamma_1^{k+1}\right)\right\}\right] = 0.$$

Define $\widehat{\eta}_1^{k+1}(\overline{a}_k, l_1) \equiv \eta_1^{k+1}\left(\overline{a}_k, l_1; \widehat{\psi}^{k+1,(1)}, \widehat{\gamma}_1^{k+1,(1)}\right)$.

(v) Define $\overset{\bullet}{m}^{k+1}\left(\overline{a}_k, z; \psi^{k+1}\right) \equiv \partial m^{k+1}\left(\overline{a}_k, z; \psi^{k+1}\right)/\partial\psi^{k+1}$.
Compute $\widehat{\psi}_{MR}^{k+1}$ solving

$$\mathbb{P}_n\left[\sum_{\overline{a}_1^k \in \overline{\mathcal{A}}_1^k} \overset{\bullet}{m}^{k+1}\left(\overline{a}_1^k, Z; \widehat{\psi}^{k+1,(k)}\right)\left\{\widehat{\eta}_1^{k+1}(\overline{a}_1^k, L_1) - m^{k+1}\left(\overline{a}_1^k, Z; \psi^{k+1}\right)\right\}\right] = 0.$$

The multiple robustness of each $\widehat{\psi}_{MR}^{k+1}$ is a consequence of the following facts:

(I) The identity (1.56) holds not only when $\eta^{k+1\prime} \equiv \left(\eta_1^{k+1\prime}, \ldots, \eta_k^{k+1\prime}\right)$ is equal to the true $\eta^{k+1} \equiv \left(\eta_1^{k+1}, \ldots, \eta_k^{k+1}\right)$ or $\overline{\pi}'_k \equiv (\pi'_1, \ldots, \pi'_k)$ is equal to the true $\overline{\pi}_k \equiv (\pi_1, \ldots, \pi_k)$, but also under the weaker condition that, for each $j \in \{1, \ldots, k\}$, either η'_j is equal to the true η_j or π'_j is equal to the true π_j.

(II) The estimator $\widehat{\pi}_j$ in step 1 is consistent for π_j under $\mathcal{P}_j$, $j = 1, \ldots, k$.

(III) The estimator $\widehat{\eta}_k^{k+1}$ in step 3 (if $k = 1$) or step 5.ii, (if $k \geq 2$), is consistent for the true η_k^{k+1} under model $\mathcal{R}_k^{k+1}$.

(IV) If $k \geq 2$, for each $s = 1, \ldots, (k-1)$, the estimator $\widehat{\eta}_s^{k+1}$ in step 5.iii (if $s \geq 2$) or step 5.iv (if $s = 1$) is itself multiply robust in that it is consistent for the true η_s^{k+1} under the model that assumes that $\mathcal{R}_s^{k+1}$ holds and that, for each $j \in \{s+1, \ldots, k\}$, either $\mathcal{R}_j^{k+1}$ or $\mathcal{P}_j$ holds.

(V) The estimator $\widehat{\psi}_{MR}^{k+1}$ in step 4 (if $k = 1$) or step 5.v (if $k \geq 2$) actually solves the equation $\mathbb{P}_n\left\{U_{\widehat{d}_k}^{k+1}\left(\psi^{k+1}, \widehat{\eta}^{k+1}, \widehat{\overline{\pi}}_k\right)\right\} = 0$ for $\widehat{d}_k\left(\overline{A}_k, Z\right) = \overset{\bullet}{m}^{k+1}\left(\overline{A}_k, Z; \widehat{\psi}^{k+1,(k)}\right)$,
$\widehat{\eta}^{k+1} \equiv \left(\widehat{\eta}_1^{k+1}, \ldots, \widehat{\eta}_k^{k+1}\right)$ computed in steps 3 (if $k = 1$) or steps 5.ii-iv (if $k \geq 2$), and $\widehat{\overline{\pi}}_k \equiv (\widehat{\pi}_1, \ldots, \widehat{\pi}_k)$ computed in step 1.

Facts (I)-(V) imply that, under regularity conditions, $\widehat{\psi}_{MR}^{k+1}$ is consistent and asymptotically for $\psi^{k+1,*}$ under the model that assumes that $\mathcal{M}^{k+1}$ holds and that, for each $j \in \{1, \ldots, k\}$, either $\mathcal{R}_j^{k+1}$ or $\mathcal{P}_j$ holds. Because ultimately $\widehat{\psi}_{MR}^{k+1}, k = 1, ..., K$ and all nuisance parameters ultimately solve a system of estimating equations, then a consistent estimator of the asymptotic variance estimator of the entire vector of parameters (i.e. $\widehat{\psi}_{MR}^{k+1}, k = 1, ..., K$ and all nuisance parameters) can be obtained by the usual sandwich variance estimator. A consistent estimator of the asymptotic variance of each $\widehat{\psi}_{MR}^{k+1}$ can then be obtained by extracting the specific entry of the sandwich variance estimator matrix. Because of the complexity of the estimating functions, this procedure might be impractical. Nevertheless, just as in the case of a single outcome, the bootstrap can be used instead to estimate the variance of each $\widehat{\psi}_{MR}^{k+1}$.

As noted in Section 1.7.2, fact (I) is a consequence of the likelihood factorization that takes place in coarsened at random models in [24].

To prove fact (III), we must show that, for every $k \geq 1$, $\left(\widehat{\overline{\psi}}^{k+1,(k)}, \widehat{\overline{\gamma}}^{k+1,(k)}, \widehat{\overline{\tau}}_{k,MR}\right)$ converges to $\left(\overline{\psi}^{k+1,*}, \overline{\gamma}^{k+1,*}, \overline{\tau}_k^*\right)$ under $\mathcal{R}_k^{k+1}$. We prove it by induction in k. When $k = 1$, the estimator $\widehat{\tau}_{1,MR}$ is consistent for τ_1^* under $\mathcal{R}_1^2$ because (t_1) is a conditional mean model for the outcome $t_1(V_1)$ on Z and $\mathcal{R}_1^2$ implies (t_1). Then, since $\mathcal{R}_1^2$ also implies $\left(\eta_1^2\right)$, which is a regression model for the

outcome Y_2 on covariates A_1 and L_1, $\left(\widehat{\psi}^{2,(1)}, \widehat{\gamma}^{2,(1)}\right)$ is consistent for $\left(\psi^{2,*}, \gamma^{2,*}\right)$ under that model. Now, for $k \geq 2$, assume that $\left(\widehat{\overline{\psi}}^{k,(k-1)}, \widehat{\overline{\gamma}}^{k,(k-1)}, \widehat{\overline{\tau}}_{k-1,MR}\right)$ is consistent for $\left(\overline{\psi}^{k,*}, \overline{\gamma}^{k,*}, \overline{\tau}^{*}_{k-1}\right)$ under $\mathcal{R}^{k}_{k-1}$. Since $\mathcal{R}^{k+1}_{k}$ implies $\mathcal{R}^{k}_{k-1}$, to prove that $\left(\widehat{\overline{\psi}}^{k+1,(k)}, \widehat{\overline{\gamma}}^{k+1,(k)}, \widehat{\overline{\tau}}_{k,MR}\right)$ is consistent for $\left(\overline{\psi}^{k+1,*}, \overline{\gamma}^{k+1,*}, \overline{\tau}^{*}_{k}\right)$ under $\mathcal{R}^{k+1}_{k}$, it suffices to prove that $\left(\widehat{\psi}^{k+1,(k)}, \widehat{\gamma}^{k+1,(k)}, \widehat{\tau}_{k,MR}\right)$ is consistent for $\left(\psi^{k+1,*}, \gamma^{k+1,*}, \tau^{*}_{k}\right)$ under $\mathcal{R}^{k+1}_{k}$. The fact that $\widehat{\tau}_{k,MR}$ is consistent for τ^{*}_{k} under $\mathcal{R}^{k+1}_{k}$ follows from the facts that: (1) (t_k) is a conditional mean model for the outcome $t_k\left(L_k\right)$ on covariates $\overline{A}_k$ and $\overline{L}_k$, (2) $\left(\widehat{\overline{\psi}}^{k,(k-1)}, \widehat{\overline{\gamma}}^{k,(k-1)}, \widehat{\overline{\tau}}_{k-1,MR}\right)$ is consistent for $\left(\overline{\psi}^{k,*}, \overline{\gamma}^{k,*}, \overline{\tau}^{*}_{k-1}\right)$ under $\mathcal{R}^{k}_{k-1}$ by inductive hypothesis, and (3) $\mathcal{R}^{k+1}_{k}$ implies restriction (t_k) and model $\mathcal{R}^{k}_{k-1}$. Finally, $\left(\widehat{\psi}^{k+1,(k)}, \widehat{\gamma}^{k+1,(k)}\right)$ is consistent for $\left(\psi^{k+1,*}, \gamma^{k+1,*}\right)$ under $\mathcal{R}^{k+1}_{k}$ because (1) $\left(\eta^{k+1}_{k}\right)$ is a regression model for the outcome Y_{k+1} on covariates $\overline{A}_k$ and $\overline{L}_k$, (2) $\left(\widehat{\overline{\psi}}^{k,(k-1)}, \widehat{\overline{\gamma}}^{k,(k-1)}, \widehat{\overline{\tau}}_{k-1,MR}\right)$ is consistent for $\left(\overline{\psi}^{k,*}, \overline{\gamma}^{k,*}, \overline{\tau}^{*}_{k-1}\right)$ under $\mathcal{R}^{k}_{k-1}$ by inductive hypothesis, (3) $\widehat{\tau}_{k,MR}$ is consistent for τ^{*}_{k} under $\mathcal{R}^{k+1}_{k}$, and (4) $\mathcal{R}^{k+1}_{k}$ implies $\left(\eta^{k+1}_{k}\right)$ and $\mathcal{R}^{k}_{k-1}$.

Since each $\mathcal{M}^{k+1}$ is a MSMM for a single outcome, fact (IV) follows from arguments analogous to those of Section 1.7.2 and from the facts that, for each $k = 2, \ldots, K$, (1) $\widehat{\tau}_{1,MR}$ is consistent for τ^{*}_{1} under $\mathcal{R}^{k+1}_{1}$ and (2) for each $s = 2, \ldots, k-1$, $\left(\widehat{\overline{\psi}}^{s,(s-1)}, \widehat{\overline{\gamma}}^{s,(s-1)}, \widehat{\overline{\tau}}_{s,MR}\right)$ converges to $\left(\overline{\psi}^{s,*}, \overline{\gamma}^{s,*}, \overline{\tau}^{*}_{s}\right)$ under $\mathcal{R}^{k+1}_{s}$. Fact (1) holds because $\mathcal{R}^{k+1}_{1}$ implies (t_1). Since for each $k = 2, \ldots, K, s = 2, \ldots, k-1$, $\left(\widehat{\overline{\psi}}^{s,(s-1)}, \widehat{\overline{\gamma}}^{s,(s-1)}, \widehat{\overline{\tau}}_{s-1,MR}\right)$ is consistent for $\left(\overline{\psi}^{s,*}, \overline{\gamma}^{s,*}, \overline{\tau}^{*}_{s-1}\right)$ under $\mathcal{R}^{s}_{s-1}$ and $\mathcal{R}^{k+1}_{s}$ implies $\mathcal{R}^{s}_{s-1}$, then to prove fact (2) it suffices to prove that $\widehat{\tau}_{s,MR}$ is consistent for τ^{*}_{s} under $\mathcal{R}^{k+1}_{s}$ for every $k = 2, \ldots, K, s = 2, \ldots, k-1$. It holds because (a) (t_s) is a conditional mean model for the outcome $t_s\left(L_s\right)$ on covariates $\overline{A}_s$ and $\overline{L}_s$, (b) $\left(\widehat{\overline{\psi}}^{s,(s-1)}, \widehat{\overline{\gamma}}^{s,(s-1)}, \widehat{\overline{\tau}}_{s-1,MR}\right)$ is consistent for $\left(\overline{\psi}^{s,*}, \overline{\gamma}^{s,*}, \overline{\tau}^{*}_{s-1}\right)$ under $\mathcal{R}^{k+1}_{s}$, and (c) $\mathcal{R}^{k+1}_{s}$ implies (t_s).

Finally, fact (V) follows from the facts that, when $\widehat{d}_k\left(\overline{A}_k, Z\right) = \overset{\bullet}{m}{}^{k+1}\left(\overline{A}_k, Z; \widehat{\psi}^{k+1,(k)}\right)$, (i) $\widehat{\psi}^{k+1}_{MR}$ solves $\mathbb{P}_n\left\{S^{k+1,0}_{\widehat{d}_k}\left(\psi^{k+1}, \widehat{\eta}^{k+1}_{1}\right)\right\} = 0$ by step 4 (if $k = 1$) or step 5.v (if $k \geq 2$), (ii) if $k \geq 2$, for each $s \in \{1, \ldots, k-1\}$, $\mathbb{P}_n\left\{S^{k+1,s}_{\widehat{d}_k}\left(\widehat{\eta}^{k+1}_{s}, \widehat{\eta}^{k+1}_{s+1}, \widehat{\pi}_1, \ldots, \widehat{\pi}_s\right)\right\} = 0$, by step 5.iii-iv and the fact that $\overset{\bullet}{m}{}^{k+1}$ is a subvector of $\overset{\bullet}{\eta}{}^{k+1}_{s}$, and (iii) $\mathbb{P}_n\left\{S^{k+1,k}_{\widehat{d}_k}\left(\widehat{\eta}^{k+1}_{k}, \widehat{\overline{\pi}}_k\right)\right\} = 0$ by step 3 (if $k = 1$) or step 5.ii (if $k \geq 2$) and the fact that $\overset{\bullet}{m}{}^{k+1}$ is also a subvector of $\overset{\bullet}{\eta}{}^{k+1}_{k}$.

1.12 Resumen

Los modelos marginales estructurales para la media (MMEM) son herramientas populares para modelar el efecto causal de tratamientos variantes en el tiempo en presencia variables confusoras variantes en el tiempo que están afectadas por el tratamiento recibido en el pasado. Desde que fueron propuestos por primera vez por Robins ([30]), los MMEM se han aplicado para analizar numerosos estudios relacionados con la salud. Por ejemplo, estudios sobre el efecto del tratamiento antirretroviral altamente activo en el recuento de CD4 ([9]), el efecto del uso del organizador de pastillas sobre la adherencia a los medicamentos antirretrovirales y la carga viral ([28]) y el efecto de la soledad en los síntomas depresivos ([59]).

En este capítulo proponemos un estimador paramétrico múltiple robusto para el parámetro de un MMEM en el caso particular en que la variable de respuesta en continua y no acotada. Nuestro estimador es de fácil implementación ya que requiere simplemente del ajuste de una serie de regresiones iteradas por el método de mínimos cuadrados pesados.

Para ser concretos sobre las contribuciones de este capítulo comenzamos por definir formalmente los modelos marginales estructurales para la media. Supongamos que los datos observados son n replicaciones i.i.d. un vector

$$O \equiv (L_1, A_1, \ldots, L_K, A_K, Y)$$

donde Y es una variables de respuesta de interés medida en el instante de tiempo t_{K+1} que es no acotada, es decir con rango en la recta real. Para cada $k = 1, \ldots, K$, A_k es el tratamiento recibido en el tiempo t_k que toma valores en un conjunto finito $\mathcal{A}_k$ y L_k es un vector de covariables medido en el tiempo t_k^- , es decir un instante previo a t_k $(t_{k-1} < t_k)$, que toma valores en un subconjunto $\mathcal{L}_k$ de un espacio euclídeo. En lo que sigue, para $k \in \{1, ..., K\}$ y para cualquier $\{v_r\}_{1 \leq r \leq K}$, denotamos $\overline{v}_k \equiv (v_1, \ldots, v_k)$ y $\underline{v}_k \equiv (v_k, \ldots, v_K)$. Así mismo, para $k \in \{1, ..., K\}$ y cualquier colección de conjuntos $\{\mathcal{C}_r\}_{1 \leq r \leq K}$, denotamos $\overline{\mathcal{C}}_k \equiv \mathcal{C}_1 \times \cdots \times \mathcal{C}_k$ y $\underline{\mathcal{C}}_k \equiv \mathcal{C}_k \times \cdots \times \mathcal{C}_K$.

Para cada historia de tratamiento $\overline{a}_K = (a_1, \ldots, a_K)$, sea $Y_{\overline{a}_K}$ la respuesta del individuo si éste hubiese seguido el régimen de tratamiento $\overline{a}_K$. Bajo las suposiciones de

(1) consistencia:

$$Y_{\overline{a}_K} = Y \text{ si } \overline{A}_K = \overline{a}_K$$

(2) inexistencia de variables confusoras no medidas (no unmeasured confounding, NUC): para todo $\overline{a}_K$ y k,

$$Y_{\overline{a}_K} \amalg A_k | \overline{L}_k, \overline{A}_{k-1} = \overline{a}_{k-1}$$

y

(3) positividad: para todo k y $\overline{a}_k$, si $f(\overline{a}_{k-1}, \overline{l}_k) > 0$ entonces $f(a_k | \overline{a}_{k-1}, \overline{l}_k) > 0$,

es bien sabido ([32]) que $E(Y_{\overline{a}_K} | Z)$ está identificado, donde Z es un subvector de L_1. En lo que sigue, nos referiremos a (1) - (3) como las *condiciones de indentificabilidad.*

En este capítulo, asumimos las condiciones de identificabilidad y hacemos propuestas de estimación para el parámetro $\psi^* \in R^p$ del MMEM ([30])

$$E\left(Y_{\overline{a}_K} | Z\right) = m\left(\overline{a}_K, Z; \psi^*\right) \quad \text{para todo } \overline{a}_K, \tag{1.57}$$

donde $m(\cdot, \cdot; \cdot)$ es una función conocida.

A continuación, discutimos las propuestas existentes para la estimación de los parámetros de los MMEM.

Bajo las condiciones de identificabilidad, Robins ([35]) probó que (1.57) es equivalente al modelo para los datos observados O definido por

$$E\left[\pi^p\left(\overline{A}_K,\overline{L}_K\right)^{-1}\left\{Y-m\left(\overline{A}_K,Z;\psi^*\right)\right\}\middle|\,\overline{A}_K,Z\right]=0$$

donde $\pi^p\left(\overline{a}_K,\overline{l}_K\right)\equiv\prod_{j=1}^{K}\pi_j\left(\overline{a}_j,\overline{l}_j\right)$ con $\pi_1\left(a_1,l_1\right)\equiv\Pr\left(A_1=a_1|L_1=l_1\right)$ y $\pi_j\left(\overline{a}_j,\overline{l}_j\right)\equiv\Pr\left(A_j=a_j|\overline{A}_{j-1}=\overline{a}_{j-1},\overline{L}_j=\overline{l}_j\right),2\leq j\leq K$. Esta observación dio origen al estimador IPTW, $\widehat{\psi}_{IPTW}$, que se obtiene ajustando una regresión por mínimos cuadrados pesados con variable de respuesta Y y covariables $(\overline{A}_K,Z)$. Los pesos de la regresión están dados por la inversa de $\widehat{\pi}^p\left(\overline{A}_K,\overline{L}_K\right)\equiv\prod_{j=1}^{K}\widehat{\pi}_j\left(\overline{A}_j,\overline{L}_j\right)$ donde, para cada $1\leq j\leq K$, $\widehat{\pi}_j\left(\overline{A}_j,\overline{L}_j\right)$ es el estimador de máxima verosimilitud de $\pi_j\left(\overline{A}_j,\overline{L}_j\right)$ bajo un modelo paramétrico $\mathcal{P}_j$ para la probabilidad de tratamiento π_j. Bajo condiciones de regularidad, el estimador $\widehat{\psi}_{IPTW}$ es consistente y asintóticamente normal (CAN) bajo el modelo que asume (1.57) y que todos los modelos $\mathcal{P}_j,1\leq j\leq K$, se satisfacen. Sin embargo, si alguna de las π_j fue modelada incorrectamente, es posible que este estimador ni siquiera converja en probabilidad a ψ^*.

La siguiente observación sugiere una estrategia alternativa para estimar a ψ^*. Sea

$$\eta_K(\overline{a}_K,\overline{l}_K)\equiv E\left(Y|\overline{A}_K=\overline{a}_K,\overline{L}_K=\overline{l}_K\right)$$

y, para $k=K-1,K-2,\ldots,1$, sea

$$\eta_k(\overline{a}_K,\overline{l}_k)\equiv E\left\{\eta_{k+1}(\overline{a}_K,\overline{L}_{k+1})|\overline{A}_k=\overline{a}_k,\overline{L}_k=\overline{l}_k\right\}.$$

También, definamos

$$\eta_0\left(\overline{a}_K,z\right)\equiv E\left\{\eta_1(\overline{a}_K,L_1)|Z=z\right\}.$$

El Teorema 3.2 de [29], (ver también [33]) implica que, bajo las condiciones de identificabilidad,

$$\eta_k\left(\overline{a}_K,\overline{L}_k\right)=E\left(Y_{\overline{a}_K}|\overline{A}_k=\overline{a}_k,\overline{L}_k\right),\tag{1.58}$$

$k=1,\ldots,K$, y

$$\eta_0\left(\overline{a}_K,Z\right)=E\left(Y_{\overline{a}_K}|Z\right).$$

Por lo tanto, bajo estas condiciones, el modelo (1.57) es equivalente al modelo para los datos observados O definido por

$$\eta_0\left(\overline{a}_K,Z\right)=m\left(\overline{a}_K,Z;\psi^*\right)\text{ para todo }\overline{a}_K.\tag{1.59}$$

Esta observación sugiere postular modelos paramétricos $\mathcal{R}_k$, $\eta_k(\overline{a}_K,\overline{L}_k)=\eta_k(\overline{a}_K,\overline{L}_k;\delta_k^*),1\leq k\leq K$, y calcular un estimador "iterated conditional expectation" (ICE) $\widehat{\psi}_{ICE}$ de ψ^* mediante el siguiente procedimiento. Primero, calculamos una solución $\widehat{\delta}_K$ de

$$\mathbb{P}_n\left[\frac{\partial}{\partial\delta_K}\eta_K(\overline{A}_K,\overline{L}_K;\delta_K)\left\{Y-\eta_K(\overline{A}_K,\overline{L}_K;\delta_K)\right\}\right]=0.$$

Luego, para $k = K-1, K-2, \ldots, 1$ iterativamente calculamos una solución $\widehat{\delta}_k$ de

$$\mathbb{P}_n \left[\sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{\partial}{\partial \delta_k} \eta_k(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \delta_k) \left\{ \eta_{k+1}(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1}; \widehat{\delta}_{k+1}) - \eta_k(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \delta_k) \right\} \right] = 0.$$

Finalmente, $\widehat{\psi}_{ICE}$ resuelve

$$\mathbb{P}_n \left[\sum_{\underline{a}_1 \in \underline{\mathcal{A}}_1} \frac{\partial}{\partial \psi} m(\underline{a}_1, Z; \psi) \left\{ \eta_1(\underline{a}_1, L_1; \widehat{\delta}_1) - m(\underline{a}_1, Z; \psi) \right\} \right] = 0.$$

Bajo condiciones de regularidad, este estimador es CAN bajo el modelo que asume (1.57) y que todos los modelos $\mathcal{R}_j, 1 \leq j \leq K$, se satisfacen. Sin embargo, es posible que este estimador ni siquiera converja en probabilidad a ψ^* si alguna de las η_k fue modelada incorrectamente.

Bang y Robins ([1]) propusieron un estimador doble robusto (DR) de ψ^* que ofrece más oportunidades de tener una inferencia correcta que los estimadores IPTW e ICE al evitar comprometerse con una estrategia de modelado específica. Otros estimadores DR fueron descriptos en [36], [25], [14] y [27]. Estos estimadores requieren que el analista postule dos secuencias de modelos: una secuencia de modelos para las probabilidades de tratamiento $\pi_k, 1 \leq k \leq K$, y otra secuencia de modelos para los funcionales $\eta_k, 1 \leq k \leq K$. El estimador es CAN para ψ^* bajo el modelo unión que asume (1.57) y que o bien la secuencia de modelos para las $\pi_k, 1 \leq k \leq K$ o bien la secuencia de modelos para las $\eta_k, 1 \leq k \leq K$, es correcta, pero no necesariamente ambas.

Un inconveniente de los estimadores DR es que los funcionales en el conjunto $\{\eta_k : 0 \leq k \leq K\}$ no son de variación independiente. Una consecuencia de este hecho es que las restricciones impuestas por un modelo para η_k pueden no ser compatibles con las restricciones impuestas por un modelo para $\eta_{k'}$, $k' \neq k$. Más aún, cualquier modelo para η_k puede no ser compatible con el modelo marginal estructural de interés para $E\left(Y_{\overline{a}_K}|Z\right)$. Esto implica que el analista no puede elegir libremente los modelos para cada η_k porque de hacerlo corre el riesgo de construir un estimador que no sea genuinamente doble robusto.

Según nuestro conocimiento, en el contexto de la estimación doble robusta basada en modelos para la medias contrafactuales η_k, no existe ninguna propuesta general para asegurar la compatibilidad de los modelos. Uno de los objetivos de esta tesis es llenar este vacío en el caso especial en el que la variable de respuesta en continua y no acotada. Extendiendo el trabajo de Robins, Rotnitzky y Scharfstein ([42]), en este capítulo proponemos una nueva clase de modelos anidados flexibles para los funcionales η_k que son siempre compatibles entre sí y compatibles con el MMEM. Estos modelos son lo suficientemente flexibles como para permitir que L_1 sea un modificador del efecto aditivo de $\overline{A}_K$ en Y, es decir que las diferencias $E\left(Y_{\overline{a}_K}|L_1\right) - E\left(Y_{\overline{a}^*_K}|L_1\right)$ dependan de L_1 para $\overline{a}_K \neq \overline{a}^*_K$. Así mismo, estos modelos permiten que $\left(\overline{A}_{j-1}, \overline{L}_j\right)$ sea un modificador del efecto aditivo de $\underline{A}_j$ en Y. Explotando esta propuesta de modelado, construimos un estimador DR, $\widehat{\psi}_{DR}$, basado en modelos compatibles. Más aún, proponemos un estimador múltiple robusto (MR) que otorga aún mayor protección ante la incorrecta especificación de los modelos que el DR. Este estimador, al igual que el estimador DR, requiere que el analista postule dos secuencias de modelos: una secuencia de modelos $\mathcal{P}_k$ para las probabilidades de tratamiento $\pi_k, 1 \leq k \leq K$, y otra secuencia de modelos $\mathcal{R}_k$ para los funcionales $\eta_k, 1 \leq k \leq K$. El estimador MR, $\widehat{\psi}_{MR}$, resuelve la ecuación

$$\mathbb{P}_n \left\{ U_{\widehat{d}}(\psi, \widehat{\eta}, \widehat{\pi}) \right\} = 0 \tag{1.60}$$

donde $\widehat{d}$ es una función estimada a partir de los datos,

$$U_d\left(\psi,\eta,\pi\right)\equiv S_d^K\left(\eta_K,\pi\right)+\sum_{k=1}^{K-1}S_d^k\left(\eta_k,\eta_{k+1},\pi_1,...,\pi_k\right)+S_d^0\left(\psi,\eta_1\right),$$

con

$$S_d^K\left(\eta_K,\pi\right)\equiv\frac{d\left(\overline{A}_K,Z\right)}{\prod_{j=1}^K\pi_j\left(\overline{A}_j,\overline{L}_j\right)}\left\{Y-\eta_K\left(\overline{A}_K,\overline{L}_K\right)\right\},$$

para $k=1,...,K-1$,

$$S_d^k\left(\eta_k,\eta_{k+1},\pi_1,...,\pi_k\right)\equiv\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}}\frac{d(\overline{A}_k,\underline{a}_{k+1},Z)}{\prod_{j=1}^k\pi_j\left(\overline{A}_j,\overline{L}_j\right)}\left\{\eta_{k+1}(\overline{A}_k,\underline{a}_{k+1},\overline{L}_{k+1})-\eta_k(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k)\right\},$$

y

$$S_d^0\left(\psi,\eta_1\right)\equiv\sum_{\underline{a}_1\in\underline{\mathcal{A}}_1}d\left(\underline{a}_1,Z\right)\left\{\eta_1(\underline{a}_1,L_1)-m(\underline{a}_1,Z;\psi)\right\},$$

Además, en la ecuación (1.60), $\widehat{\eta}\equiv(\widehat{\eta}_1,...,\widehat{\eta}_K)$ y $\widehat{\pi}\equiv(\widehat{\pi}_1,...,\widehat{\pi}_K)$ donde $\widehat{\pi}_k$ es un estimador consistente para π_k bajo $\mathcal{P}_k$, $k=1,...,K$, $\widehat{\eta}_K$ es un estimador consistente para η_K bajo $\mathcal{R}_K$ y, para cada $k=1,...,K-1$, $\widehat{\eta}_k$ es un estimador múltiple robusto de η_k en el sentido de que es consistente para η_k bajo el modelo que asume que $\mathcal{R}_k$ se satisface y que, para cada $j\in\{k+1,...,K\}$, alguno de los modelos $\mathcal{R}_j$ ó $\mathcal{P}_j$ también se satisface. De ahora en más, para cualquier funcional χ de la ley de los datos observados O, cualquier modelo paramétrico $\mathcal{G}$ que asume que $\chi=\chi_{\beta^*}$ para algún $\beta^*\in\Gamma$ (con Γ un subconjunto de un espacio euclídeo) y cualquier estimador $\widehat{\beta}$ de β^*, diremos que $\widehat{\chi}\equiv\chi_{\widehat{\beta}}$ es consistente para χ bajo $\mathcal{G}$ si $\widehat{\beta}$ es consistente para β^* bajo $\mathcal{G}$.

El estimador $\widehat{\psi}_{MR}$ es múltiple robusto en el sentido de que, bajo condiciones de regularidad, es CAN para ψ^* bajo el modelo que asume que (1.57) se satisface y que, para cada $j\in\{1,\ldots,K\}$, alguno de los modelos $\mathcal{R}_j$ o $\mathcal{P}_j$ se satisface. El hecho de que los modelos $\mathcal{R}_k$ sean anidados (es decir, para cada $k=1,\ldots,K-1$, $\mathcal{R}_{k+1}$ implica $\mathcal{R}_k$ y $\mathcal{R}_1$ implica (1.57)), implica que $\widehat{\psi}_{MR}$ es CAN para ψ^* bajo el modelo que asume (1.57) y que alguna de las tres condiciones siguientes (i), (ii) o (iii) se satisface : (i) los modelos $\mathcal{R}_1,\ldots,\mathcal{R}_K$ son correctos; (ii) para algún $k\in\{1,\ldots,K-1\}$, los modelos $\mathcal{R}_1,\ldots,\mathcal{R}_k$ y los modelos $\mathcal{P}_{k+1},\ldots,\mathcal{P}_K$ son correctos; (iii) los modelos $\mathcal{P}_1,\ldots,\mathcal{P}_K$ son correctos. Por lo tanto, mientras que $\widehat{\psi}_{DR}$ produce inferencias válidas si (i) o (iii) se satisface, $\widehat{\psi}_{MR}$ también produce inferencias válidas si (ii) se satisface incluso cuando (i) y (iii) no se verifican. Dicho de otro modo, $\widehat{\psi}_{MR}$ ofrece $K+1$ oportunidades para una inferencia correcta, en lugar dos como lo hace $\widehat{\psi}_{DR}$. La múltiple robustez de $\widehat{\psi}_{MR}$ es esencialmente una consecuencia de los siguientes hechos:

(I) Como se demuestra en la Proposición 1 de la sección 1.10, para cualquier $\eta'\equiv(\eta'_1,\ldots,\eta'_K)$ y cualquier $\pi'\equiv(\pi'_1,\ldots,\pi'_K)$, $E\left\{U_d\left(\psi^*,\eta',\pi'\right)\right\}=0$ bajo (1.57) si para cada $k=1,...,K$, o bien η'_k es igual al verdadero η_k o bien π'_k es igual al verdadero π_k.

(II) El estimador $\widehat{\eta}_K$ es consistente para η_K bajo $\mathcal{R}_K$ y, para cada $k=1,...,K-1$, $\widehat{\eta}_k$ es un estimador múltiple robusto de η_k en el sentido mencionado anteriormente.

Por otra parte, para cada $k = 1, ..., K-1$, la estimación múltiple robusta de η_k fue posible también gracias a la Proposición 1. Esto se debe a que (1.58) implica que el modelo para η_k que se deduce de $\mathcal{R}_k$ puede considerarse como un MMEM en un estudio longitudinal con $K-k$ instantes de tiempo, variable de respuesta Y, variables de tratamiento $A_{k+1}, \ldots, A_K$, y con $\left(\overline{A}_k, \overline{L}_k\right)$ en lugar de Z y L_{k+1} en lugar de V.

Por otro lado, la elección particular de la función $\widehat{d}$ en la ecuación de estimación (1.60) y las ecuaciones de estimación particulares utilizadas para estimar a cada η_k, son tales que los primeros K términos del promedio muestral de $U_{\widehat{d}}(\psi, \widehat{\eta}, \widehat{\pi})$ (es decir del lado izquierdo de la ecuación (1.60)) se anulan. De modo que $\widehat{\psi}_{MR}$ de hecho se obtiene, al igual que el estimador ICE, mediante una regresión por el método de mínimos cuadrados no pesados reemplazando a Y por el valor predicho de η_1. Así mismo, como se detalla en la subsección 1.7.2, cada $\widehat{\eta}_k$ se obtiene, al igual que en el método ICE, mediante una regresión por el método de mínimos cuadrados donde la variable de respuesta es el estimador de η_{k+1}, pero a diferencia del método anterior, esta regresión es pesada.

Las secciones 1.3 y 1.4 describen los MMEM y los estimadores existentes respectivamente. En la sección 1.5 proponemos modelos compatibles para la secuencia de medias contrafactuales y en la sección 1.6 presentamos estimadores DR y MR que usan esos modelos para caso en que el número de instantes de tiempo de exposición K es igual a 2. En la sección 1.7, generalizamos nuestra propuesta para el caso en que K es arbitrario. En la sección 1.8, ilustramos nuestros métodos mediante el análisis de un conjunto de datos del "National Heart Lung and Blood Institute Growth and Health Study". En la sección 1.9 presentamos un estudio de simulación. En la sección 1.10, probamos la consistencia y la normalidad asintótica de nuestro estimador MR. Finalmente, en la sección 1.11, generalizamos nuestra propuesta al caso de variables de respuesta repetidas.

Chapter 2

On non-parametric doubly and multiply robust estimation of the g-formula

2.1 Introduction

The goal of this chapter is to investigate and contrast the asymptotic properties of double and multiple robust estimation of the longitudinal g-computation formula parameter (aka g-formula) of Robins ([29]) from n i.i.d. copies of a vector $O = (O_1, \ldots, O_K, L_{K+1})$ where $O_k = (L_k, A_k), k = 1, ..., K, A_k$ is a discrete variable (representing treatment received at time t_k) and L_k is a, possibly multivariate, random vector (representing the data recorded on a subject just prior to receiving treatment A_k).

Letting p denote the density of the law P of O, with respect to some dominating measure, write

$$p(o) = \prod_{k=0}^{K} g_k\left(l_{k+1}|\bar{l}_k, \bar{a}_k\right) \prod_{k=1}^{K} h_k\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right),$$

or for short $p = gh$, where $h_k\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right) \equiv P\left(A_k = a_k|\overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \bar{a}_{k-1}\right)$ and $g_k\left(l_{k+1}|\bar{l}_k, \bar{a}_k\right)$ is (a version of) the conditional density of L_{k+1}. Here and throughout, for $1 \leq k \leq K$ and any $\{v_j\}_{1\leq j\leq K}$, we let $\bar{v}_k \equiv (v_1, \ldots, v_k)$.

The g-computation formula ([29]) is defined as

$$\theta(p) \equiv E_{gh^*}\left\{\kappa\left(\overline{L}_{K+1}\right)\right\}$$

where κ is a given real valued function and $h_k^*\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right)$ is a given, i.e. known, probability mass density for each $k = 1, ..., K$, such that $p^* = gh^*$ is absolutely continuous with respect to $p = gh$ and $E_{gh^*}(\cdot)$ denotes expectation under $p^* = gh^*$. Explicitly,

$$\theta(p) = \int \varphi(o) \prod_{k=0}^{K} g_k\left(l_{k+1}|\bar{l}_k, \bar{a}_k\right) d\mu(o) \tag{2.1}$$

where $\left(\bar{l}_0, \bar{a}_0\right) \equiv$nill and

$$\varphi(o) \equiv \left\{ \prod_{k=1}^{K} h_k^* \left(a_k | \bar{l}_k, \bar{a}_{k-1}\right) \right\} \kappa\left(\bar{l}_{K+1}\right). \tag{2.2}$$

is a known, i.e. specified, function of o.

Special choices of h_k^* yield $\theta(p)$ equal to parameters which are of interest in causal inference and in missing data analysis and which are reviewed in Appendix B.1. A leading special example is when

$$h_k^*\left(a_k | \bar{l}_k, \bar{a}_{k-1}\right) = I_{\left\{a_k^*\right\}}(a_k) \tag{2.3}$$

which, under the assumption of no unmeasured confounding, corresponds to the expectation of a counterfactual response when a particular fixed, i.e. non-dynamic, treatment strategy $A_k = a_k^*, k = 1, ..., K$, is forced in the population ([29], [32] and [33]). To avoid distracting technicalities and alleviate the notation, in this chapter we will focus on this special case, i.e. we will assume that h_k^* is the mass point probability (2.3) . Our results easily generalize to arbitrary $h_k^{*\prime}s$.

For the point mass probability h_k^* of (2.3) the g-computation formula reduces to

$$\theta(p) \equiv E_{g_0}\left[E_{g_1}\left[\ldots E_{g_{K-1}}\left[E_{g_K}\left\{\kappa\left(\overline{L}_{K+1}\right) | \overline{A}_K = \overline{a}_K^*, \overline{L}_K\right\} \middle| \overline{A}_{K-1} = \overline{a}_{K-1}^*, \overline{L}_{K-1}\right] \ldots | A_1 = a_1^*, L_1\right]\right]$$

where $E_{g_k}(\cdot|\cdot)$ denotes conditional expectation under g_k. The expression makes it clear that $\theta(p)$ depends on p only through g; more precisely only through the marginal law g_0 of L_1, through the conditional expectation

$$\eta_K\left(\overline{L}_K\right) \equiv E_{g_K}\left\{\kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_K = \overline{a}_K^*, \overline{L}_K\right\}$$

and through the iterated conditional expectations defined sequentially for $k = K-1, \ldots, 1$, as

$$\eta_k\left(\overline{L}_k\right) \equiv E_{g_k}\left\{\eta_{k+1}\left(\overline{L}_{k+1}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k\right\}. \tag{2.4}$$

So, letting $\eta \equiv (g_0, \eta_1, ..., \eta_k)$, from now on, we will denote $\theta(p)$ as $\theta(\eta)$.

A natural first choice for estimating $\theta(\eta)$ is with the so called plug-in estimator $\theta(\widehat{\eta})$ where $\widehat{\eta} \equiv (\widehat{g}_0, \widehat{\eta}_1, ..., \widehat{\eta}_K)$ and

1. $\widehat{g}_0 = d\widehat{G}_0$ where $\widehat{G}_0$ is the empirical law of L_1,
2. $\widehat{\eta}_K$ is a preferred estimator of the conditional mean of $\kappa\left(\overline{L}_{K+1}\right)$ given $\overline{L}_K$ among subjects with $\overline{A}_K = \overline{a}_K^*$, and,
3. sequentially for $k = K-1, \ldots, 1$, $\widehat{\eta}_k\left(\overline{L}_k\right)$ is a preferred estimator of the conditional mean of $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ given $\overline{L}_k$ among subjects with $\overline{A}_k = \overline{a}_k^*$, obtained by pretending that the unknown "outcome" $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ is equal to its estimator $\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)$.

When the estimators $\widehat{\eta}_k\left(\overline{L}_k\right), k = K+1, ..., 1$, are computed under parametric regression models, the plug in estimator is known as the parametric g-formula estimator ([37]). Under regularity conditions, the parametric g-formula estimator $\theta(\widehat{\eta})$ satisfies that $\sqrt{n}\{\theta(\widehat{\eta}) - \theta(\eta)\}$ converges to a mean zero Normal distribution if the parametric regression models assumed for each $\eta_k, k = K, ..., 1$ are all correct. However, $\theta(\widehat{\eta})$ is not even consistent if one of these models is misspecified. A well known strategy that yields an estimator that confers some protection against mispecification of the

models for η_k is to add to $\theta(\widehat{\eta})$ the quantity $\mathbb{P}_n\left[M\left(\widehat{h},\widehat{\eta}\right)\right]$ where $\widehat{h} \equiv \left(\widehat{h}_1, ..., \widehat{h}_K\right)$ is a vector of preferred estimators of the $h_k's$ and for any $h^\dagger \equiv \left(h_1^\dagger, ..., h_K^\dagger\right)$ and $\eta^\dagger \equiv \left(g_0^\dagger, \eta_1^\dagger, ..., \eta_K^\dagger\right)$,

$$M\left(h^\dagger, \eta^\dagger\right) \equiv m\left(O; h^\dagger, \eta^\dagger\right) \tag{2.5}$$
$$\equiv \sum_{k=1}^{K}\left\{\prod_{j=1}^{k}\frac{I_{\{a_j^*\}}(A_j)}{h_j^\dagger\left(A_j|\overline{A}_{j-1}, \overline{L}_j\right)}\right\}\left\{\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right) - \eta_k^\dagger\left(\overline{L}_k\right)\right\}$$

with $\eta_{K+1}^\dagger\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$ ([25], [1]). Note that $M\left(h^\dagger, \eta^\dagger\right)$ does not depend on $g_0^\dagger$ and that when $h^\dagger = h$ and $\eta_k^\dagger = \eta_k$ for all $k = 1, ..., K$, i.e. when $\eta_k^\dagger$ is equal to the true iterated conditional expectation under $p = gh$, then $E_p\{M(h, \eta)\} = 0$, where here and throughout $E_p(\cdot)$ denotes expectation under the law $p = gh$.

This strategy yields the estimator

$$\widehat{\theta} \equiv \theta(\widehat{\eta}) + \mathbb{P}_n\left[M\left(\widehat{h},\widehat{\eta}\right)\right].$$

It is well known that the random variable

$$IF(h, \eta) \equiv M(h, \eta) + \eta_1(L_1) - \theta(\eta)$$

is the, unique, influence function of the parameter $\theta(\eta)$ under a non-parametric model for the law P of O ([41], [40]). Then, since $\mathbb{P}_n[\widehat{\eta}_1(L_1)] = \theta(\widehat{\eta})$, we observe that the estimator $\widehat{\theta}$ is equal to the semiparametric efficient one step estimator under a non-parametric model, i.e.

$$\widehat{\theta} \equiv \theta(\widehat{\eta}) + \mathbb{P}_n\left[IF\left(\widehat{h},\widehat{\eta}\right)\right].$$

Another important algebraic identity shows that the one step estimator $\widehat{\theta}$ is, in fact, the so-called Augmented Inverse Probability Weighted (AIPW) estimator, familiar in the missing data and causal inference literature ([44]). Specifically, some algebra gives

$$\eta_1^\dagger(L_1) + M\left(h^\dagger, \eta^\dagger\right) = \left\{\prod_{j=1}^{K}\frac{I_{\{a_j^*\}}(A_j)}{h_j^\dagger\left(A_j|\overline{A}_{j-1}, \overline{L}_j\right)}\right\}\kappa\left(\overline{L}_{K+1}\right)$$
$$- \sum_{k=1}^{K}\left[\prod_{j=1}^{k-1}\frac{I_{\{a_j^*\}}(A_j)}{h_j^\dagger\left(A_j|\overline{A}_{j-1}, \overline{L}_j\right)}\left\{\left[\frac{I_{\{a_k^*\}}(A_k)}{h_k^\dagger\left(A_k|\overline{A}_{k-1}, \overline{L}_k\right)}\right] - 1\right\}\eta_k^\dagger\left(\overline{L}_k\right)\right]$$

where $\prod_{j=1}^{0}(\cdot) \equiv 1$. The AIPW estimator is precisely the sample average of the right hand side of the equality when, for each k, $h_k^\dagger$ and $\eta_k^\dagger$ are replaced by estimators $\widehat{h}_k, \widehat{\eta}_k$.

Under regularity conditions, if the $\eta_k's$ and the $h_k's$ are estimated under parametric, generalized linear, regression models, e.g. the $h_k's$ are estimated under logistic regression models if the $A_k's$ are binary, then $\sqrt{n}\left\{\widehat{\theta} - \theta(\eta)\right\}$ converges to a mean zero Normal distribution if either (i) the parametric regression models assumed for each $\eta_k, k = K, ..., 1$ are all correct, or (ii) the parametric

models assumed for each h_k $k = K, ..., 1$ are all correct, but not necessarily both (i) and (ii) hold simultaneously. This property, whose heuristic proof is reviewed in Section 2.4, is referred to as double-robustness, since $\widehat{\theta}$ confers the data analyst two opportunities of obtaining correct inferences about $\theta(\eta)$, one by modeling the $\eta_k's$ correctly and another by modelling the $h_k's$ correctly.

The one step estimator becomes specially attractive when the $\widehat{\eta}_k, k = 1, ..., K$, in steps 2 and 3 above are estimated under a non-parametric model defined solely by smoothness or sparsity assumptions. In such case, the plug in estimator $\theta(\widehat{\eta})$ may not be a useful estimation option as it may not even converge at rate $\sqrt{n}$ ([31]). In contrast, provided one uses an appropriate sample splitting approach explained in detail in Section 2.5 below, the one step estimator $\widehat{\theta}$ that uses nonparametric estimators $\widehat{\eta}_k$ and $\widehat{h}_k$ is $\sqrt{n}-$consistent for $\theta(\eta)$ and asymptotically normal with mean zero provided $\widehat{\eta}_k$ and $\widehat{h}_k$ converge sufficiently fast to η_k and h_k ([39], [8], [11], [57]). Moreover, when $K = 1$, it has been well established that convergence at rate $\sqrt{n}$ of the one step estimator can be obtained by trading off slower rates of convergence for the estimator of one of the nuisance functions, η_1 or h_1, with faster rates for the estimator of the other nuisance ([39], [31], [8], [11], [57]). In contrast, little has been reported in the literature about the specific trade offs in rates of convergence for estimation of the nuisances η and h conferred by the one step estimator when $K > 1$, the exception being [31]. One goal of this chapter is to study which trade offs, if any, in convergence rates of the nuisance function estimators are conferred by $\widehat{\theta}$ when $K > 1$.

Recently, a number of articles have pointed out ([53], [24]) that when the $\eta_k's$ are estimated under parametric models, one can obtain estimators that confer even more protection against model misspecification than the preceding one step estimator. The following modification to step 3 above yields one such estimator:

3_MR. sequentially for $k = K-1, \ldots, 1$, compute $\widehat{\eta}_{k,MR}\left(\overline{L}_k\right)$ as a preferred estimator of the conditional mean of $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ given $\overline{L}_k$ among subjects with $\overline{A}_k = \overline{a}_k^*$, obtained by pretending that the unknown "outcome" $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ is equal to the pseudo-outcome

$$\widehat{Q}_{k+1} \equiv \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right) + \frac{I_{\left\{a_{k+1}^*\right\}}\left(A_{k+1}\right)}{\widehat{h}_{k+1}\left(a_{k+1}^* | \overline{a}_k^*, \overline{L}_{k+1}\right)} \left\{\widehat{Q}_{k+2} - \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right)\right\}$$

where $\widehat{Q}_{K+1} \equiv \kappa\left(\overline{L}_{K+1}\right)$.

The papers of [46] and [47] have defined and advocated the use of the pseudo-outcomes $\widehat{Q}_{k+1}$ to produce double robust estimators. It was not until the article of [53] that it was noticed that indeed using these pseudo outcomes produces estimators that confer further protection against model misspecification.

Denote the one step estimator that uses $\widehat{\eta}_{k,MR}$ instead of $\widehat{\eta}_k$ (as in step 3 above) with $\widehat{\theta}_{MR}$, i.e.

$$\widehat{\theta}_{MR} \equiv \theta\left(\widehat{\eta}_{MR}\right) + \mathbb{P}_n\left[M\left(\widehat{h}, \widehat{\eta}_{MR}\right)\right]$$

where $\widehat{\eta}_{MR} \equiv \left(\widehat{g}_0, \widehat{\eta}_{1,MR}, ..., \widehat{\eta}_{K-1,MR}, \widehat{\eta}_K\right)$.

It can be shown that, when the models used to compute $\widehat{\eta}_{k,MR}$, and those used to compute $\widehat{h}_k$, $k = 1, ..., K$, are parametric, $\widehat{\theta}_{MR}$ in fact agrees with the estimator of the coefficient associated with $\overline{a}_K^*$ in the MSMM of Chapter 1 of this thesis, in the special case in which the baseline covariates (denoted in that chapter as Z) are nill, the MSMM is saturated in $\overline{a}_K$ and the functions

$g_k\left(\overline{a}_K, \overline{L}_{j-1}; \gamma_k\right), k = 1, ..., K$ (used in Subsection 1.7.1 to model the $\rho_k' s$) are also saturated in $\overline{a}_K$. The former is true if, in addition, (1) the models for the $\eta_k\left(\overline{L}_k\right)' s$ are those implied by the compatible parametric models for the $\eta_k\left(\overline{a}_K^*, \overline{L}_k\right)' s$ defined in Section 1.7.1 of that chapter and (2) the parameters indexing the models for the $\eta_k' s$ are computed via weighted regression as in that chapter. Then, it follows from that chapter that, under regularity conditions, $\sqrt{n}\left\{\widehat{\theta}_{MR} - \theta(\eta)\right\}$ converges to a mean zero Normal distribution when, for each $k = 1, ..., K$, either the parametric model for h_k used to compute $\widehat{h}_k$ is correct or the parametric model for η_k used to compute $\widehat{\eta}_{k,MR}$ is correct. This property, whose heuristic proof is provided in Section 2.4.2 for arbitrary parametric models and arbitrary parameter estimators, not just those used in Chapter 1, is known as multiple robustness ([53], [24]) and has been called sequential double robustness in ([22]). It implies that $\widehat{\theta}_{MR}$ confers more protection to model misspecification than $\widehat{\theta}$ because it ensures valid inferences not only when all the $\eta_k' s$ are correctly modeled, or all the $h_k' s$ are correctly modeled, but also when a subset of the $\eta_k' s$ are correctly modeled so long as for the $k' s$ for which the $\eta_k' s$ are incorrectly modeled, the $h_k' s$ are correctly modeled.

Whereas the robustness benefits of $\widehat{\theta}_{MR}$ over and above those of $\widehat{\theta}$ appear to be well understood and documented in the literature when the nuisance functions h and η are estimated under parametric models, the same is not true for the case in which h and η are estimated under non-parametric models. Thus, a second goal of this chapter is to investigate, when h and η are estimated under non-parametric models defined solely by smoothness or sparsity assumptions, whether $\widehat{\eta}_{k,MR}$ confers additional trade offs in the requirements on the rates of convergence of the nuisance parameter estimators over and above those already conferred by the one step estimator that uses $\widehat{\eta}_k$.

To be concrete about the contributions of this chapter, in order to ensure $\sqrt{n}-$ consistent estimation of $\theta(\eta)$, we start by writing a decomposition of the centered difference between the one step estimator and the true parameter which is typically used when analyzing the asymptotic properties of the one step estimator. In what follows, $h^\dagger \equiv \left(h_1^\dagger, ..., h_K^\dagger\right)$ and $\eta^\dagger \equiv \left(g_0^\dagger, \eta_1^\dagger, ..., \eta_K^\dagger\right)$ stand for placeholders for arbitrary estimators of h and η and

$$\widehat{\theta}^\dagger \equiv \theta\left(\eta^\dagger\right) + \mathbb{P}_n\left[M\left(h^\dagger, \eta^\dagger\right)\right].$$

Note that, when $g_0^\dagger = \widehat{g}_0$, $\theta\left(\eta^\dagger\right) = \mathbb{P}_n\left\{\eta_1^\dagger(L_1)\right\}$, so that

$$\begin{aligned}\widehat{\theta}^\dagger &\equiv \theta\left(\eta^\dagger\right) + \mathbb{P}_n\left[M\left(h^\dagger, \eta^\dagger\right)\right] \\ &= \mathbb{P}_n\left[\eta_1^\dagger(L_1) + M\left(h^\dagger, \eta^\dagger\right)\right] \\ &\equiv \mathbb{P}_n\left[Q\left(h^\dagger, \eta^\dagger\right)\right]\end{aligned}$$

where

$$\begin{aligned}Q\left(h^\dagger, \eta^\dagger\right) &\equiv q\left(O; h^\dagger, \eta^\dagger\right) \\ &\equiv \eta_1^\dagger(L_1) + M\left(h^\dagger, \eta^\dagger\right)\end{aligned}$$

Note that $Q\left(h^\dagger, \eta^\dagger\right)$ does not depend on $g_0^\dagger$.

When $h^\dagger = \widehat{h}$ and $\eta^\dagger = \widehat{\eta}$ then $\widehat{\theta}^\dagger$ coincides with the doubly robust estimator $\widehat{\theta}$ and when $h^\dagger = \widehat{h}$

and $\eta^{\dagger} = \widehat{\eta}_{MR}$ then $\widehat{\theta}^{\dagger}$ is equal to the multiply robust estimator $\widehat{\theta}_{MR}$. Write

$$\begin{aligned}\sqrt{n}\left\{\widehat{\theta}^{\dagger} - \theta(\eta)\right\} &= \mathbb{G}_n\left\{Q(h,\eta)\right\} + \mathbb{G}_n\left\{Q\left(h^{\dagger},\eta^{\dagger}\right) - Q(h,\eta)\right\} \\ &\quad + \sqrt{n}\left[E_p\left\{Q\left(h^{\dagger},\eta^{\dagger}\right)\right\} - \theta(\eta)\right] \\ &= \Upsilon_{1,n} + \Upsilon_{2,n} + \Upsilon_{3,n}\end{aligned}$$

where

$$E_p\left\{Q\left(h^{\dagger},\eta^{\dagger}\right)\right\} \equiv \int q\left(o;h^{\dagger},\eta^{\dagger}\right) dP(o)$$

$$\mathbb{G}_n(\cdot) \equiv \sqrt{n}\mathbb{P}_n\left\{\cdot - E_p(\cdot)\right\}$$

is the centered empirical process, $\Upsilon_{1,n} \equiv \mathbb{G}_n\left\{Q(h,\eta)\right\}, \Upsilon_{2,n} \equiv \mathbb{G}_n\left\{Q\left(h^{\dagger},\eta^{\dagger}\right) - Q(h,\eta)\right\}$ and $\Upsilon_{3,n} \equiv \sqrt{n}\left\{E_p\left\{Q\left(h^{\dagger},\eta^{\dagger}\right)\right\} - \theta(\eta)\right\}$

By the Central Limit Theorem, the term $\Upsilon_{1,n}$ converges to a mean 0 normal distribution provided $Var_{gh}\left[Q(h,\eta)\right] < \infty$.

The term $\Upsilon_{2,n}$ is the difference of two centered empirical processes, one evaluated at (h,η) and the other evaluated at its estimator $\left(h^{\dagger},\eta^{\dagger}\right)$. If sufficient smoothness or sparsity conditions are placed in the functions (h,η), then one should be able to construct estimators $\left(h^{\dagger},\eta^{\dagger}\right)$ converging to (h,η) at sufficiently fast rate, so that $\Upsilon_{2,n}$ would be $o_p(1)$. One can make this term $o_p(1)$ under very mild regularity conditions, even without restrictions on smoothness or sparsity by employing the following strategy known as cross-fitting. First split the sample into a finite number, say $\mathbf{U}$, of equal, or nearly equal, sized subsamples, designating one of them as the "main estimation subsample" and the remaining as the "nuisance estimation subsamples". Next compute $\eta^{\dagger}$ and $h^{\dagger}$ using data from the union of the nuisance estimation subsamples and compute the one step estimator from the main estimation subsample replacing the unknown η and h with their estimators computed from the union of the nuisance estimation subsamples. Next, repeat the procedure, $\mathbf{U}-1$ times, each time designating a distinct subsample as the main estimation subsample. Finally, compute the estimator $\widehat{\theta}^{\dagger}$of $\theta(\eta)$ as the average of the $\mathbf{U}$ one step estimators. The use of cross-fitting to avoid imposing conditions on the model complexity has been noticed long ago ([49], Chapter 25 of [58]) but has been emphasized and advocated only lately (see, for instance, [45], [63] and [8]). In Section 2.5 we describe the precise steps of this procedure.

Assuming cross-fitting has been employed, then $\sqrt{n}\left\{\widehat{\theta}^{\dagger} - \theta(\eta)\right\}$ will be bounded in probability provided

$$\Upsilon_{3,n} \equiv \sqrt{n}\left[E_p\left\{Q\left(h^{\dagger},\eta^{\dagger}\right)\right\} - \theta(\eta)\right] \tag{2.6}$$

is $O_p(1)$. Furthermore, $\sqrt{n}\left\{\widehat{\theta}^{\dagger} - \theta(\eta)\right\}$ will be asymptotically normal with mean zero and variance $Var_{gh}\left\{Q(h,\eta)\right\}$ if this variance is finite and $\Upsilon_{3,n} = o_p(1)$. Thus, the term (2.6) which is often referred to as the "drift" term or the "bias" term ([58]), is crucial in determining the asymptotic distribution of $\widehat{\theta}^{\dagger}$.

The key contribution of this chapter is the derivation of distinct expressions for the drift term. To our knowledge, none of the expressions that will be derived in this chapter have been reported earlier in the literature. Each of these expressions helps visualize the general structure of the robustness properties -in terms of trade offs of the rates of convergence of the non-parametric estimators of η and h- conferred by the doubly robust estimator $\widehat{\theta}$ and the multiply robust estimator $\widehat{\theta}_{MR}$. For the special case in which the estimators of each η_k are linear in the outcome, we will additionally

provide a further expression for the drift that will allow us to investigate in detail and compare the asymptotic behavior of $\widehat{\theta}$ and $\widehat{\theta}_{MR}$ when the η_k are estimated by series estimation.

2.2 Notation

In this section we summarize the notation that will be used throughout the chapter.

For any $k \in \mathbb{N}, [k] \equiv \{1, ..., k\}$.

For $1 \leq j \leq k \leq K$ and any $\{v_j\}_{1 \leq j \leq K}$, we let

$$\overline{v}_k \equiv (v_1, \ldots, v_k), \underline{v}_k \equiv (v_k, \ldots, v_K) \text{ and } \overline{v}_j^k \equiv (v_j, \ldots, v_k)$$

As in the introduction, we let $h^\dagger \equiv \left(h_1^\dagger, ..., h_K^\dagger\right)$ and $\eta^\dagger \equiv \left(g_0^\dagger, \eta_1^\dagger, ..., \eta_K^\dagger\right)$ stand for placeholders for arbitrary estimators of h and η. In a slight abuse of notation, we write for any $k \in [K]$

$$h_k^\dagger\left(\overline{L}_k\right) \equiv h_k^\dagger\left(a_k^* | \overline{L}_k, \overline{a}_{k-1}^*\right)$$

Furthermore,

$$\eta_{K+1}^\dagger\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$$

and, for $k \in [K]$,

$$\underline{h}_k^\dagger \equiv \overline{h}_k^{\dagger K} \equiv \left(h_k^\dagger, ..., h_K^\dagger\right) \text{ and } \underline{\eta}_k^\dagger \equiv \left(\eta_k^\dagger, ..., \eta_K^\dagger\right).$$

For $k \in [K]$ and $j \in [k]$, we let

$$I_k \equiv I_{\left\{a_k^*\right\}}\left(A_k\right), \ \overline{I}_k \equiv \prod_{r=1}^{k} I_{\{a_r^*\}}\left(A_r\right) \text{ and } \overline{I}_j^k \equiv \prod_{r=j}^{k} I_{\{a_r^*\}}\left(A_r\right),$$

$$\pi^{\dagger k} \equiv \prod_{r=1}^{k} h_r^\dagger\left(\overline{L}_r\right) \text{ and } \pi_j^{\dagger k} \equiv \prod_{r=j}^{k} h_r^\dagger\left(\overline{L}_r\right),$$

$$\pi^k \equiv \prod_{r=1}^{k} h_r\left(\overline{L}_r\right) \text{ and } \pi_j^k \equiv \prod_{r=j}^{k} h_r\left(\overline{L}_r\right)$$

and for any $j \in [K]$, $\overline{I}_{j+1}^j \equiv 1$, $\prod_{r=j+1}^{j}(\cdot) \equiv 1$ and $\sum_{r=j+1}^{j}(\cdot) \equiv 0$.

For any collection of sets $\{\mathcal{C}_j\}_{1 \leq j \leq K}$ and any $k \in [K]$, we write

$$\overline{\mathcal{C}}_k = \mathcal{C}_1 \times \cdots \times \mathcal{C}_k \text{ and } \underline{\mathcal{C}}_k = \mathcal{C}_k \times \cdots \times \mathcal{C}_K$$

For any vector $v = [v_i]_{1 \leq i \leq p} \in R^p$, $\|v\|$ denotes its Euclidean norm $\left(\sum_i v_i^2\right)^{1/2}$.

For any matrix $A = [a_{ij}]_{1 \leq i \leq p, 1 \leq j \leq p} \in R^{p \times q}$, $\|A\|$ denotes its Euclidean norm $\sup\{\|Av\| : v \in \mathbb{R}^p \text{ with } \|v\| = 1\}$.

For any function $f : \mathcal{X} \subseteq R^p \to R^q$, $\|f\|_\infty$ denotes $\sup_{x \in \mathcal{X}} \|f(x)\|$.

We use the notation $a_n \lesssim b_n$ to denote $a_n \leq cb_n$ for some constant $c > 0$ and $a_n \lesssim_P b_n$ to denote $a_n = O_p(b_n)$. Moreover, we say that $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

2.3 Parametric vs non-parametric estimation of h and η

For ease of reference in the rest of the chapter, it will be convenient to define here what we will mean by parametric and non-parametric estimation of the functions h_k and η_k.

Suppose that we postulate working models $h_{k,\nu_k}\left(\overline{L}_k\right)$ for $h_k\left(\overline{L}_k\right) \equiv P\left(A_k = a_k^* | \overline{A}_{k-1} = \overline{a}_{k-1}^*, \overline{L}_k\right)$ and $\eta_{k,\psi_k}\left(\overline{L}_k\right)$ for $\eta_k\left(\overline{L}_k\right) \equiv E\left[\eta_{k+1}\left(\overline{L}_{k+1}\right) | \overline{A}_k = \overline{a}_k^*, \overline{L}_k\right]$ indexed by finite dimensional, i.e. Euclidean, parameters. For instance, we might postulate the logistic regression model for $h_k\left(\overline{L}_k\right)$:

$$h_{k,\nu_k}\left(\overline{L}_k\right) = \frac{\exp\left\{\nu_k' \phi_k\left(\overline{L}_k\right)\right\}}{1 + \exp\left\{\nu_k' \phi_k\left(\overline{L}_k\right)\right\}} \tag{2.7}$$

and, for η_k we might postulate the model:

$$\eta_{k,\psi_k}\left(\overline{L}_k\right) = \Psi\left\{\psi_k' \phi_k\left(\overline{L}_k\right)\right\} \tag{2.8}$$

where $\phi_k\left(\overline{L}_k\right) \equiv \left(\phi_{k,1}\left(\overline{L}_k\right), \ldots, \phi_{k,m_k}\left(\overline{L}_k\right)\right)'$ is some given vector function of $\overline{L}_k$ of dimension m_k and $\Psi(\cdot)$ is a given link function -for example $\Psi(u) = u$ if $\kappa\left(\overline{L}_{K+1}\right)$ is a continuous random variable. Here, we assume the same transformation $\phi_k\left(\overline{L}_k\right)$ of $\overline{L}_k$ is used in both working models to simplify the presentation in this section only.

If we assume that the dimension m_k does not change with the sample size n, then when the working models are correct, under regularity conditions on the smoothness of the maps $\nu_k \to h_{k,\nu_k}\left(\overline{L}_k\right)$ and $\psi_k \to \eta_{k,\psi_k}\left(\overline{L}_k\right)$, on the moments of $\phi_k\left(\overline{L}_k\right)$ and on the error variance $E_p\left[\left\{\eta_{k+1}\left(\overline{L}_{k+1}\right) - \eta_k\left(\overline{L}_k\right)\right\}^2 | \overline{A}_k = \overline{a}_k^*, \overline{L}_k\right]$ it is possible to find estimators $\widehat{\nu}_k$ and $\widehat{\psi}_k$, such that the functions $\widehat{h}_k(\cdot) = h_{k,\widehat{\nu}_k}(\cdot)$ and $\widehat{\eta}_k(\cdot) = \eta_{k,\widehat{\psi}_k}(\cdot)$ satisfy for all $k = 1, ..., K$

$$\left\|\widehat{h}_k - h_k\right\|_{L_2(P)} \equiv \left[\int \left\{\widehat{h}_k\left(\overline{l}_k\right) - h_k\left(\overline{l}_k\right)\right\}^2 dP\left(\overline{l}_k\right)\right]^{1/2} = O_p\left(n^{-1/2}\right) \tag{2.9}$$

and

$$\left\|\widehat{\eta}_k - \eta_k\right\|_{L_2(P)} \equiv \left[\int \left\{\widehat{\eta}_k\left(\overline{l}_k\right) - \eta_k\left(\overline{l}_k\right)\right\}^2 dP\left(\overline{l}_k\right)\right]^{1/2} = O_p\left(n^{-1/2}\right) \tag{2.10}$$

For instance, since the true values of ν_k and ψ_k solve the population moment equations

$$E_p\left[t_k\left(\overline{L}_k; \nu_k\right) \overline{I}_{k-1}\left\{I_k - h_{k,\nu_k}\left(\overline{L}_k\right)\right\}\right] = 0, k = 1, ..., K$$

and

$$E_p\left[t_k\left(\overline{L}_k; \psi_k\right) \overline{I}_k\left\{\eta_{k+1,\psi_{k+1}}\left(\overline{L}_{k+1}\right) - \eta_{k,\psi_k}\left(\overline{L}_k\right)\right\}\right] = 0, k = 1, ..., K$$

for any given arbitrary $m_k \times 1$ vector function $t_k(\cdot;\cdot)$, then solving the empirical version of these moment equations, i.e. solving

$$\mathbb{P}_n\left[t_k\left(\overline{L}_k; \nu_k\right) \overline{I}_{k-1}\left\{I_k - h_{k,\nu_k}\left(\overline{L}_k\right)\right\}\right] = 0, k = 1, ..., K \tag{2.11}$$

and recursively solving for $k = K, ..., 1$,

$$\mathbb{P}_n \left[t_k \left(\overline{L}_k; \psi_k \right) \overline{I}_k \left\{ \eta_{k+1,\psi_{k+1}} \left(\overline{L}_{k+1} \right) - \eta_{k,\psi_k} \left(\overline{L}_k \right) \right\} \right] = 0, k = 1, ..., K \tag{2.12}$$

where $\eta_{K+1,\psi_{K+1}} \left(\overline{L}_{K+1} \right) \equiv \kappa \left(\overline{L}_{K+1} \right)$, results in estimators $\widehat{\nu}_k$ and $\widehat{\psi}_k$, $k = 1, ..., K$, that, under regularity conditions, satisfy that $\sqrt{n} \left\{ \widehat{\nu}_k - \nu_k \right\}$ and $\sqrt{n} \left\{ \widehat{\psi}_k - \psi_k \right\}$ converge to mean zero Normal random variables (Section 5.3 of [58]). This, in turn, implies that under regularity conditions, (2.9) and (2.10) hold.

In this chapter, we will say that h_k (likewise η_k) is estimated *parametrically* or at *parametric rates* if the estimator $\widehat{h}_k$ (likewise $\widehat{\eta}_k$) satisfies (2.9) (likewise (2.10)) and we will refer to $\widehat{h}_k$ (likewise $\widehat{\eta}_k$) as a *parametric estimator* of h_k (likewise of η_k). In an admittedly abuse of terminology, we will refer to any estimator $\widehat{h}_k$ (likewise $\widehat{\eta}_k$) that does not meet the condition (2.9) (likewise (2.10)) as a *non-parametric estimator* of h_k (likewise of η_k).

Admittedly, the appellative non-parametric is an abuse of terminology because even when the *parametric* models (2.7) and (2.8) are correct, $\widehat{h}_k (\cdot) = h_{k,\widehat{\nu}_k} (\cdot)$ may fail to satisfy (2.9) and $\widehat{\eta}_k (\cdot) = \eta_{k,\widehat{\psi}_k} (\cdot)$ may fail to satisfy (2.10) under a triangular array asymptotics in which model (2.8) is correct but the dimension m_k increases with n. For instance, the condition (2.10) fails even in the very simple case in which $K = 1$, model (2.8) holds with $\Psi (u) = u$, L_1 is a $Unif (0, 1)$ random variable, $\phi_1 (L_1) = (\phi_{1,1} (L_1), \ldots, \phi_{1,m_1} (L_1))$ are the first m_1 elements of the Fourier basis $(1, \cos (2\pi_j L_1), \sin (2\pi_j L_1)), j = 1, 2, ...$ and $\widehat{\psi}_1$ is the ordinary least squares estimator of ψ_1, i.e. solving (2.12) with $t_1 (L_1; \psi_1) = \phi_1 (L_1)$. It is well known (see, for example, [2]) that in such case, $\widehat{\eta}_1 (\cdot) = \eta_{1,\widehat{\psi}_1} (\cdot)$ satisfies $\|\widehat{\eta}_1 - \eta_1\|_{L_2(P)} = O_p \left(\sqrt{\frac{m_1}{n}} \right)$ but does not satisfy $\|\widehat{\eta}_1 - \eta_1\|_{L_2(P)} = o_p \left(\sqrt{\frac{m_1}{n}} \right)$. So, if m_1 increases with n and $\frac{m_1}{n} = o(1)$, then $\|\widehat{\eta}_1 - \eta_1\|_{L_2(P)}$ converges to 0 in probability but not at the parametric rate $O_p \left(n^{-1/2} \right)$.

As another example, suppose again that $K = 1$, model (2.8) holds with $\Psi (u) = u$ but now with $\phi_1 (L_1) = L_1$ where L_1 is an $m_1 \times 1$ vector. Consider the Lasso estimator

$$\widehat{\psi}_1^{LASSO} = \arg\min_{\psi_1} \left\{ \mathbb{P}_n \left[\left\{ \kappa \left(\overline{L}_2 \right) - \psi_1' L_1 \right\}^2 \right] + \lambda \sum_{j=1}^{m_1} |\psi_{1,j}| \right\}$$

where λ is a tunning parameter. Letting $\|\psi_1\|_0 \equiv \# \{ j : \psi_{1,j} \neq 0 \}$ it is well known that when $\lambda \asymp \sqrt{\frac{\log(m_1)}{n}}$ and $\|\psi_1\|_0 \leq s_1$ then under regularity conditions, $\widehat{\eta}_1^{LASSO} (\cdot) = \eta_{1,\widehat{\psi}_1^{LASSO}} (\cdot)$ satisfies $\left\|\widehat{\eta}_1^{LASSO} - \eta_1\right\|_{L_2(P)} = O_p \left(\sqrt{\frac{s_1 \log(m_1)}{n}} \right)$ but $\left\|\widehat{\eta}_1^{LASSO} - \eta_1\right\|_{L_2(P)} = o_p \left(\sqrt{\frac{s_1 \log(m_1)}{n}} \right)$ does not hold (see, for example, [2] and [4]). So, when $\frac{s_1 \log(m_1)}{n} = o(1)$ but s_1 and m_1 grow with n, $\left\|\widehat{\eta}_1^{LASSO} - \eta_1\right\|_{L_2(P)}$ converges to 0 but not at the parametric rate $O \left(n^{-1/2} \right)$.

The appellative non-parametric applies more generally to *any* procedure yielding an estimator $\widehat{\eta}_k$ ($\widehat{h}_k$) that does not meet the condition (2.10) ((2.9)) under the assumed model for η_k (h_k). For instance, the appellative applies to *any* estimator of η_1, under a model that assumes only conditions on the smoothness of η_1 because in such case it is well known that there exists no estimator $\widehat{\eta}_1$ that converges at the parametric rate $O_p \left(n^{-1/2} \right)$, i.e. that satisfies $\|\widehat{\eta}_1 - \eta_1\|_{L_2(P)} = O_p \left(n^{-1/2} \right)$ ([51]). We will also refer to estimators of the nuisance functions h_k and η_k obtained by modern machine learning algorithms as *non-parametric*. These include algorithms for which results on their $L_2 (P)$

convergence rates have only recently started to become available in the literature such as neural networks ([12]) and boosting ([23]).

2.4 The expressions for the drift

In this section we derive, as anticipated in the introduction, different expressions for the drift term

$$\Upsilon_{3,n} \equiv \sqrt{n}\left[E_p\left\{Q\left(h^{\dagger},\eta^{\dagger}\right)\right\}-\theta\left(\eta\right)\right]$$

Although $h^{\dagger} \equiv \left(h_1^{\dagger},...,h_K^{\dagger}\right)$ and $\eta^{\dagger} \equiv \left(g_0^{\dagger},\eta_1^{\dagger},...,\eta_K^{\dagger}\right)$ are placeholders for estimators of the unknown nuisance functions $h \equiv (h_1,...,h_K)$ and $\eta \equiv (g_0,\eta_1,...,\eta_K)$, nevertheless in the expectation that appears in the expression for $\Upsilon_{3,n}$, the functions $\left(h^{\dagger},\eta^{\dagger}\right)$ are regarded as fixed and known, since recall that

$$E_p\left\{Q\left(h^{\dagger},\eta^{\dagger}\right)\right\} \equiv \int q\left(o;h^{\dagger},\eta^{\dagger}\right)dP\left(o\right)$$

Thus, throughout this section, $h^{\dagger}$ and $\eta^{\dagger}$ will be regarded as fixed and known functions.

The first expression for the drift that we will describe is stated in part (i) of Lemma 7 below. Although this expression has not appeared in printing, it can be deduced rather easily from Lemma A.2 of [40]. Such Lemma provides a special decomposition of the influence function $Q\left(h,\eta\right)-\theta\left(\eta\right)$ for $\theta\left(\eta\right)$ under a non-parametric model. The form of the drift given in Lemma 7 below, can be deduced rather easily from such decomposition. Nevertheless, in the Appendix B.2, we provide an alternative derivation of this expression which does not require invoking the fact that $Q\left(h,\eta\right)-\theta\left(\eta\right)$ is the influence function for $\theta\left(\eta\right)$.

Define $Q_{K+1}\left(\underline{h}_{K+1}^{\dagger},\underline{\eta}_{K+1}^{\dagger}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$ and sequentially for $j \in [K]$ define,

$$\begin{aligned} Q_j\left(\underline{h}_j^{\dagger},\underline{\eta}_j^{\dagger}\right) &\equiv q_j\left(\overline{L}_{K+1},\overline{I}_j^K;\underline{h}_j^{\dagger},\underline{\eta}_j^{\dagger}\right) \\ &\equiv \eta_j^{\dagger}\left(\overline{L}_j\right)+\sum_{k=j}^{K}\frac{\overline{I}_j^k}{\pi_j^{\dagger k}}\left\{\eta_{k+1}^{\dagger}\left(\overline{L}_{k+1}\right)-\eta_k^{\dagger}\left(\overline{L}_k\right)\right\} \end{aligned} \tag{2.13}$$

Notice that $\underline{h}_1^{\dagger}$ is equal to the entire function vector h. Furthermore, $Q_1\left(\underline{h}_1^{\dagger},\underline{\eta}_1^{\dagger}\right)$ agrees with $\eta_1^{\dagger}\left(L_1\right)+M\left(h^{\dagger},\eta^{\dagger}\right)$ where $M\left(h^{\dagger},\eta^{\dagger}\right)$ was defined in the introduction section. Also recall that $M\left(h^{\dagger},\eta^{\dagger}\right)$ depends on $\eta^{\dagger}$ only through $\underline{\eta}_1^{\dagger}$. Thus, we conclude that with $Q\left(\cdot,\cdot\right)$ defined as in the introduction,

$$Q_1\left(\underline{h}_1^{\dagger},\underline{\eta}_1^{\dagger}\right) = Q\left(h^{\dagger},\eta^{\dagger}\right)$$

A quick calculation shows that for any $j \in [K]$, $Q_j\left(\underline{h}_j^{\dagger},\underline{\eta}_j^{\dagger}\right)$ admits the following alternative expression

$$Q_j\left(\underline{h}_j^{\dagger},\underline{\eta}_j^{\dagger}\right) = \frac{\overline{I}_j^K}{\pi_j^{\dagger K}}\kappa\left(\overline{L}_{K+1}\right)-\sum_{k=j}^{K}\left\{\frac{\overline{I}_j^k}{\pi_j^{\dagger k}}-\frac{\overline{I}_j^{k-1}}{\pi_j^{\dagger(k-1)}}\right\}\eta_k^{\dagger}\left(\overline{L}_k\right).$$

Furthermore, $Q_j\left(\underline{h}_j^{\dagger},\underline{\eta}_j^{\dagger}\right)$ satisfies the following recursive formula, for $j = K, K-1, ..., 1$,

$$Q_j\left(\underline{h}_j^{\dagger},\underline{\eta}_j^{\dagger}\right) = \eta_j^{\dagger}\left(\overline{L}_j\right)+\frac{I_j}{h_j^{\dagger}\left(\overline{L}_j\right)}\left\{Q_{j+1}\left(\underline{h}_{j+1}^{\dagger},\underline{\eta}_{j+1}^{\dagger}\right)-\eta_j^{\dagger}\left(\overline{L}_j\right)\right\}$$

Thus, when $\underline{\eta}_1^\dagger$ is the vector of estimators obtained in steps 2 and 3_MR described in the introduction, and $h_j^\dagger = \widehat{h}_j$, then $Q_j\left(\underline{h}_j^\dagger, \underline{\eta}_j^\dagger\right)$ coincides with the "pseudo outcome" $\widehat{Q}_j$ of step 3_MR of the introduction.

Next, define for any $p = gh$, and any $\left(h^\dagger, \eta^\dagger\right)$,

$$a^p\left(h^\dagger, \eta^\dagger\right) \equiv \sum_{k=1}^{K} E_p\left[\frac{\overline{I}_k}{\pi^{\dagger(k-1)}}\left(\frac{1}{h_k\left(\overline{L}_k\right)} - \frac{1}{h_k^\dagger\left(\overline{L}_k\right)}\right)\left\{\eta_k^\dagger\left(\overline{L}_k\right) - \eta_k\left(\overline{L}_k\right)\right\}\right] \tag{2.14}$$

and, for $j \in [K]$,

$$a_j^p\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger; \overline{L}_j\right) \equiv \sum_{k=j+1}^{K} E_{\underline{g}_j, \underline{h}_{j+1}}\left[\frac{\overline{I}_{j+1}^{k-1}}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{I_k}{h_k\left(\overline{L}_k\right)} - \frac{I_k}{h_k^\dagger\left(\overline{L}_k\right)}\right)\left\{\eta_k^\dagger\left(\overline{L}_k\right) - \eta_k\left(\overline{L}_k\right)\right\}\middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right]$$

Part (i) of the following lemma establishes that $\sqrt{n}a^p\left(h^\dagger, \eta^\dagger\right)$ is equal to the drift (2.6). Part (ii) provides an important generalization which is used to prove two further expressions for the drift that are stated in the subsequent Lemma 8. Lemmas 7 and 8 are proved in Appendix B.2.

Lemma 7 *For $p = gh$, the following holds:*

i)

$$E_p\left\{Q\left(h^\dagger, \eta^\dagger\right)\right\} - \theta\left(\eta\right) = a^p\left(h^\dagger, \eta^\dagger\right), \tag{2.15}$$

ii) for any $j \in [K]$,

$$E_{\underline{g}_j, \underline{h}_{j+1}}\left\{Q_{j+1}\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger\right)\middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\} - \eta_j\left(\overline{L}_j\right) = a_j^p\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger; \overline{L}_j\right). \tag{2.16}$$

Notice that part (i) of Lemma 7 implies that the drift of $\widehat{\theta}$ $\left(\widehat{\theta}_{MR}\right)$ is equal to $\sqrt{n}a^p\left(h^\dagger, \eta^\dagger\right)$ when we replace $h^\dagger$ with $\widehat{h}$, and $\eta^\dagger$ with $\widehat{\eta}$ $\left(\widehat{\eta}_{MR}\right)$.

The expression $\sqrt{n}a^p\left(h^\dagger, \eta^\dagger\right)$ for the drift of the one step estimator is useful for analyzing the properties of the double robust estimator $\widehat{\theta}$ when the nuisance functions h and η are estimated under parametric models, as illustrated in subsection 2.4.1. There exist, however, two other alternative expressions for $a^p\left(h^\dagger, \eta^\dagger\right)$ which appear to better highlight and point out to the properties of the estimators $\widehat{\theta}$ and $\widehat{\theta}_{MR}$, because they are written in terms of the differences between the estimated values $\eta_k^\dagger\left(\overline{L}_k\right)$ and the conditional mean of the "pseudo outcomes" used to compute them. To define these alternative expressions we introduce the following notation.

For any $k \in [K]$, let

$$\Delta_k\left(\eta_k^\dagger, \eta_{k+1}^\dagger; g_k\right) \equiv \overline{I}_k\left[\eta_k^\dagger\left(\overline{L}_k\right) - E_{g_k}\left\{\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right)\middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k\right\}\right]$$

and

$$\Gamma_k\left(\underline{h}_{k+1}^\dagger, \underline{\eta}_k^\dagger; \underline{g}_k, \underline{h}_{k+1}\right) \equiv \overline{I}_k\left[\eta_k^\dagger\left(\overline{L}_k\right) - E_{\underline{g}_k, \underline{h}_{k+1}}\left\{Q_{k+1}\left(\underline{h}_{k+1}^\dagger, \underline{\eta}_{k+1}^\dagger\right)\middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k\right\}\right]$$

where recall $\eta_{K+1}^\dagger\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$.

Define

$$c^p\left(h^\dagger,\eta^\dagger\right)\equiv\sum_{k=1}^{K}E_p\left\{\left(\frac{1}{\pi^k}-\frac{1}{\pi^{\dagger k}}\right)\Delta_k\left(\eta_k^\dagger,\eta_{k+1}^\dagger;g_k\right)\right\} \tag{2.17}$$

and

$$b^p\left(h^\dagger,\eta^\dagger\right)\equiv\sum_{k=1}^{K}E_p\left\{\frac{1}{\pi^{(k-1)}}\left(\frac{1}{h_k\left(\overline{L}_k\right)}-\frac{1}{h_k^\dagger\left(\overline{L}_k\right)}\right)\Gamma_k\left(\underline{h}_{k+1}^\dagger,\underline{\eta}_k^\dagger;\underline{g}_k,\underline{h}_{k+1}\right)\right\}. \tag{2.18}$$

Note that if for each $k\in[K]$, $\eta_k^\dagger$ stands for the estimator $\widehat{\eta}_k$ from step 3 in the algorithm of the introduction section, then for $\overline{I}_k=1$, $\Delta_k\left(\eta_k^\dagger,\eta_{k+1}^\dagger;g_k\right)$ is precisely the difference between the estimated mean and the true conditional mean of the pseudo outcome $\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)$. Likewise, if instead $\eta_k^\dagger$ stands for the estimator $\widehat{\eta}_{k,MR}$ from step 3_MR in the algorithm of the introduction section, then $\Gamma_k\left(\underline{h}_{k+1}^\dagger,\underline{\eta}_k^\dagger;\underline{g}_k,\underline{h}_{k+1}\right)$ for $\overline{I}_k=1$ is precisely the difference between the estimated mean and the true conditional mean of the pseudo outcome $\widehat{Q}_{k+1}$. So, $c^p\left(h^\dagger,\eta^\dagger\right)$ and $b^p\left(h^\dagger,\eta^\dagger\right)$ depend on these differences respectively.

The following Lemma, proved in Appendix B.2, provides two alternative expressions for the drift of the one step estimator.

Lemma 8 $a^p\left(h^\dagger,\eta^\dagger\right)=b^p\left(h^\dagger,\eta^\dagger\right)=c^p\left(h^\dagger,\eta^\dagger\right)$.

2.4.1 Heuristic argument for the double robustness of $\widehat{\theta}$ when h and η are estimated parametrically

As indicated in the introduction, the double robustness of $\widehat{\theta}$ when h and η are estimated under parametric working models for them has been well documented in the literature ([36], [25], [1]). Nevertheless, in this section we illustrate the usefulness of the expression $a^p\left(h^\dagger,\eta^\dagger\right)$ for the drift term to derive the asymptotic property of $\widehat{\theta}$, by providing a heuristic explanation for why $\widehat{\theta}$ is doubly robust when $\widehat{h}$ and $\widehat{\eta}$ are parametric estimators of h and η.

Specifically, suppose that we postulate working parametric models $h_{k,\nu_k}\left(\overline{L}_k\right)$ for $h_k\left(\overline{L}_k\right)$ and $\eta_{k,\psi_k}\left(\overline{L}_k\right)$ for $\eta_k\left(\overline{L}_k\right)\equiv E\left[\eta_{k+1}\left(\overline{L}_{k+1}\right)|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right]$, i.e. models indexed by Euclidean parameters whose dimension does not vary with n. Let $\widehat{\nu}_k$ be the maximum likelihood estimator (MLE) of ν_k. Then $\widehat{\nu}_k$ solves (2.11) for some $t_k(\cdot,\cdot)$. For instance, under model (2.7) the MLE $\widehat{\nu}_k$ of ν_k solves (2.11) for $t_k\left(\overline{L}_k,\nu_k\right)=\phi_k\left(\overline{L}_k\right)$. On the other hand, suppose that the vector $\widehat{\psi}\equiv\left(\widehat{\psi}_1,...,\widehat{\psi}_K\right)$, solves the system of equations (2.12) for $k=1,...,K$ and given functions $t_k\left(\overline{L}_k,\psi_k\right)$. Then, $\widehat{\eta}_k\left(L_k\right)=\eta_{k,\widehat{\psi}_k}\left(\overline{L}_k\right)$ meets the definition given in step 3 for estimating $\eta_k(\cdot)$ in the introduction when the preferred estimator is the, possibly weighted -depending on the function $t_k(\cdot,\cdot)-$ least squares estimator of the regression function $\eta_k(\cdot)$ pretending that the outcome is $\widehat{\eta}_{k+1}\left(L_k\right)=\eta_{k,\widehat{\psi}_{k+1}}\left(\overline{L}_{k+1}\right)$.

Unlike the case in which h and η are estimated nonparametrically, we will argue below that cross-fitting is not needed, so we will assume in this subsection that $\widehat{h}_k\left(\overline{L}_k\right)$ and $\widehat{\eta}_k\left(L_k\right),k=1,...,K$ and $\widehat{\theta}$ are all computed from the entire sample $\mathcal{D}\equiv\{O_i:i=1,...,n\}$.

The estimators $\widehat{\nu} \equiv (\widehat{\nu}_1, ..., \widehat{\nu}_K)'$ and $\widehat{\psi} \equiv \left(\widehat{\psi}_1, ..., \widehat{\psi}_K\right)'$ ultimately solve a system of estimating equations. So, under standard regularity conditions for solutions of estimating equations, there exist $\nu^* \equiv (\nu_1^*, ..., \nu_K^*)$ and $\psi^* \equiv (\psi_1^*, ..., \psi_K^*)$ such that $\sqrt{n}\left\{\left(\widehat{\psi}, \widehat{\nu}\right) - (\psi^*, \nu^*)\right\}$ converges in law to a mean zero normal random vector. Furthermore, ν_k^* is equal to the true parameter value ν_k if the model $h_{k,\nu_k}\left(\overline{L}_k\right)$ is correct and, if the models $\eta_{k,\psi_k}\left(\overline{L}_k\right), k \in [K]$ are all correct then ψ_k^* is equal to the true parameter value ψ_k for every $k \in [K]$. Now, letting $\eta_{\psi^*} \equiv \left(g_0, \eta_{1,\psi_1^*}\left(L_1\right), ..., \eta_{K,\psi_K^*}\left(\overline{L}_K\right)\right)$, $h_{\nu^*} \equiv \left(h_{1,\nu_1^*}, ..., h_{K,\nu_K^*}\right), \eta_{\widehat{\psi}} \equiv \left(g_0, \eta_{1,\widehat{\psi}_1}\left(L_1\right), ..., \eta_{K,\widehat{\psi}_K}\left(\overline{L}_K\right)\right)$ and $h_{\widehat{\nu}} \equiv \left(h_{1,\widehat{\nu}_1}, ..., h_{K,\widehat{\nu}_K}\right)$ write

$$\begin{aligned}\sqrt{n}\left\{\widehat{\theta} - \theta(\eta)\right\} &= \mathbb{G}_n\left\{Q\left(h_{\nu^*}, \eta_{\psi^*}\right)\right\} + \mathbb{G}_n\left\{Q\left(h_{\widehat{\nu}}, \eta_{\widehat{\psi}}\right) - Q\left(h_{\nu^*}, \eta_{\psi^*}\right)\right\} \\ &+ \sqrt{n}\left[E_p\left\{Q\left(h_{\widehat{\nu}}, \eta_{\widehat{\psi}}\right)\right\} - \theta(\eta)\right] \end{aligned} \tag{2.19}$$

Assuming that, for all $k \in [K]$, the maps $\psi_k \to \eta_{\psi_k}\left(\overline{L}_k\right)$ and $\nu_k \to h_{\nu_k}\left(\overline{L}_k\right)$ are continuously differentiable a.s.$\left(\overline{L}_k\right)$ and that the parameter spaces for ψ_k and ν_k are compact, the term $\mathbb{G}_n\left\{Q\left(h_{\widehat{\nu}}, \eta_{\widehat{\psi}}\right) - Q\left(h_{\nu^*}, \eta_{\psi^*}\right)\right\}$ can be shown to be $o_p(1)$ (see Example 19.7 of [58]). On the other hand,

$$\begin{aligned}\sqrt{n}\left[E_p\left\{Q\left(h_{\widehat{\nu}}, \eta_{\widehat{\psi}}\right)\right\} - \theta(\eta)\right] &= \sqrt{n}\left[E_p\left\{Q\left(h_{\widehat{\nu}}, \eta_{\widehat{\psi}}\right)\right\} - E_p\left\{Q\left(h_{\nu^*}, \eta_{\psi^*}\right)\right\}\right] \\ &+ \sqrt{n}\left[E_p\left\{Q\left(h_{\nu^*}, \eta_{\psi^*}\right)\right\} - \theta(\eta)\right] \\ &= \sqrt{n}\left[E_p\left\{Q\left(h_{\widehat{\nu}}, \eta_{\widehat{\psi}}\right)\right\} - E_p\left\{Q\left(h_{\nu^*}, \eta_{\psi^*}\right)\right\}\right] \\ &+ \sqrt{n} a^p\left(h_{\nu^*}, \eta_{\psi^*}\right) \end{aligned} \tag{2.20}$$

If for all k, the model $\eta_{k,\psi_k}\left(\overline{L}_k\right)$ for $\eta_k\left(\overline{L}_k\right)$ is correct, then η_{ψ^*} is equal to the true η. In such case, $a^p\left(h_{\nu^*}, \eta_{\psi^*}\right) = a^p\left(h_{\nu^*}, \eta\right) = 0$. Likewise, if for all k, the model h_{k,ν_k} for h_k is correct, then h_{ν^*} is equal to the true h. In such case, $a^p\left(h_{\nu^*}, \eta_{\psi^*}\right) = a^p\left(h, \eta_{\psi^*}\right) = 0$. So, in both cases, the drift is equal to

$$\sqrt{n}\left[E_p\left\{Q\left(h_{\widehat{\nu}}, \eta_{\widehat{\psi}}\right)\right\} - E_p\left\{Q\left(h_{\nu^*}, \eta_{\psi^*}\right)\right\}\right] \tag{2.21}$$

If the map $(\psi, \nu) \to E_p\left\{Q\left(h_\nu, \eta_\psi\right)\right\}$ is continuously differentiable, then the delta method gives that (2.21) converges to 0 if all working models are correct (since in such case, $\frac{\partial}{\partial(\psi,\nu)} E_p\left\{Q\left(h_\nu, \eta_\psi\right)\right\}\Big|_{(\psi,\nu)=(\psi^*,\nu^*)} = 0$) or to a mean zero Normal distribution otherwise. This then concludes the heuristic proof that $\widehat{\theta}$ is doubly robust, i.e. that $\sqrt{n}\left\{\widehat{\theta} - \theta(\eta)\right\}$ converges to a mean zero Normal distribution if either (i) the parametric regression models assumed for each $\eta_k, k \in [K]$ are all correct, or (ii) the parametric models assumed for each $h_k, k \in [K]$ are all correct, but not necessarily both (i) and (ii) hold simultaneously.

2.4.2 Heuristic argument for the multiple robustness of $\widehat{\theta}_{MR}$ when h and η are parametrically estimated

The multiple robustness property of $\widehat{\theta}_{MR}$ when $\widehat{h}$ and $\widehat{\eta}_{MR}$ are computed under parametric models has been shown for $K = 2$ in [53] and for arbitrary K it can be derived from Lemma 6 of [24].

In this section we use the expression $b^p\left(h^\dagger,\eta^\dagger\right)$ for the drift to provide a heuristic explanation for why $\widehat{\theta}_{MR}$ is multiply robust. As in the preceding subsection, we will assume that h, η and θ are estimated from the entire same sample, i.e. no cross-fitting is employed. Recall that the only difference between $\widehat{\theta}_{MR}$ and $\widehat{\theta}$ are the "pseudo outcomes" used in the estimation of the iterated conditional expectations. Thus, as in the preceding section, we shall assume parametric models η_{k,ψ_k} and h_{k,ν_k} for η_k and h_k. For concreteness, we shall also assume the working model (2.8) and that $\widehat{\psi}_{k,MR}$ is computed recursively, for $k=K,K-1,...,1$, as the, possibly weighted, least squares estimator of a finite dimensional parameter ψ_k in the regression of the "pseudo-outcome"

$$\widehat{Q}_{k+1}\equiv\eta_{k+1,\widehat{\psi}_{k+1,MR}}\left(\overline{L}_{k+1}\right)+\frac{I_{k+1}}{\widehat{h}_{k+1}\left(\overline{L}_{k+1}\right)}\left\{\widehat{Q}_{k+2}-\eta_{k+1,\widehat{\psi}_{k+1,MR}}\left(\overline{L}_{k+1}\right)\right\}$$

on $\overline{L}_k$ under model (2.8) using units that satisfy $\overline{A}_k=\overline{a}_k^*$. That is, $\widehat{\psi}_{k,MR}$ solves

$$\mathbb{P}_n\left[\overline{I}_k t_k\left(\overline{L}_k,\psi_k\right)\left\{\widehat{Q}_{k+1}-\Psi\left\{\psi_k'\phi_k\left(\overline{L}_k\right)\right\}\right\}\right]=0$$

for some vector-valued function $t_k(\cdot,\cdot)$ of the same dimension as ψ_k. Then, just as we reasoned earlier, to analyze the limiting distribution of $\widehat{\theta}_{MR}$ we first note that the vectors $\widehat{\psi}_{MR}=\left(\widehat{\psi}_{1,MR},...,\widehat{\psi}_{K,MR}\right)'$ and $\widehat{\nu}_{ML}$ ultimately solve a joint system of estimating equations, so under regularity conditions, there exist ψ_{MR}^* and ν^* such that $\sqrt{n}\left\{\left(\widehat{\psi}_{MR},\widehat{\nu}\right)-\left(\psi_{MR}^*,\nu^*\right)\right\}$ converges to a mean zero Normal distribution. Next, repeating the expansion (2.19) but with $\widehat{\theta}_{MR}$ instead of $\widehat{\theta}$, $\widehat{\psi}_{MR}$ instead of $\widehat{\psi}$ and ψ_{MR}^* instead of ψ^*, and arguing as in the preceding subsection that under regularity conditions, $\mathbb{G}_n\left\{Q\left(h_{\widehat{\nu}},\eta_{\widehat{\psi}}\right)-Q\left(h_{\nu^*},\eta_{\psi_{MR}^*}\right)\right\}=o_p(1)$, the asymptotic distribution of $\sqrt{n}\left\{\widehat{\theta}_{MR}-\theta(\eta)\right\}$ depends on the asymptotic behavior of $\sqrt{n}\left[E_p\left\{Q\left(h_{\widehat{\nu}},\eta_{\widehat{\psi}_{MR}}\right)\right\}-\theta(\eta)\right]$. Writing the expansion (2.20) but with $\widehat{\psi}_{MR}$ instead of $\widehat{\psi}$ and ψ_{MR}^* instead of ψ^*, we conclude just as in the preceding subsection that if the map $(\psi,\nu)\rightarrow E_p\{Q(h_\nu,\eta_\psi)\}$ is continuously differentiable then by the delta method, $\sqrt{n}\left[E_p\left\{Q\left(h_{\widehat{\nu}},\eta_{\widehat{\psi}_{MR}}\right)\right\}-E_p\left\{Q\left(h_{\nu^*},\eta_{\psi_{MR}^*}\right)\right\}\right]$ converges to either 0 or to a mean zero Normal distribution. Thus, $\sqrt{n}\left\{\widehat{\theta}_{MR}-\theta(\eta)\right\}$ converges to a mean zero Normal distribution if and only if $a^p\left(h_{\nu^*},\eta_{\psi_{MR}^*}\right)=0$.

We will next argue that $\widehat{\theta}_{MR}$ is multiple robust by arguing that, under regularity conditions, $a^p\left(h_{\nu^*},\eta_{\psi_{MR}^*}\right)$, or equivalently $b^p\left(h_{\nu^*},\eta_{\psi_{MR}^*}\right)$ satisfies

$$b^p\left(h_{\nu^*},\eta_{\psi_{MR}^*}\right)=0 \text{ if for each } k\in[K], \text{ either model } \eta_{k,\psi_k} \text{ or model } h_{k,\nu_k} \text{ is correct} \tag{2.22}$$

To argue why (2.22) should be true under regularity conditions, it will be convenient to define the collection $\mathcal{H}_k$ of laws for the observed data O such that h_k is equal to h_{k,ν_k} for some ν_k and likewise to define the collection $\mathcal{G}_k$ of laws for the observed data O such that η_k is equal to η_{k,ψ_k} for some ψ_k. Notice that the assertion that model h_{k,ν_k} is correct or model η_{k,ψ_k} is correct is the same as the assertion that the true data generating law p of O belongs to $\mathcal{H}_k\cup\mathcal{G}_k$. Also, define

$$\eta_{MR}^*\equiv\eta_{\psi_{MR}^*},h^*\equiv h_{\nu^*}$$

We will argue that (2.22) should be true under regularity conditions by arguing that, under regularity conditions, the following fact should hold for each $k\in[K]$.

Fact 1 $\eta_{k,\psi^*_{k,MR}} = \eta_k$ if (i) model η_{k,ψ_k} is correct and (ii) for each $k < j \leq K$ either the model η_{j,ψ_j} is correct or the model h_{j,ν_j} is correct

Heuristic argument of why fact 1 should hold under regularity conditions. We argue by reverse induction in k. For $k = K, \widehat{\psi}_{K,MR}$ is the estimated coefficient from the, possibly non-linear, least squares procedure with outcome $\kappa\left(\overline{L}_{K+1}\right)$ and covariates $\phi_K\left(\overline{L}_K\right)$ among units with $\overline{A}_K = \overline{a}^*_K$, under the model $\eta_{K,\psi_K}\left(\overline{L}_K\right)$ for $E\left[\kappa\left(\overline{L}_{K+1}\right)|\overline{A}_K = \overline{a}^*_K, \overline{L}_K\right]$. If this model is correct, then under standard regularity conditions, the probability limit $\psi^*_{K,MR}$ of $\widehat{\psi}_{K,MR}$ is equal to the true value ψ_K. Consequently, $\eta_{K,\psi^*_{K,MR}}$ is equal to η_K and therefore fact 1 holds for $k = K$. Next, suppose that fact 1 holds for $k = K, \ldots, j+1$. Noticing that, by construction, $\widehat{Q}_{j+1} = Q_{j+1}\left(\underline{\widehat{h}}_{j+1}, \underline{\widehat{\eta}}_{MR,j+1}\right)$, we conclude that under regularity conditions, $\widehat{\psi}_{j,MR}$ solves

$$0 = \mathbb{P}_n\left[\overline{I}_j\phi_j\left(\overline{L}_j\right)\left\{Q_{j+1}\left(\underline{h}^*_{j+1}, \underline{\eta}^*_{MR,j+1}\right) - \Psi\left\{\psi_j^T\phi_j\left(\overline{L}_j\right)\right\}\right\}\right] + o_p(1)$$

Suppose $p \in \mathcal{G}_j \cap \left[\cap_{k=j+1}^K\left(\mathcal{H}_k \cup \mathcal{G}_k\right)\right]$. Then, for each $k = j+1, \ldots, K$, either $p \in \mathcal{H}_k$ or $p \in \mathcal{G}_k$. If $p \in \mathcal{G}_k$ then since p also belongs to $\cap_{r=k+1}^K\left(\mathcal{H}_r \cup \mathcal{G}_r\right)$ we have, by inductive hypothesis, that $\eta^*_{k,MR} = \eta_k$. If $p \in \mathcal{H}_k$, then $h^*_k = h_k$. Thus, for every $k = j+1, \ldots, K, h^*_k = h_k$ or $\eta^*_{k,MR} = \eta_k$. Consequently, by part (ii) of Lemma 7, $E_g\left\{Q_{j+1}\left(\underline{h}^*_{j+1}, \underline{\eta}^*_{MR,j+1}\right)\middle|\overline{A}_j = \overline{a}^*_j, \overline{L}_j\right\} = \eta_j\left(\overline{L}_j\right)$. Furthermore, since $p \in \mathcal{G}_j, \eta_j = \eta_{j,\psi_j}$ for some "true" ψ_j and therefore the equation

$$0 = E_{\overline{g}_j,\overline{h}_j}\left[\overline{I}_j\phi_j\left(\overline{L}_j\right)\left\{Q_{j+1}\left(\underline{h}^*_{j+1}, \underline{\eta}^*_{MR,j+1}\right) - \Psi\left\{\psi_j^T\phi_j\left(\overline{L}_j\right)\right\}\right\}\right]$$

is solved at the true ψ_j. Then, under regularity conditions for the consistency of $M-$ estimators, the probability limit $\psi^*_{j,MR}$ of $\widehat{\psi}_{j,MR}$ is equal to the "true" ψ_j which shows fact 1 holds for $k = j$.

We now argue why (2.22) should hold under regularity conditions, by arguing that for $p \in \cap_{r=k}^K\left(\mathcal{H}_r \cup \mathcal{G}_r\right)$ it should hold that

$$E_{\overline{g}_{k-1},\overline{h}_k}\left[\frac{1}{\pi^{(k-1)}}\left(\frac{1}{h_k} - \frac{1}{h^*_k}\right)\Gamma_k\left(\underline{h}^*_{k+1}, \underline{\eta}^*_{MR,k}; \underline{g}_k, \underline{h}_{k+1}\right)\right] = 0 \tag{2.23}$$

Suppose then that $p \in \cap_{r=k}^K\left(\mathcal{H}_r \cup \mathcal{G}_r\right)$. If $p \in \mathcal{H}_k$ then under regularity conditions $h^*_k = h_k$ and thus (2.23) holds. If $p \notin \mathcal{H}_k$ then $p \in \mathcal{G}_k \cap \left[\cap_{r=k+1}^K\left(\mathcal{H}_r \cup \mathcal{G}_r\right)\right]$. Then, under regularity conditions, fact 1 implies that $\eta^*_{k,MR} = \eta_k$. In addition, for $r = k+1, \ldots, K$, either $p \in \mathcal{H}_r$ in which case, again under reg. conditions, $h^*_r = h_r$ or $p \in \mathcal{G}_r \cap \left[\cap_{s=r+1}^K\left(\mathcal{H}_s \cup \mathcal{G}_s\right)\right]$ in which case, again by fact $1, \eta^*_{r,MR} = \eta_r$. Thus, we conclude that when $p \in \mathcal{G}_k \cap \left[\cap_{r=k+1}^K\left(\mathcal{H}_r \cup \mathcal{G}_r\right)\right]$, $\eta^*_{k,MR} = \eta_k$ and for $r = k+1, \ldots, K$, either $\eta^*_{r,MR} = \eta_r$ or $h^*_r = h_r$. Thus, since by Lemma 7 we know that

$$\Gamma_k\left(\underline{h}^\dagger_{k+1}, \underline{\eta}^\dagger_k; \underline{g}_k, \underline{h}_{k+1}\right) = 0 \text{ if } \eta^\dagger_k = \eta_k \text{ and for } j > k, \text{either } \eta^\dagger_j = \eta_j \text{ or } h^\dagger_j = h_j, \tag{2.24}$$

we conclude that $\Gamma_k\left(\underline{h}^*_{k+1}, \underline{\eta}^*_{MR,k}; \underline{g}_k, \underline{h}_{k+1}\right) = 0$, which then implies that (2.23) holds. This ends the heuristic proof of the multiple robustness of $\widehat{\theta}_{MR}$.

2.5 The proposed estimators of θ when the functions h and η are estimated non-parametrically

In this section we provide the precise steps for computing the estimators $\widehat{\theta}$ and $\widehat{\theta}_{MR}$ when h and η are estimated non-parametrically.

In what follows, given a sample $\mathcal{S}$ and a positive integer J, the operation of randomly splitting $\mathcal{S}$ into J equally or nearly equal sized subsamples stands for the operation of randomly partitioning $\mathcal{S}$ into J disjoint subsamples, $\mathcal{S}^1, ..., \mathcal{S}^J$, such that the size of each $\mathcal{S}^j$ is either $\lfloor N/J \rfloor$ or $\lfloor N/J \rfloor + 1$, where $\lfloor \cdot \rfloor$ is the floor function, i.e. $\lfloor x \rfloor$ is the greatest integer less than or equal x.

The algorithm for computing $\widehat{\theta}$ and $\widehat{\theta}_{MR}$ starts by randomly splitting the entire sample $\mathcal{D} \equiv \{O_i : i = 1, ..., n\}$ into a fixed number $\boldsymbol{U}$ of equally or nearly equally sized subsamples $\mathcal{D}^u, u = 1, ..., \boldsymbol{U}$. The algorithm then computes one estimator $\widehat{\theta}^u$ and one estimator $\widehat{\theta}^u_{MR}$ for each u following the procedure indicated below. The final estimators are defined as

$$\widehat{\theta} \equiv \frac{1}{\boldsymbol{U}} \sum_{u=1}^{\boldsymbol{U}} \widehat{\theta}^u$$

and

$$\widehat{\theta}_{MR} \equiv \frac{1}{\boldsymbol{U}} \sum_{u=1}^{\boldsymbol{U}} \widehat{\theta}^u_{MR}$$

To facilitate the understanding of the algorithm for computing the proposed estimators $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ we first describe it for the special cases $K = 2$ and $K = 3$ and subsequently we state it for an arbitrary K.

Procedure for computing $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ when $K = 2$.

- Designate sample $\mathcal{D}^u$ as the main estimation sample and designate its complement $\mathcal{N} \equiv \mathcal{D} - \mathcal{D}^u$ as the nuisance estimation sample. Randomly split $\mathcal{N}$ into 2 equally or nearly equally sized, subsamples $\mathcal{N}^1$ and $\mathcal{N}^2$.
- *Estimation of h_1 and h_2.*
  - using data from subsample $\mathcal{N}^1$ compute the preferred non-parametric estimator $\widehat{h}_1(\cdot)$ of $h_1(\cdot)$.
  - using data from subsample $\mathcal{N}^2$ compute the preferred non-parametric estimator $\widehat{h}_2(\cdot)$ of $h_2(\cdot)$.
- *Estimation of η_2 to be used in constructing both $\widehat{\theta}$ and $\widehat{\theta}_{MR}$.* Using data from subsample $\mathcal{N}^2$, compute the preferred non-parametric estimator $\widehat{\eta}_2(\cdot)$ of $\eta_2(\cdot) \equiv E\left[\kappa\left(\overline{L}_3\right) | \overline{A}_2 = \overline{a}^*_2, \overline{L}_2 = \cdot\right]$. Let $\widehat{\eta}_{2,MR}(\cdot) \equiv \widehat{\eta}_2(\cdot)$.
- *Estimation of η_1*
  - *Estimation of η_1 to be used in constructing $\widehat{\theta}$.* Using data from units in subsample $\mathcal{N}^1$ with $A_1 = a^*_1$ and the already estimated function $\widehat{\eta}_2(\cdot)$ compute the preferred non-parametric estimator $\widehat{\eta}_1(\cdot)$ of $\eta_1(\cdot) \equiv E\left[\eta_2\left(\overline{L}_2\right) | A_1 = a^*_1, L_1 = \cdot\right]$ obtained by pretending that the unknown "outcome" $\eta_2\left(\overline{L}_2\right)$ is equal to the "pseudo outcome" $\widehat{\eta}_2\left(\overline{L}_2\right)$.

– *Estimation of* η_1 *to be used in constructing* $\widehat{\theta}_{MR}$. Using data from units in subsample $\mathcal{N}^1$ with $A_1 = a_1^*$ and the already estimated functions $\left(\widehat{\eta}_{2,MR}(\cdot), \widehat{h}_2(\cdot)\right)$ compute the preferred non-parametric estimator $\widehat{\eta}_{1,MR}(\cdot)$ of $\eta_1(\cdot) \equiv E\left[\eta_2\left(\overline{L}_2\right) | A_1 = a_1^*, L_1 = \cdot\right]$, obtained by pretending that the unknown "outcome" $\eta_2\left(\overline{L}_2\right)$ is equal to the "pseudo-outcome"

$$\begin{aligned}\widehat{Q}_2 &\equiv q_2\left(\overline{L}_3, \overline{I}_2^2; \underline{\widehat{h}}_2, \underline{\widehat{\eta}}_{2,MR}\right) \\ &= \widehat{\eta}_{2,MR}\left(\overline{L}_2\right) + \frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right) - \widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\end{aligned}$$

- Using the already estimated functions $\widehat{h}_1, \widehat{h}_2, \widehat{\eta}_1$ and $\widehat{\eta}_2$ compute

$$\widehat{\theta}^u = \frac{1}{\#\mathcal{D}^u} \sum_{i:O_i \in \mathcal{D}^u} q_1\left(\overline{L}_{3,i}, \overline{I}_{1,i}^2; \widehat{h}, \widehat{\eta}\right)$$

- Using the already estimated functions $\widehat{h}_1, \widehat{h}_2, \widehat{\eta}_{1,MR}$ and $\widehat{\eta}_{2,MR}$, compute

$$\widehat{\theta}^u_{MR} = \frac{1}{\#\mathcal{D}^u} \sum_{i:O_i \in \mathcal{D}^u} q_1\left(\overline{L}_{3,i}, \overline{I}_{1,i}^2; \widehat{h}, \widehat{\eta}_{MR}\right)$$

Procedure for computing $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ when $K = 3$.

- Designate sample $\mathcal{D}^u$ as the main estimation sample and designate its complement $\mathcal{N} \equiv \mathcal{D} - \mathcal{D}^u$ as the nuisance estimation sample. Randomly split $\mathcal{N}$ into 3 equally or nearly equally sized, subsamples $\mathcal{N}^1, \mathcal{N}^2$ and $\mathcal{N}^3$.

- *Estimation of* h_1,h_2 and h_3.

  – using data from subsample $\mathcal{N}^k$ compute the preferred non-parametric estimator $\widehat{h}_k(\cdot)$ of $h_k(\cdot), k = 1, 2, 3$.

- *Estimation of* η_3 *to be used in constructing both* $\widehat{\theta}$ *and* $\widehat{\theta}_{MR}$. Using data from subsample $\mathcal{N}^3$, compute the preferred non-parametric estimator $\widehat{\eta}_3(\cdot)$ of $\eta_3(\cdot) \equiv E\left[\kappa\left(\overline{L}_4\right) | \overline{A}_3 = \overline{a}_3^*, \overline{L}_3 = \cdot\right]$. Let $\widehat{\eta}_{3,MR}(\cdot) \equiv \widehat{\eta}_3(\cdot)$.

- *Estimation of* η_2 and η_1

  – *Estimation of* η_2 and η_1 *to be used in constructing* $\widehat{\theta}$.

    * Using data from units in subsample $\mathcal{N}^2$ with $\overline{A}_2 = \overline{a}_2^*$ and the already estimated function $\widehat{\eta}_3(\cdot)$ compute the preferred non-parametric estimator $\widehat{\eta}_2(\cdot)$ of $\eta_2(\cdot) \equiv E\left[\eta_3\left(\overline{L}_3\right) | \overline{A}_2 = \overline{a}_2^*, \overline{L}_2 = \cdot\right]$ obtained by pretending that the unknown "outcome" $\eta_3\left(\overline{L}_3\right)$ is equal to the "pseudo outcome" $\widehat{\eta}_3\left(\overline{L}_3\right)$.

* Using data from units in subsample $\mathcal{N}^1$ with $A_1 = a_1^*$ and the already estimated function $\widehat{\eta}_2(\cdot)$ compute the preferred non-parametric estimator $\widehat{\eta}_1(\cdot)$ of $\eta_1(\cdot) \equiv E\left[\eta_2\left(\overline{L}_2\right)|A_1 = a_1^*, L_1 = \cdot\right]$ obtained by pretending that the unknown "outcome" $\eta_2\left(\overline{L}_2\right)$ is equal to the "pseudo outcome" $\widehat{\eta}_2\left(\overline{L}_2\right)$.

– *Estimation of* η_2 and η_1 to be used in constructing $\widehat{\theta}_{MR}$.

* Using data from units in subsample $\mathcal{N}^2$ with $\overline{A}_2 = \overline{a}_2^*$ and the already estimated functions $\left(\widehat{\eta}_{3,MR}(\cdot), \widehat{h}_3(\cdot)\right)$ compute the preferred non-parametric estimator $\widehat{\eta}_{2,MR}(\cdot)$ of $\eta_2(\cdot) \equiv E\left[\eta_3\left(\overline{L}_3\right)|\overline{A}_2 = \overline{a}_2^*, \overline{L}_2 = \cdot\right]$, obtained by pretending that the unknown "outcome" $\eta_3\left(\overline{L}_3\right)$ is equal to the "pseudo-outcome"

$$\begin{aligned}\widehat{Q}_3 &\equiv q_3\left(\overline{L}_4, \overline{I}_3^3; \underline{\widehat{h}}_3, \underline{\widehat{\eta}}_{3,MR}\right) \\ &= \widehat{\eta}_{3,MR}\left(\overline{L}_3\right) + \frac{I_3}{\widehat{h}_3\left(\overline{L}_3\right)}\left\{\kappa\left(\overline{L}_4\right) - \widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\end{aligned}$$

* Using data from units in subsample $\mathcal{N}^1$ with $A_1 = a_1^*$ and the already estimated functions $\left(\widehat{\eta}_{3,MR}(\cdot), \widehat{h}_3(\cdot)\right)$ and $\left(\widehat{\eta}_{2,MR}(\cdot), \widehat{h}_2(\cdot)\right)$ compute the preferred non-parametric estimator $\widehat{\eta}_{1,MR}(\cdot)$ of $\eta_1(\cdot) \equiv E\left[\eta_2\left(\overline{L}_2\right)|A_1 = a_1^*, L_1 = \cdot\right]$, obtained by pretending that the unknown "outcome" $\eta_2\left(\overline{L}_2\right)$ is equal to the "pseudo-outcome"

$$\begin{aligned}\widehat{Q}_2 &\equiv q_2\left(\overline{L}_4, \overline{I}_2^3; \underline{\widehat{h}}_2, \underline{\widehat{\eta}}_{2,MR}\right) \\ &= \widehat{\eta}_{2,MR}\left(\overline{L}_2\right) + \frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\widehat{\eta}_{3,MR}\left(\overline{L}_3\right) - \widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\} \\ &+ \frac{I_2 I_3}{\widehat{h}_2\left(\overline{L}_2\right)\widehat{h}_3\left(\overline{L}_3\right)}\left\{\kappa\left(\overline{L}_3\right) - \widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\} \\ &= \widehat{\eta}_{2,MR}\left(\overline{L}_2\right) + \frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{q_3\left(\overline{L}_4, \overline{I}_3^3; \underline{\widehat{h}}_3, \underline{\widehat{\eta}}_{3,MR}\right) - \widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\end{aligned}$$

- Using the already estimated functions $\widehat{h}_1, \widehat{h}_2, \widehat{h}_3, \widehat{\eta}_1, \widehat{\eta}_2$ and $\widehat{\eta}_3$ compute

$$\widehat{\theta}^u = \frac{1}{\#\mathcal{D}^u}\sum_{i:O_i \in \mathcal{D}^u} q_1\left(\overline{L}_{4,i}, \overline{I}_{1,i}^3; \widehat{h}, \widehat{\eta}\right)$$

- Using the already estimated functions $\widehat{h}_1, \widehat{h}_2, \widehat{h}_3, \widehat{\eta}_{1,MR}, \widehat{\eta}_{2,MR}$ and $\widehat{\eta}_{3,MR}$ compute

$$\widehat{\theta}^u_{MR} = \frac{1}{\#\mathcal{D}^u}\sum_{i:O_i \in \mathcal{D}^u} q_1\left(\overline{L}_{4,i}, \overline{I}_{1,i}^3; \widehat{h}, \widehat{\eta}_{MR}\right)$$

Procedure for computing $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ for arbitrary K.

- Designate sample $\mathcal{D}^u$ as the main estimation sample and designate its complement $\mathcal{N} \equiv \mathcal{D} - \mathcal{D}^u$ as the nuisance estimation sample. Randomly split $\mathcal{N}$ into K equally or nearly equally sized, subsamples $\mathcal{N}^1, ..., \mathcal{N}^K$.

- *Estimation of* h.
  - using data from subsample $\mathcal{N}^k$ compute the preferred non-parametric estimator $\widehat{h}_k(\cdot)$ of $h_k(\cdot), k = 1, ..., K$.

- *Estimation of* η_K *to be used in constructing both* $\widehat{\theta}$ *and* $\widehat{\theta}_{MR}$. Using data from subsample $\mathcal{N}^K$, compute the preferred non-parametric estimator $\widehat{\eta}_K(\cdot)$ of $\eta_K(\cdot) \equiv E\left[\kappa\left(\overline{L}_{K+1}\right)|\overline{A}_K = \overline{a}_K^*, \overline{L}_K = \cdot\right]$. Let $\widehat{\eta}_{K,MR}(\cdot) \equiv \widehat{\eta}_K(\cdot)$.

- *Estimation of* η_k for $k = K-1, ..., 1$
  - *Estimation of* η_k for $k = K-1, ..., 1$, *to be used in constructing* $\widehat{\theta}$. For $k = K-1, ..., 1$ repeat
    - Using data from units in subsample $\mathcal{N}^k$ with $\overline{A}_k = \overline{a}_k^*$ and the already estimated function $\widehat{\eta}_{k+1}(\cdot)$ compute the preferred non-parametric estimator $\widehat{\eta}_k(\cdot)$ of $\eta_k(\cdot) \equiv E\left[\eta_{k+1}\left(\overline{L}_{k+1}\right)|\overline{A}_k = \overline{a}_k^*, \overline{L}_k = \cdot\right]$ obtained by pretending that the unknown "outcome" $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ is equal to the "pseudo outcome" $\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)$.
  - *Estimation of* η_k for $k = K-1, ..., 1$,*to be used in constructing* $\widehat{\theta}_{MR}$. For $k = K-1, ..., 1$ repeat
    - Using data from units in subsample $\mathcal{N}^k$ with $\overline{A}_k = \overline{a}_k^*$ and the already estimated functions $\left(\widehat{\eta}_{k+1,MR}(\cdot), \widehat{h}_{k+1}(\cdot)\right), ..., \left(\widehat{\eta}_{K,MR}(\cdot), \widehat{h}_K(\cdot)\right)$ compute the preferred non-parametric estimator $\widehat{\eta}_{k,MR}(\cdot)$ of $\eta_k(\cdot) \equiv E\left[\eta_{k+1}\left(\overline{L}_{k+1}\right)|\overline{A}_k = \overline{a}_k^*, \overline{L}_k = \cdot\right]$, obtained by pretending that the unknown "outcome" $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ is equal to the "pseudo-outcome"

$$\begin{aligned}
\widehat{Q}_{k+1} &\equiv q_{k+1}\left(\overline{L}_{K+1}, \overline{I}_{k+1}^K; \underline{\widehat{h}}_{k+1}, \underline{\widehat{\eta}}_{k+1,MR}\right) \\
&= \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right) + \sum_{r=k+1}^{K-1} \left\{\prod_{j=k+1}^{r} \frac{I_j}{\widehat{h}_j\left(\overline{L}_j\right)}\right\} \left\{\widehat{\eta}_{r+1,MR}\left(\overline{L}_{r+1}\right) - \widehat{\eta}_{r,MR}\left(\overline{L}_r\right)\right\} \\
&+ \left\{\prod_{j=k+1}^{K} \frac{I_j}{\widehat{h}_j\left(\overline{L}_j\right)}\right\} \left\{\kappa\left(\overline{L}_{K+1}\right) - \widehat{\eta}_{K,MR}\left(\overline{L}_K\right)\right\} \\
&= \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right) \\
&+ \frac{I_{k+1}}{\widehat{h}_{k+1}\left(\overline{L}_{k+1}\right)} \left\{q_{k+2}\left(\overline{L}_{K+1}, \overline{I}_{k+2}^K; \underline{\widehat{h}}_{k+2}, \underline{\widehat{\eta}}_{k+2,MR}\right) - \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right)\right\}
\end{aligned}$$

    where $q_{K+1}\left(\overline{L}_{K+1}, \overline{I}_{K+1}^K; \underline{\widehat{h}}_{K+1}, \underline{\widehat{\eta}}_{K+1,MR}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$.

- Using the already estimated functions $\widehat{h} = \left(\widehat{h}_1, ..., \widehat{h}_K\right)$ and $\widehat{\eta} = (\widehat{\eta}_1, ..., \widehat{\eta}_K)$, compute

$$\widehat{\theta}^u = \frac{1}{\#\mathcal{D}^u} \sum_{i: O_i \in \mathcal{D}^u} q_1\left(\overline{L}_{K+1,i}, \overline{I}_{1,i}^K; \widehat{h}, \widehat{\eta}\right)$$

- Using the already estimated functions $\widehat{h} = \left(\widehat{h}_1, ..., \widehat{h}_K\right)$ and $\widehat{\eta}_{MR} = \left(\widehat{\eta}_{1,MR}, ..., \widehat{\eta}_{K,MR}\right)$, compute

$$\widehat{\theta}^u_{MR} = \frac{1}{\#\mathcal{D}^u} \sum_{i:O_i \in \mathcal{D}^u} q_1\left(\overline{L}_{K+1,i}, \overline{I}^K_{1,i}; \widehat{h}, \widehat{\eta}_{MR}\right)$$

The sample splitting strategy employed by our algorithm is important and worth summarizing for clarity. Specifically, for each u, $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ are computed by averaging the estimates $q_1\left(\overline{L}_{K+1}, \overline{I}^K_1; \widehat{h}, \widehat{\eta}\right)$ and $q_1\left(\overline{L}_{K+1}, \overline{I}^K_1; \widehat{h}, \widehat{\eta}_{MR}\right)$ over units in the subsample $\mathcal{D}^u$. The functions $h_k, k = 1, ..., K$ are each estimated from an independent subsample $\mathcal{N}^k$ of $\mathcal{N} = \mathcal{D} - \mathcal{D}^u$. The functions $\eta_k, k = 1, ..., K$ are estimated, in turn, from units in $\mathcal{N} = \mathcal{D} - \mathcal{D}^u$ as the result of a recursive process. The recursive process differs depending on whether the goal is to construct $\widehat{\theta}^u$ or $\widehat{\theta}^u_{MR}$. In both cases, in the recursive process, for $k = K, ..., 1$, each function $\eta_k(\cdot)$ is estimated from units in an independent subsample $\mathcal{N}^k$ of $\mathcal{N}$ using, either just the already estimated function $\widehat{\eta}_{k+1}(\cdot)$ to construct the pseudo outcome -if the ultimate goal is to compute $\widehat{\theta}^u$-, or using the function $q_{k+1}\left(\cdot, \cdot; \underline{\widehat{h}}_{k+1}, \underline{\widehat{\eta}}_{k+1,MR}\right)$, that depends on the already estimated functions $\widehat{\eta}_{k+1,MR}(\cdot), ..., \widehat{\eta}_{K,MR}(\cdot)$ and $\widehat{h}_{k+1}(\cdot), ..., \widehat{h}_K(\cdot)$, to construct the pseudo outcome -if the ultimate goal is to compute $\widehat{\theta}^u_{MR}$-.

The estimation of each h_k and each η_k from a distinct subsample $\mathcal{N}^k$ of the nuisance estimation sample $\mathcal{N}$ is carried out to allow the derivation of bounds on the rates of convergence of the drifts of $\widehat{\theta}_{DR}$ and $\widehat{\theta}_{MR}$ when the η_k are estimated via series estimation. These drifts depend on quantities that include least squares fits of outcomes that are, in turn, data dependent functions of O. Without sample splitting, these outcomes cannot be treated as i.i.d. and the drifts become analytically intractable without making further assumptions (see the Remark 1 at the end of Section 2.7.2 and Remark 2 of Appendix B.6).

To explain the reason why we compute $\widehat{\theta}^u$ ($\widehat{\theta}^u_{MR}$) by averaging over units in $\mathcal{D}^u$ the values of $q_1\left(\overline{L}_{K+1}, \overline{I}^K_1; \widehat{h}, \widehat{\eta}\right)$ ($q_1\left(\overline{L}_{K+1}, \overline{I}^K_1; \widehat{h}, \widehat{\eta}_{MR}\right)$) using estimated functions $\widehat{h}, \widehat{\eta}$ ($\widehat{h}, \widehat{\eta}_{MR}$) computed from an independent sample $\mathcal{N} = \mathcal{D} - \mathcal{D}^u$, denote with N_u the sample size of $\mathcal{D}^u$ and let $\mathbb{G}_{N_u} \equiv \sqrt{N_u}\mathbb{P}_{N_u}\{\cdot - E_p(\cdot)\}$ be the centered empirical process over the sample $\mathcal{D}^u$. Because the subsamples $\mathcal{D}^u$ were obtained from a process that sample splitted $\mathcal{D}$ into equal or nearly equal size subsamples, each N_u is either equal to $\lfloor n/\mathbf{U} \rfloor$ or to $\lfloor n/\mathbf{U} \rfloor + 1$. Then, letting

$$N \equiv \lfloor n/\mathbf{U} \rfloor$$

we have that

$$N \leq N_u \leq N + 1, \tag{2.25}$$

$$\mathbf{U}(N_u - 1) \leq n \leq \mathbf{U}(N_u + 1) \tag{2.26}$$

Since $\mathbf{U}$ is a fixed constant, then from (2.26) we have that

$$N_u \to \infty \text{ as } n \to \infty \tag{2.27}$$

and consequently from (2.25) we have that $N \to \infty$ as $n \to \infty$ and

$$\frac{N_u}{N} \underset{n\to\infty}{\to} 1 \tag{2.28}$$

Furthermore, from (2.26) we have

$$o_p(1) \text{ as } N_u \to \infty \Leftrightarrow o_p(1) \text{ as } n \to \infty \tag{2.29}$$

Now, noting that $q_1\left(\overline{L}_{K+1}, \overline{I}_1^K; \widehat{h}, \widehat{\eta}\right)$ ($q_1\left(\overline{L}_{K+1}, \overline{I}_1^K; \widehat{h}, \widehat{\eta}_{MR}\right)$) agrees with $Q\left(h^\dagger, \eta^\dagger\right)$ as defined in the introduction when $\left(h^\dagger, \eta^\dagger\right)$ is replaced with $\left(\widehat{h}, \widehat{\eta}\right)$ ($\left(\widehat{h}, \widehat{\eta}_{MR}\right)$), then just as in the introduction, we can write

$$\begin{aligned}\sqrt{N_u}\left\{\widehat{\theta}^{\dagger,u} - \theta(\eta)\right\} &= \mathbb{G}_{N_u}\{Q(h,\eta)\} + \mathbb{G}_{N_u}\left\{Q\left(h^\dagger, \eta^\dagger\right) - Q(h,\eta)\right\} \\ &\quad + \sqrt{N_u}\left[E_p\left\{Q\left(h^\dagger, \eta^\dagger\right)\right\} - \theta(\eta)\right]\end{aligned}$$

where $\widehat{\theta}^{\dagger,u}$ denotes $\widehat{\theta}^u$ if $\left(h^\dagger, \eta^\dagger\right)$ is replaced with $\left(\widehat{h}, \widehat{\eta}\right)$, $\widehat{\theta}^{\dagger,u}$ denotes $\widehat{\theta}^u_{MR}$ if $\left(h^\dagger, \eta^\dagger\right)$ is replaced with $\left(\widehat{h}, \widehat{\eta}_{MR}\right)$ and

$$\begin{aligned}&\mathbb{G}_{N_u}\left\{Q\left(h^\dagger, \eta^\dagger\right) - Q(h,\eta)\right\} \equiv \\ &\equiv \frac{1}{\sqrt{N_u}} \sum_{i: O_i \in \mathcal{D}^u} \left\{Q\left(h^\dagger, \eta^\dagger\right)_i - Q(h,\eta)_i - E_p\left[Q\left(h^\dagger, \eta^\dagger\right) - Q(h,\eta) \,|\mathcal{N}\right]\right\}\end{aligned}$$

with $Q\left(h^\dagger, \eta^\dagger\right)_i \equiv q_1\left(\overline{L}_{K+1,i}, \overline{I}_{1,i}^K; h^\dagger, \eta^\dagger\right)$ and $Q(h,\eta)_i \equiv q_1\left(\overline{L}_{K+1,i}, \overline{I}_{1,i}^K; h, \eta\right)$. Sample splitting $\mathcal{D}$ into $\mathcal{D}^u$ and $\mathcal{N} = \mathcal{D} - \mathcal{D}^u$ and computing $\left(h^\dagger, \eta^\dagger\right)$ from the independent sample $\mathcal{D} - \mathcal{D}^u$ results in $\mathbb{G}_{N_u}\left\{Q\left(h^\dagger, \eta^\dagger\right) - Q(h,\eta)\right\}$ to be equal to $\sqrt{N_u}$ times an average of N_u random variables that, conditionally on the data in $\mathcal{N}$, are independent and identically distributed, and have mean zero. Thus, under the very mild requirement that

$$E_p\left[\left\{Q\left(h^\dagger, \eta^\dagger\right) - Q(h,\eta)\right\}^2 |\mathcal{N}\right] \underset{N_u \to \infty}{\overset{P}{\to}} 0 \tag{2.30}$$

and consequently, by (2.27), as the sample size n of $\mathcal{D}$ converges to ∞, an application of the Dominated Convergence Theorem gives that

$$\mathbb{G}_{N_u}\left\{Q\left(h^\dagger, \eta^\dagger\right) - Q(h,\eta)\right\} \underset{n \to \infty}{\overset{P}{\to}} 0. \tag{2.31}$$

For completeness, in Appendix B.3 we show that (2.30) implies (2.31). We qualify the requirement (2.30) as very mild because it typically holds under weaker conditions than those required to ensure that the drift term $\sqrt{N_u}\left[E_p\left\{Q\left(h^\dagger, \eta^\dagger\right)\right\} - \theta(\eta)\right]$ is $o_p(1)$. In Section 2.8, we illustrate this point for $K = 2$ and $K = 3$ when $\eta_k^\dagger, k = 1, .., K$, are series estimators of η_k as defined in the following subsection.

Now, suppose that having established that (2.31) holds for the particular estimators $\left(h^\dagger, \eta^\dagger\right)$ of (h, η) ($\left(\widehat{h}, \widehat{\eta}\right)$ or $\left(\widehat{h}, \widehat{\eta}_{MR}\right)$), we could find conditions that ensure that the drift term $\sqrt{N_u}\left[E_p\left\{Q\left(h^\dagger, \eta^\dagger\right)\right\} - \theta(\eta)\right]$ converges to zero in probability as N_u converges to ∞, then we would conclude that under such conditions,

$$\sqrt{N_u}\left\{\widehat{\theta}^{\dagger,u} - \theta(\eta)\right\} = \mathbb{G}_{N_u}\{Q(h,\eta)\} + o_p(1) \tag{2.32}$$

showing then that asymptotically, $\widehat{\theta}^{\dagger,u} - \theta(\eta)$ behaves like a sample average of N_u mean zero random variables. This highlights the obvious fact that by sample splitting we loose the information about θ available in the $n - N_u$ units in the nuisance estimation sample. We recover the information lost due to sample splitting by computing one $\widehat{\theta}^{\dagger,u}$ for each $u = 1, ..., \boldsymbol{U}$ and designating our final estimator of θ as the average of the $\widehat{\theta}^{\dagger,u}$ over u. To show this, assuming that (2.32) holds, we first write

$$\begin{aligned}
\sqrt{N}\left\{\widehat{\theta}^{\dagger,u} - \theta(\eta)\right\} &= \sqrt{N_u}\left\{\widehat{\theta}^{\dagger,u} - \theta(\eta)\right\} + \left\{\sqrt{\frac{N}{N_u}} - 1\right\}\sqrt{N_u}\left\{\widehat{\theta}^{\dagger,u} - \theta(\eta)\right\} \\
&= \sqrt{N_u}\left\{\widehat{\theta}^{\dagger,u} - \theta(\eta)\right\} + o_p(1) \\
&= \frac{1}{\sqrt{N_u}} \sum_{i:O_i \in \mathcal{D}^u} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\} + o_p(1) \\
&= \frac{1}{\sqrt{N}} \sum_{i:O_i \in \mathcal{D}^u} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\} + \\
&\left(1 - \sqrt{\frac{N_u}{N}}\right) \frac{1}{\sqrt{N_u}} \sum_{i:O_i \in \mathcal{D}^u} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\} + o_p(1) \\
&= \frac{1}{\sqrt{N}} \sum_{i:O_i \in \mathcal{D}^u} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\} + o_p(1)
\end{aligned}$$

where the second equality and the last equality hold if $var_{gh}[Q(h,\eta)] < \infty$ because by the Central Limit Theorem, $\frac{1}{\sqrt{N_u}} \sum_{i:O_i \in \mathcal{D}^u} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\} = O_p(1)$ and thus $\sqrt{N_u}\left\{\widehat{\theta}^{\dagger,u} - \theta(\eta)\right\} = O_p(1)$ and, by (2.28), $\left(\sqrt{\frac{N}{N_u}} - 1\right) = o_p(1)$ and $\left(1 - \sqrt{\frac{N_u}{N}}\right) = o_p(1)$.

Now let $\widehat{\theta}^{\dagger} \equiv \boldsymbol{U}^{-1} \sum_{u=1}^{\boldsymbol{U}} \widehat{\theta}^{\dagger,u}$. If (2.32) holds for all $u = 1, ..., \boldsymbol{U}$ then the last display gives

$$\begin{aligned}
\sqrt{n}\left\{\widehat{\theta}^{\dagger} - \theta(\eta)\right\} &= \sqrt{\frac{n}{N}}\left\{\boldsymbol{U}^{-1} \sum_{u=1}^{\boldsymbol{U}} \sqrt{N}\left[\widehat{\theta}^{\dagger,u} - \theta(\eta)\right]\right\} \\
&= \left\{\sqrt{\frac{n}{N}}\boldsymbol{U}^{-1} \sum_{u=1}^{\boldsymbol{U}} \frac{1}{\sqrt{N}} \sum_{i:O_i \in \mathcal{D}^u} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\}\right\} + o_p(1) \\
&= \left\{\frac{\sqrt{n}}{N\boldsymbol{U}} \sum_{i=1}^{n} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\}\right\} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\} \\
&+ \left(\frac{n}{N\boldsymbol{U}} - 1\right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{Q(h,\eta)_i - E_p[Q(h,\eta)]\} + o_p(1) \\
&= \mathbb{G}_n\{Q(h,\eta)\} + o_p(1)
\end{aligned}$$

where the last equality in the last display follows because $\left(\frac{n}{N\boldsymbol{U}} - 1\right) = \left(\frac{n}{\lfloor \frac{n}{\boldsymbol{U}} \rfloor \boldsymbol{U}} - 1\right) = o(1)$ and

by the Central Limit Theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{Q\left(h,\eta\right)_i - E_p\left[Q\left(h,\eta\right)\right]\right\} = O_p\left(1\right)$.

We have then arrived at the conclusion that, asymptotically, $\widehat{\theta}^{\dagger} - \theta\left(\eta\right)$ behaves, as $\widehat{\theta}^{\dagger,u} - \theta\left(\eta\right)$, like a sample average of $Q\left(h,\eta\right) - E_p\left[Q\left(h,\eta\right)\right]$, except that in the latter the average is over units in $\mathcal{D}^u$ whereas in the former the average is over units in the entire sample $\mathcal{D}$, thus proving the announced recovery of lost information obtained by $\widehat{\theta}^{\dagger}$.

The preceding analysis makes clear that the choice of $\boldsymbol{U}$ does not impact on the asymptotic distribution of $\widehat{\theta}^{\dagger}$so long as $\boldsymbol{U}$ is fixed. For the goal of this chapter, the choice of $\boldsymbol{U}$ is therefore inconsequential. However, the finite sample performance of $\widehat{\theta}^{\dagger}$ will be affected by the choice of $\boldsymbol{U}$. As $\boldsymbol{U}$ grows, the size of each nuisance estimation sample $\mathcal{D} - \mathcal{D}^u$ increases thus improving the estimation of the functions h and η. However, as $\boldsymbol{U}$ grows, the size of each main estimation sample $\mathcal{D}^u$ decreases, thus making the normal approximation of the distribution of $\widehat{\theta}^{\dagger,u}$ less accurate. The investigation of methods for selecting $\boldsymbol{U}$ so as to improve the finite sample performance of $\widehat{\theta}^{\dagger}$ is an interesting topic but it is beyond the scope of this work.

2.5.1 Series estimation of η

A concrete example of a "preferred estimator" used to estimate each function η_k in the preceding algorithm is the series estimator. Series estimation of the function $\eta_k\left(\cdot\right)$ is estimation via ordinary least squares on a covariate vector $\phi_k\left(\overline{L}_k\right) \equiv \left(\phi_{k,1}\left(\overline{L}_k\right),\ldots,\phi_{k,m_k}\left(\overline{L}_k\right)\right)'$ comprised by the first m_k elements of a dictionary of approximating functions $\left\{\phi_{k,j}\left(\cdot\right)\right\}_{1\leq j\leq\infty}$. In series estimation, m_k changes with the sample size n. We discuss the selection of m_k later in this subsection.

For the purposes of our calculations in the next sections, it is convenient to define the series estimator in the following way. Given the k^{th} nuisance estimation sample $\mathcal{N}^k$, define for any scalar function $r\left(O\right)$ of $O = \left(\overline{L}_{K+1}, \overline{A}_K\right)$,

$$\Pi^k\left[r\right]\left(\cdot\right) \equiv \widetilde{\psi}_k'\left(r\right)\phi_k\left(\cdot\right) \tag{2.33}$$

where

$$\widetilde{\psi}_k\left(r\right) \equiv \left[\sum_{i:O_i\in\mathcal{N}^k}\overline{I}_{k,i}\phi_k\left(\overline{L}_{k,i}\right)\phi_k\left(\overline{L}_{k,i}\right)'\right]^{-1}\left[\sum_{i:O_i\in\mathcal{N}^k}\overline{I}_{k,i}\phi_k\left(\overline{L}_{k,i}\right)r\left(O_i\right)\right]$$

Then, if in the algorithm for computing $\widehat{\theta}^u$ one employs series estimation, the estimated function $\widehat{\eta}_k\left(\cdot\right)$ is computed recursively for $k = K,...,1$, as

$$\widehat{\eta}_k\left(\cdot\right) \equiv \widehat{\psi}_k'\phi_k\left(\cdot\right) \tag{2.34}$$

where $\widehat{\psi}_k \equiv \widetilde{\psi}_k\left(\widehat{\eta}_{k+1}\right)$. Likewise, if $\widehat{\eta}_{k,MR}\left(\cdot\right)$ is computed by series estimation, then

$$\widehat{\eta}_{k,MR}\left(\cdot\right) \equiv \widehat{\psi}_{k,MR}'\phi_k\left(\cdot\right) \tag{2.35}$$

where $\widehat{\psi}_{k,MR} \equiv \widetilde{\psi}_k\left(q_{k+1}\left(\cdot,\cdot;\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right)\right)$.

The map $r \to \Pi^k\left[r\right]$ is an operator that maps functions of O to functions of $\overline{L}_k$. If the outcome $r\left(O\right)$ is not used in the selection of the dimension m_k of the vector $\phi_k\left(\cdot\right)$, then the operator Π^k is linear, that is,

$$\Pi^k\left[r_1 + r_2\right] = \Pi^k\left[r_1\right] + \Pi^k\left[r_2\right]$$

In Subsection 2.7.1 we will derive yet another expression for the drifts of $\widehat{\theta}$ and $\widehat{\theta}_{MR}$ when each $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ are computed as the result of recursively applying arbitrary linear operators $\Pi^k[\cdot]$. Using these expressions when $\Pi^k[\cdot]$ is the linear regression operator (2.33) we will calculate bounds on the rates of convergence of the drifts of $\widehat{\theta}$ and $\widehat{\theta}_{MR}$ under the assumption that the functions $\eta_k(\cdot)$ belong to Hölder balls with known smoothness order.

For $\mathcal{X} \subseteq \mathbb{R}^d$ and for $s \in (0,1]$, the Hölder ball of radius $\rho > 0$ and smoothness order s, denoted as $\mathcal{H}(\mathcal{X}; s, \rho)$, is defined as the set of all functions $f : \mathcal{X} \to \mathbb{R}$ such that

$$|f(x) - f(\widetilde{x})| \leq \rho \|x - \widetilde{x}\|^s$$

for all $x, \widetilde{x} \in \mathcal{X}$. For $s > 1$, $\mathcal{H}(\mathcal{X}; s, \rho)$ is defined as follows. For a $d-tuple$ $\alpha = (\alpha_1, \ldots, \alpha_d)$ of nonnegative integers, let

$$D^\alpha \equiv \partial_{x_1}^{\alpha_1} \ldots \partial_{x_d}^{\alpha_d}.$$

Then, $\mathcal{H}(\mathcal{X}; s, \rho)$ is defined as the set of all functions $f : \mathcal{X} \to \mathbb{R}$ such that f is $\lfloor s \rfloor$ times continuously differentiable and

$$|D^\alpha f(x) - D^\alpha f(\widetilde{x})| \leq \rho \|x - \widetilde{x}\|^{s - [s]} \quad \text{and} \quad \left|D^\beta f(x)\right| \leq \rho$$

for all $x, \widetilde{x} \in \mathcal{X}$. and all $d-tuples$ $\alpha = (\alpha_1, \ldots, \alpha_d)$ and $\beta = (\beta_1, \ldots, \beta_d)$ of nonnegative integers satisfying $\alpha_1 + \cdots + \alpha_d = \lfloor s \rfloor$ and $\beta_1 + \cdots + \beta_d \leq \lfloor s \rfloor$.

The asymptotic behavior of series estimators of a conditional mean function has been widely studied by several authors for different dictionaries, under the assumption that the conditional mean function belongs to a Hölder ball. Here, we briefly review the findings, which we will use in the derivation of our bounds for the drifts. Given a generic outcome Z, a $d \times 1$ vector X and n i.i.d. copies of (Z, X) with unknown cdf F, consider the series estimator $\widehat{g}(X)$ of $g(X) \equiv E(Z|X)$ defined as

$$\widehat{g}(X) \equiv \widehat{\beta}' p(X)$$

where $p(X) \equiv (p_1(X), \ldots, p_m(X))'$ is the vector of the first m elements of a dictionary $\{p_j(x)\}_{j \geq 1}$ and $\widehat{\beta} = \left[\mathbb{P}_n \left\{p(X) p(X)'\right\}\right]^{-1} \mathbb{P}_n \{p(X) Z\}$. Different authors have investigated the rate of convergence of

$$\|\widehat{g} - g\|_{L_2(F)}^2 \equiv \int \{\widehat{g}(x) - g(x)\}^2 dF_X(x)$$

for different dictionaries, under the assumption that $g(x)$ belongs to a Hölder ball $\mathcal{H}(\mathcal{X}; s, \rho)$ for some finite ρ (see, for example, [26], [7], [5]). In a recent article, Belloni et al. ([3]) provide a unifying theory which demonstrates that for dictionaries $\{p_j(x)\}_{j \geq 1}$ satisfying certain optimal approximation properties, which include Cohen-Daubechies-Vial wavelet, B-splines and local polynomial partition series, and with $m \asymp n^{\frac{d}{d+2s}}$, under regularity conditions, the series estimator satisfies

$$\|\widehat{g} - g\|_{L_2(F)}^2 = O_p\left(n^{-\frac{2s}{d+2s}}\right). \tag{2.36}$$

In the literature on non-parametric estimation the rate $n^{-\frac{2s}{d+2s}}$ is referred to as the "optimal" convergence rate for estimating conditional mean functions in a Hölder ball $\mathcal{H}(\mathcal{X}; s, \rho)$. This is because, if X has compact support and $\text{var}(Z|X) \leq \sigma^2$, the rate is minimax in that

$$\inf_{\widetilde{g}} \sup_{g : g \in \mathcal{H}(\mathcal{X}; s, \rho)} E_F\left\{\|\widetilde{g} - g\|_{L_2(F)}^2\right\} \gtrsim n^{-\frac{2s}{d+2s}}.$$

See Chapter 3.2 of [15].

In our derivation of the bounds for the drifts of $\widehat{\theta}$ and $\widehat{\theta}_{MR}$ that use series estimation of η_k we will assume that each η_k belongs to a Hölder ball $\mathcal{H}\left(\overline{\mathcal{L}}_k; s_k, \rho_k\right)$ with known smoothness s_k. Here, $\mathcal{L}_k$ is the sample space of L_k and, as defined in the notation section, $\overline{\mathcal{L}}_k \equiv \mathcal{L}_1 \times \cdots \times \mathcal{L}_k, k \in [K]$. We will assume that the number of dictionary elements m_k is non-data driven and grows with n at the rate $n^{\frac{d_k}{d_k+2s_k}}$ where $d_k = \dim\left(\overline{L}_k\right)$, so that the optimal rate $n^{-\frac{2s_k}{d_k+2s_k}}$ would be achieved by the series estimator if it could be computed using the -unavailable- outcomes $\eta_{k+1}\left(\overline{L}_{k+1}\right)$. By assuming that m_k is pre-specified, i.e. non-data driven, we can then express the series estimator as a linear operator Π^k and then exploit the special expressions for the drifts of $\widehat{\theta}$ and $\widehat{\theta}_{MR}$ when the estimators of η_k are obtained through recursive application of linear operators.

Admittedly, the results we will obtain are of theoretical interest but do not cover the realistic scenario in which the smoothness s_k order is unknown and m_k is selected adaptively, i.e. via a data driven procedure. In Section 2.9 we briefly discuss the reasons why our results do not extend straightforwardly to scenarios in which m_k is data driven, as would be the case if, for instance, m_k was selected from V-fold cross validation. Although we have not succeeded in producing rigorous results for the case in which m_k is selected adaptively, in Section 2.9 we will point to some results in the literature on cross validation, which suggest that our conclusions about the relative merits of $\widehat{\theta}$ vs $\widehat{\theta}_{MR}$ should remain valid when m_k is selected by V-fold cross-validation.

2.6 Global comparison of the drifts of the split-specific estimators $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ that use arbitrary non-parametric estimators of h and η

In this section we will apply the expressions for the drifts of arbitrary one step estimators of θ derived in Section 2.4 to make some general remarks on the comparison of the drifts of the split specific estimators $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ that use arbitrary non-parametric estimators of h and η.

In what follows, we fix u and we let $\mathcal{M} = \mathcal{D}^u$ denote the main estimation sample. We let N denote the size of $\mathcal{M}$. This is a minor change of notation with respect to the notation used in Section 2.5 where we have used N_u to denote the size of $\mathcal{D}^u$ and N to denote $\lfloor n/\mathbf{U} \rfloor$. This change greatly alleviates the notation from inconsequential subscripts. As in Section 2.5, we let the set $\mathcal{N} = \mathcal{D} - \mathcal{D}^u$ denote the nuisance estimation sample, $\mathcal{N}^1, ..., \mathcal{N}^K$ denote the partition of $\mathcal{N}$ into equal or nearly equal subsamples, and $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ denote the two distinct estimators of $\eta_k, k = 1, ..., K$, defined in the algorithm for computing $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ described in Section 2.5.

Recall from the introduction section that the drift of the one step estimator of θ based on N i.i.d. copies of $O = \left(\overline{L}_{K+1}, \overline{A}_K\right)$, that uses arbitrary estimators of $(h_1, ..., h_K)$ and $(\eta_1, ..., \eta_K)$, say $\left(h_1^\dagger, ..., h_K^\dagger\right)$ and $\left(\eta_1^\dagger, ..., \eta_K^\dagger\right)$, and with the law of L_0 estimated by its empirical cumulative distribution $\widehat{G}_0$, is equal to $\sqrt{N}$ times

$$E_p\left\{Q\left(h^\dagger, \eta^\dagger\right)\right\} - \theta\left(\eta\right) \tag{2.37}$$

where

$$\begin{aligned} Q\left(h^\dagger, \eta^\dagger\right) &\equiv q\left(O; h^\dagger, \eta^\dagger\right) \\ &\equiv \eta_1^\dagger\left(L_1\right) + \sum_{k=1}^{K}\left\{\prod_{j=1}^{k}\frac{I_j}{h_j^\dagger\left(\overline{L}_j\right)}\right\}\left\{\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right) - \eta_k^\dagger\left(\overline{L}_k\right)\right\} \end{aligned}$$

$\eta_{K+1}^\dagger\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$ and

$$E_p\left\{Q\left(h^\dagger, \eta^\dagger\right)\right\} \equiv \int q\left(o; h^\dagger, \eta^\dagger\right) dP\left(o\right)$$

The specific estimators $\left(h_1^\dagger, ..., h_K^\dagger\right)$ and $\left(\eta_1^\dagger, ..., \eta_K^\dagger\right)$ used by one step estimators $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$, namely $\left(\widehat{h}_1, ..., \widehat{h}_K\right)$ and $\left(\widehat{\eta}_1, ..., \widehat{\eta}_K\right)$, and $\left(\widehat{h}_1, ..., \widehat{h}_K\right)$ and $\left(\widehat{\eta}_{1,MR}, ..., \widehat{\eta}_{K,MR}\right)$ respectively, are computed using data from the nuisance estimation sample $\mathcal{N}$, so

$$\int q\left(o; \widehat{h}, \widehat{\eta}\right) dP\left(o\right) = E_p\left[Q\left(\widehat{h}, \widehat{\eta}\right)\middle|\mathcal{N}\right]$$

and

$$\int q\left(o; \widehat{h}, \widehat{\eta}_{MR}\right) dP\left(o\right) = E_p\left[Q\left(\widehat{h}, \widehat{\eta}_{MR}\right)\middle|\mathcal{N}\right].$$

Likewise, if in the expression $c^p\left(h^\dagger, \eta^\dagger\right)$ for (2.37) established in Lemma 8, the outcome $\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right)$ in $E_{g_k}\left\{\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right)\middle|\overline{A}_k = \overline{a}_k^*, \overline{L}_k\right\}$ is replaced with $\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)$ one obtains

$$\int \widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right) dF\left(\overline{L}_{k+1}|\overline{A}_k = \overline{a}_k^*, \overline{L}_k\right) = E_{g_k}\left\{\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)\middle|\overline{A}_k = \overline{a}_k^*, \overline{L}_k, \underline{\mathcal{N}}^{k+1}\right\}$$

where

$$\underline{\mathcal{N}}^{k+1} \equiv \cup_{j=k+1}^{K} \mathcal{N}^{j}.$$

Similarly, if in the expression $b^{p}\left(h^{\dagger},\eta^{\dagger}\right)$ for (2.37) established in Lemma 8, the outcome $Q_{k+1}\left(\underline{h}_{k+1}^{\dagger},\underline{\eta}_{k+1}^{\dagger}\right) \equiv q_{k+1}\left(\overline{L}_{K+1},\overline{I}_{k+1}^{K};\underline{h}_{k+1}^{\dagger},\underline{\eta}_{k+1}^{\dagger}\right)$ in $E_{\underline{g}_{k},\underline{h}_{k+1}}\left\{ Q_{k+1}\left(\underline{h}_{k+1}^{\dagger},\underline{\eta}_{k+1}^{\dagger}\right)\middle|\,\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k}\right\}$ is replaced by $\widehat{Q}_{k+1} \equiv Q_{k+1}\left(\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right) = q_{k+1}\left(\overline{L}_{K+1},\overline{I}_{k+1}^{K};\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right)$ one obtains

$$\begin{aligned}
&\int q_{k+1}\left(\overline{L}_{K+1},\overline{I}_{k+1}^{K};\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right) dF\left(\overline{L}_{K+1},\overline{I}_{k+1}^{K}|\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k}\right) = \\
&= E_{g_{k}}\left\{ Q_{k+1}\left(\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right)\middle|\,\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k},\underline{\mathcal{N}}^{k+1}\right\} \\
&\equiv E_{g_{k}}\left\{ \widehat{Q}_{k+1}\middle|\,\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k},\underline{\mathcal{N}}^{k+1}\right\}.
\end{aligned}$$

Now, using the expression $c^{p}\left(h^{\dagger},\eta^{\dagger}\right)$ for (2.37) evaluated at $\left(\widehat{h},\widehat{\eta}\right)$, we conclude that the drift of $\widehat{\theta}^{u}$ is $\sqrt{N}$ times the quantity

$$\begin{aligned}
&E_{p}\left[Q\left(\widehat{h},\widehat{\eta}\right)\middle|\,\mathcal{N}\right] - \theta\left(\eta\right) = \qquad (2.38)\\
&= \sum_{k=1}^{K} E_{p}\left\{ \overline{I}_{k}\left(\frac{1}{\pi^{k}}-\frac{1}{\widehat{\pi}^{k}}\right)\left[\widehat{\eta}_{k}\left(\overline{L}_{k}\right) - E_{g_{k}}\left\{\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)\middle|\,\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k},\underline{\mathcal{N}}^{k+1}\right\}\right]\middle|\,\underline{\mathcal{N}}^{k}\right\}
\end{aligned}$$

and using the expression $b^{p}\left(h^{\dagger},\eta^{\dagger}\right)$ for (2.37) evaluated at $\left(\widehat{h},\widehat{\eta}_{MR}\right)$, we conclude that the drift of $\widehat{\theta}_{MR}^{u}$ is $\sqrt{N}$ times the quantity

$$\begin{aligned}
&E\left\{ Q_{1}\left(\widehat{h},\widehat{\eta}_{MR}\right)\middle|\,\mathcal{N}\right\} - \theta\left(\eta\right) = \qquad (2.39)\\
&= \sum_{k=1}^{K} E_{p}\left\{ \frac{\overline{I}_{k}}{\pi^{k-1}}\left(\frac{1}{h_{k}\left(\overline{L}_{k}\right)}-\frac{1}{\widehat{h}_{k}\left(\overline{L}_{k}\right)}\right)\left[\widehat{\eta}_{k,MR}\left(\overline{L}_{k}\right) - E_{\underline{g}_{k},\underline{h}_{k+1}}\left\{\widehat{Q}_{k+1}\middle|\,\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k},\underline{\mathcal{N}}^{k+1}\right\}\right]\middle|\,\underline{\mathcal{N}}^{k}\right\}
\end{aligned}$$

The preceding expressions for the drifts of $\widehat{\theta}^{u}$ and $\widehat{\theta}_{MR}^{u}$ are sums, from $k=1$ to K, of terms that are expectations of objects involving the product of two estimation errors. For both $\widehat{\theta}^{u}$ and $\widehat{\theta}_{MR}^{u}$ one of these estimation errors is a *pseudo-outcome estimation error* i.e. the difference between the true and estimated conditional mean of the pseudo-outcome used by each procedure in place of the unknown outcome $\eta_{k+1}\left(\overline{L}_{k+1}\right)$:

$$\widehat{\eta}_{k}\left(\overline{L}_{k}\right) - E_{g_{k}}\left\{\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)\middle|\,\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k},\underline{\mathcal{N}}^{k+1}\right\}$$

and

$$\widehat{\eta}_{k,MR}\left(\overline{L}_{k}\right) - E_{\underline{g}_{k},\underline{h}_{k+1}}\left\{\widehat{Q}_{k+1}\middle|\,\overline{A}_{k}=\overline{a}_{k}^{*},\overline{L}_{k},\underline{\mathcal{N}}^{k+1}\right\}$$

In the drift for $\widehat{\theta}_{MR}^{u}$ the second estimation error is $\frac{1}{h_{k}}-\frac{1}{\widehat{h}_{k}}$, whereas in the drift for $\widehat{\theta}^{u}$ it is

$$\frac{1}{\pi^{k}}-\frac{1}{\widehat{\pi}^{k}} = \sum_{j=1}^{k}\frac{1}{\pi^{j-1}}\left(\frac{1}{h_{j}\left(\overline{L}_{j}\right)}-\frac{1}{\widehat{h}_{j}\left(\overline{L}_{j}\right)}\right)\frac{1}{\widehat{\pi}_{j+1}^{k}}$$

This decomposition of $\frac{1}{\pi^k} - \frac{1}{\widehat{\pi}^k}$ shows that both drifts involve a term for each k corresponding to the product of the *pseudo-outcome estimation error* for that k, times the estimation error for h_k. Yet, for each k, the drift of $\widehat{\theta}^u$ involves additional terms corresponding to the product of the *pseudo-outcome estimation error* for that k, times the estimation error for each $h_j, j < k$.

Although the drift of $\widehat{\theta}^u$ has many more terms than the drift of $\widehat{\theta}^u_{MR}$, at this level of generality, i.e. without specifying the specific non-parametric procedure used for estimating $E_{g_k}\left\{\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)\middle|\, \overline{A}_k = \overline{a}^*_k, \overline{L}_k, \underline{\mathcal{N}}^{k+1}\right\}$ and $E_{\underline{g}_k, \underline{h}_{k+1}}\left\{\widehat{Q}_{k+1}\middle|\, \overline{A}_k = \overline{a}^*_k, \overline{L}_k, \underline{\mathcal{N}}^{k+1}\right\}$, it does not seem possible to quantitatively compare the size of the drifts of the two estimators. Notice that even if one were to use the same non-parametric procedure for estimating both conditional means (but applied to the two distinct outcomes $\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)$ and $\widehat{Q}_{k+1}$), we could nevertheless not make much progress without knowing the specific procedure, since the pseudo-outcomes themselves depend on the procedures that are applied for $k+1, k+2, ..., K$.

In the next section we will show that when the non-parametric procedure used to estimate the $\eta'_k s$ is series estimation, it is possible to derive yet two more expressions for the drift that permit direct comparisons of the rates of convergence of upper bounds for the drifts of the two estimators, under the assumption that the $\eta'_k s$ lie in Hölder balls.

2.7 Analysis of the drifts of $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ when η_k is estimated via series estimation

2.7.1 Formulae for the drifts of $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ when η_k is estimated via an arbitrary linear estimator.

The aim of this section is to provide rates of convergence on bounds for the drifts of $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ that compute each $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ by series estimation. To calculate these bounds, we will start by deriving new expressions for the drifts that are valid when $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ are obtained by applying a linear map to some transformation r of the data $O = \left(\overline{A}_K, \overline{L}_{K+1}\right)$.

Given the data in the subsample $\mathcal{N}^k$, the series estimators $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ of η_k defined in (2.34) and (2.35) can be written as

$$\widehat{\eta}_k(\cdot) \equiv \Pi^k\left[\widehat{\eta}_{k+1}\right](\cdot) \tag{2.40}$$

$$\widehat{\eta}_{k,MR}(\cdot) \equiv \Pi^k\left[q_{k+1}\left(\cdot,\cdot;\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right)\right](\cdot) \tag{2.41}$$

where recall that, as defined in (2.33), for any function r of the data O,

$$\Pi^k[r](\cdot) \equiv \widetilde{\psi}'_k(r)\,\phi_k(\cdot) \tag{2.42}$$

with

$$\widetilde{\psi}_k(r) \equiv \left[\sum_{i:O_i\in\mathcal{N}^k} \overline{I}_{k,i}\phi_k\left(\overline{L}_{k,i}\right)\phi_k\left(\overline{L}_{k,i}\right)'\right]^{-1}\left[\sum_{i:O_i\in\mathcal{N}^k} \overline{I}_{k,i}\phi_k\left(\overline{L}_{k,i}\right)r(O_i)\right]$$

If, as we shall assume until Section 2.9, the dimension m_k of $\phi_k\left(\overline{L}_k\right)$ does not depend on r, -as it is the case if m_k is a predetermined, i.e. non-data driven, increasing function of $n-$, then $\widetilde{\psi}_k(r)$ is a linear function of r and therefore $\Pi^j[\cdot]$ operates linearly on r.

As a second example of an estimator obtained by applying a linear operator, consider the multivariate Nadaraya Watson kernel regression estimator $\widehat{\eta}_k$ of η_k ([20], [21]). This estimator is also of the form (2.40) where now

$$\Pi^k[r](\cdot) \equiv \frac{\sum_{i:O_i\in\mathcal{N}^k}\mathcal{K}_\delta\left(\overline{L}_{k,i}-\cdot\right)r(O_i)}{\sum_{i:O_i\in\mathcal{N}^k}\mathcal{K}_\delta\left(\overline{L}_{k,i}-\cdot\right)} \tag{2.43}$$

with $\mathcal{K}_\delta(u)$ a multidimensional kernel of dimension $d_k = \dim\left(\overline{L}_k\right)$. If the bandwidth vector δ is not data driven, then $\Pi^j[\cdot]$ defined in (2.43) operates linearly on functions r of O.

An example of a non-parametric estimator $\widehat{\eta}_k(\cdot)$ for which there exists no linear operator Π^k such that (2.40) holds, even when its tunning parameter λ is assumed to be non-data driven, is the Lasso estimator $\widehat{\eta}_k(\cdot) \equiv \widetilde{\psi}_k^{LASSO}\left(\widehat{\eta}_{k+1}\right)\phi_k(\cdot)$ where for any $r(O)$,

$$\widetilde{\psi}_k^{LASSO}(r) = \arg\min_{\psi_k}\left\{\frac{1}{\#\mathcal{N}^k}\sum_{i:O_i\in\mathcal{N}^k}\left[\left\{r(O_i)-\psi'_k\overline{L}_{k,i}\right\}^2\right] + \lambda\sum_{j=1}^{\dim\left(\overline{L}_k\right)}|\psi_{k,j}|\right\}$$

To state the next theorem which establishes the special expressions for the drift when the estimators $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ are obtained by applying a linear map, we need the following definitions.

In the following definitions, $h^{\dagger} \equiv \left(h_1^{\dagger}, \ldots, h_K^{\dagger}\right)$ is a given fixed, non-random, function. Furthermore, for $j \in [K]$, $\Pi^j[\cdot]$ stands for a given linear map, from functions of O to functions of $\overline{L}_j$.That is, $\Pi^j[\cdot]$ maps functions, say $r(\cdot)$, of $O = \left(\overline{A}_K, \overline{L}_{K+1}\right)$ to functions $\Pi^j[r](\cdot)$ of $\overline{L}_j$, and in addition, for any two functions r_1, r_2 of O, $\Pi^j[r_1 + r_2] = \Pi^j[r_1] + \Pi^j[r_2]$.

For $0 \leq j < u \leq K$, let

$$\nabla_{j,u} \equiv \frac{\overline{I}_{j+1}^{u-1}}{\pi_{j+1}^{\dagger(u-1)}} \left(\frac{I_u}{h_u\left(\overline{L}_u\right)} - \frac{I_u}{h_u^{\dagger}\left(\overline{L}_u\right)} \right)$$

In particular, $\nabla_{j,j+1} \equiv \left(\frac{I_{j+1}}{h_{j+1}\left(\overline{L}_{j+1}\right)} - \frac{I_{j+1}}{h_{j+1}^{\dagger}\left(\overline{L}_{j+1}\right)} \right)$ and $\nabla_{0,k} \equiv \frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}} \left(\frac{I_k}{h_k\left(\overline{L}_k\right)} - \frac{I_k}{h_k^{\dagger}\left(\overline{L}_k\right)} \right)$.

For any function r of O define

1. for $j \in [K]$,

$$\Pi_{DR}^j[r] \equiv \Pi^j\left[E_p\left(\frac{I_{j+1}}{h_{j+1}\left(\overline{L}_{j+1}\right)} \, r(O) \;\middle|\; \overline{A}_j = \overline{a}_j^*, \overline{L}_{j+1} = \cdot \right) \right]$$

   where $I_{K+1} \equiv h_{K+1}\left(\overline{L}_{K+1}\right) \equiv 1$.

2. for $1 \leq j < k \leq K$,

$$\Pi_{DR,j,k}[r] \equiv \left(\Pi_{DR}^j \circ \ldots \circ \Pi_{DR}^{k-1} \right)[r]$$

   where here and throughout $\circ$ denotes the composition operation. In particular, $\Pi_{DR,j,j+1}[\cdot] \equiv \Pi_{DR}^j[\cdot]$.

3. for $1 \leq j < u \leq K$,

$$\Pi_{MR,j,u}[r] \equiv \Pi^j\left[E_p\left(\nabla_{j,u} r(O) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_{j+1} = \cdot \right) \right]$$

   Note that

$$\Pi_{MR,j,j+1}[r] = \Pi^j\left[E_p\left\{ \left(\frac{I_{j+1}}{h_{j+1}\left(\overline{L}_{j+1}\right)} - \frac{I_{j+1}}{h_{j+1}^{\dagger}\left(\overline{L}_{j+1}\right)} \right) r(O) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_{j+1} = \cdot \right\} \right]$$

4. For $1 \leq r_1 < r_2 < \ldots < r_u \leq K$,

$$\Pi_{MR,r_1,r_2,\ldots,r_u}[r] \equiv \left(\Pi_{MR,r_1,r_2} \circ \ldots \circ \Pi_{MR,r_{u-2},r_{u-1}} \circ \Pi_{MR,r_{u-1},r_u} \right)[r]$$

We also need the following definitions.

a) for $j = K, K-1, ..., 1$, recursively define

$$\widetilde{\eta}_{j,DR} \equiv \Pi^j\left[\widetilde{\eta}_{j+1,DR}\right]$$

   where $\widetilde{\eta}_{K+1,DR}\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$.

b) for $j \in [K]$ define

$$\eta_{j,DR} \equiv \Pi^j \left[\eta_{j+1}\right]$$

where $\eta_{K+1}\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$ and $\eta_1, ..., \eta_K$ are the true unknown iterated conditional expectations as defined in (2.4).

c) For $j = K, K-1, ..., 1$, recursively define

$$\widetilde{\eta}_{j,MR} \equiv \Pi^j \left[q_{j+1}\left(\cdot, \cdot; \underline{h}^{\dagger}_{j+1}, \widetilde{\underline{\eta}}_{j+1,MR}\right)\right]$$

where $q_{K+1}\left(\overline{L}_{K+1}, \overline{I}^K_K; \underline{h}^{\dagger}_{K+1}, \widetilde{\underline{\eta}}_{K+1,MR}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$ and where, recall $q_j\left(\overline{L}_{K+1}, \overline{I}^K_j; \underline{h}^{\dagger}_j, \underline{\eta}^{\dagger}_j\right)$ is defined in (2.13), $j \in [K]$.

d) For $j \in [K]$, define

$$\eta_{j,MR} \equiv \eta_{j,DR} + \Pi^j \left[q_{j+1}\left(\cdot, \cdot; \underline{h}^{\dagger}_{j+1}, \widetilde{\underline{\eta}}_{j+1,MR}\right) - E_p\left\{Q_{j+1}\left(\underline{h}^{\dagger}_{j+1}, \widetilde{\underline{\eta}}_{j+1,MR}\right) \middle| \overline{A}_j = \overline{a}^*_j, \overline{L}_{j+1} = \cdot\right\}\right].$$

Note that by definition of the maps $\Pi^j, j \in [K]$, the functions $\widetilde{\eta}_{j,DR}, \eta_{j,DR}, \widetilde{\eta}_{j,MR}$ and $\eta_{j,MR}$ are functions of $\overline{L}_j$. Furthermore, when Π^j is equal to the linear operator (2.42) and $h^{\dagger}$ is replaced by the estimator $\widehat{h}$ used to compute $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$, then $\widetilde{\eta}_{j,DR}$ and $\widetilde{\eta}_{j,MR}$ coincide with the series estimators $\widehat{\eta}_j$ and $\widehat{\eta}_{j,MR}$ used to compute $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ respectively.

Throughout the rest of the chapter, unless unclear, to alleviate notation we write functions without explicitly writing the variables where they are evaluated. Thus, for instance, $E_p\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h^{\dagger}_k}\right)\Pi_{DR,k,j}\left[\eta_{j,DR} - \eta_j\right]\right\}$ stands for $E_p\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k\left(\overline{L}_k\right)} - \frac{I_k}{h^{\dagger}_k\left(\overline{L}_k\right)}\right)\Pi_{DR,k,j}\left[\eta_{j,DR} - \eta_j\right]\left(\overline{L}_k\right)\right\}$ where, recall, $\pi^{\dagger(k-1)}$ was defined in Section 2.2 as $\pi^{\dagger(k-1)} \equiv \prod_{r=1}^{k-1} h^{\dagger}_r\left(\overline{L}_r\right)$.

The proof of the following theorem is given in Appendix B.4.

Theorem 2 *Let $\widetilde{\eta}_{k,DR}$ and $\widetilde{\eta}_{k,MR}$, $k \in [K]$ be the random variables defined in (a) and (c) above and $h^{\dagger}_k, k \in [K]$, be an arbitrary probability of $A_k = a^*_k$ given $\overline{A}_{k-1} = \overline{a}^*_{k-1}$ and $\overline{L}_k, k \in [K]$. The following identities hold.*

1. *for $k \in [K]$*

$$\widetilde{\eta}_{k,DR} - \eta_k = \eta_{k,DR} - \eta_k + \sum_{j=k+1}^{K} \Pi_{DR,k,j}\left[\eta_{j,DR} - \eta_j\right] \tag{2.44}$$

2.

$$\begin{aligned} a^p\left(h^{\dagger}, \widetilde{\eta}_{DR}\right) &\equiv \sum_{k=1}^{K} E_p\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h^{\dagger}_k}\right)\left(\widetilde{\eta}_{k,DR} - \eta_k\right)\right\} \\ &= \sum_{k=1}^{K} E_p\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h^{\dagger}_k}\right)\left(\eta_{k,DR} - \eta_k\right)\right\} \\ &+ \sum_{1 \le k < j \le K} E_p\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h^{\dagger}_k}\right)\Pi_{DR,k,j}\left[\eta_{j,DR} - \eta_j\right]\right\} \end{aligned}$$

3.

$$\sum_{k=1}^{K} E_p \left\{ \frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger} \right) (\widetilde{\eta}_{k,MR} - \eta_k) \middle| L_1 \right\} \tag{2.45}$$
$$= \sum_{k=1}^{K} E_p \left\{ \frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger} \right) (\eta_{k,MR} - \eta_k) \middle| L_1 \right\}$$
$$+ \sum_{\substack{\emptyset \neq \{r_1,...,r_u\} \subseteq [K-1] \\ r_1 < r_2 < ... < r_u}} \sum_{j=r_u+1}^{K} E_p \left\{ \nabla_{0,r_1} \Pi_{MR,r_1,r_2,...,r_u,j} \left[\eta_{j,MR} - \eta_j \right] \middle| L_1 \right\}$$

4.

$$a^p \left(h^\dagger, \widetilde{\eta}_{MR} \right) \equiv \sum_{k=1}^{K} E_p \left\{ \frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger} \right) (\widetilde{\eta}_{k,MR} - \eta_k) \right\}$$
$$= \sum_{k=1}^{K} E_p \left\{ \frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger} \right) (\eta_{k,MR} - \eta_k) \right\}$$
$$+ \sum_{\substack{\emptyset \neq \{r_1,...,r_u\} \subseteq [K-1] \\ r_1 < r_2 < ... < r_u}} \sum_{j=r_u+1}^{K} E_p \left\{ \nabla_{0,r_1} \Pi_{MR,r_1,r_2,...,r_u,j} \left[\eta_{j,MR} - \eta_j \right] \right\}$$

Notational remark: in parts (3) and (4) of the Theorem, the summation $\sum_{\substack{\emptyset \neq \{r_1,...,r_u\} \subseteq [K-1] \\ r_1 < r_2 < ... < r_u}}$ is over all non-empty subsets of $[K-1] \equiv \{1, ..., K-1\}$, where $r_1 < r_2 < ... < r_u$ denote the ordered elements of a subset with cardinality u.

In the special case in which $K = 2$, the expressions for the drift given in parts (2) and (4) of Theorem 2 are

$$a^p \left(h^\dagger, \widetilde{\eta}_{DR} \right) \equiv \tag{2.46}$$
$$\equiv E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) (\eta_{1,DR} - \eta_1) \right\}$$
$$+ E_p \left\{ \frac{I_1}{h^{\dagger 1}} \left(\frac{I_2}{h_2} - \frac{I_2}{h_2^\dagger} \right) (\eta_{2,DR} - \eta_2) \right\}$$
$$+ E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) \Pi^1 \left[E_p \left\{ \frac{I_2}{h_2} (\eta_{2,DR} - \eta_2) \middle| a_1^*, \overline{L}_2 = \cdot \right\} \right] \right\}$$

and

$$
\begin{aligned}
& a^p\left(h^{\dagger}, \widetilde{\eta}_{MR}\right) \equiv \\
& \equiv E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{h_1^{\dagger}}\right)\left(\eta_{1,MR}-\eta_1\right)\right\} \\
& +E_p\left\{\frac{I_1}{h^{\dagger 1}}\left(\frac{I_2}{h_2}-\frac{I_2}{h_2^{\dagger}}\right)\left(\eta_{2,MR}-\eta_2\right)\right\} \\
& +E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{h_1^{\dagger}}\right)\Pi^1\left[E_p\left\{\left(\frac{I_2}{h_2}-\frac{I_2}{h_2^{\dagger}}\right)\left(\eta_{2,MR}-\eta_2\right)\middle| a_1^*, \overline{L}_2=\cdot\right\}\right]\right\}
\end{aligned}
\tag{2.47}
$$

When $K=3$, these formulae are

$$
\begin{aligned}
& a^p\left(h^{\dagger}, \widetilde{\eta}_{DR}\right) \equiv \sum_{k=1}^{3} E_p\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger k-1}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^{\dagger}}\right)\left(\eta_{k,DR}-\eta_k\right)\right\} \\
& +E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{h_1^{\dagger}}\right)\Pi^1\left[E_p\left\{\frac{I_2}{h_2}\left(\eta_{2,DR}-\eta_2\right)\middle| a_1^*, \overline{L}_2=\cdot\right\}\right]\right\} \\
& +E_p\left\{\frac{I_1}{h^{\dagger 1}}\left(\frac{I_2}{h_2}-\frac{I_2}{h_2^{\dagger}}\right)\Pi^2\left[E_p\left\{\frac{I_3}{h_3}\left(\eta_{3,DR}-\eta_3\right)\middle| \overline{a}_2^*, \overline{L}_3=\cdot\right\}\right]\right\} \\
& +E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{h_1^{\dagger}}\right)\Pi^1\left[E_p\left\{\frac{I_2}{h_2}\Pi^2\left[E_p\left\{\frac{I_3}{h_3}\left(\eta_{3,DR}-\eta_3\right)\middle| \overline{a}_2^*, \overline{L}_3=\cdot\right\}\right]\middle| a_1^*, \overline{L}_2=\cdot\right\}\right]\right\}
\end{aligned}
\tag{2.48}
$$

and

$$
\begin{aligned}
& a^p\left(h^{\dagger}, \widetilde{\eta}_{MR}\right) \equiv \sum_{k=1}^{3} E_p\left[\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^{\dagger}}\right)\left(\eta_{k,MR}-\eta_k\right)\right] \\
& +E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{h_1^{\dagger}}\right)\Pi^1\left[E_p\left\{\left(\frac{I_2}{h_2}-\frac{I_2}{h_2^{\dagger}}\right)\left(\eta_{2,MR}-\eta_2\right)\middle| a_1^*, \overline{L}_2=\cdot\right\}\right]\right\} \\
& +E_p\left\{\frac{I_1}{h^{\dagger 1}}\left(\frac{I_2}{h_2}-\frac{I_2}{h_2^{\dagger}}\right)\Pi^2\left[E_p\left\{\left(\frac{I_3}{h_3}-\frac{I_3}{h_3^{\dagger}}\right)\left(\eta_{3,MR}-\eta_3\right)\middle| \overline{a}_2^*, \overline{L}_3=\cdot\right\}\right]\right\} \\
& +E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{h_1^{\dagger}}\right)\Pi^1\left[E_p\left\{\frac{I_2}{h_2^{\dagger}}\left(\frac{I_3}{h_3}-\frac{I_3}{h_3^{\dagger}}\right)\left(\eta_{3,MR}-\eta_3\right)\middle| a_1^*, \overline{L}_2=\cdot\right\}\right]\right\} \\
& +E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{h_1^{\dagger}}\right)\Pi^1\left[E_p\left\{\left(\frac{I_2}{h_2}-\frac{I_2}{h_2^{\dagger}}\right)\Pi^2\left[E_p\left\{\left(\frac{I_3}{h_3}-\frac{I_3}{h_3^{\dagger}}\right)\left(\eta_{3,MR}-\eta_3\right)\middle| \overline{a}_2^*, \overline{L}_3=\cdot\right\}\right]\middle| a_1^*, \overline{L}_2=\cdot\right\}\right]\right\}
\end{aligned}
\tag{2.49}
$$

For calculating bounds on the rates of convergence of the drift of $\widehat{\theta}^u$ under smoothness assumptions on the functions η_k, the formula in part (2) of Theorem 2 is more convenient than the formulae given in Section 2.4, because it is expressed in terms of the estimation errors $\eta_{k,DR}-\eta_k$ of the ideal, but unfeasible, estimator $\eta_{k,DR}=\Pi^k\left[\eta_{k+1}\right]$ of $\eta_k \equiv E_p\left\{\eta_{k+1}\left(\overline{L}_{k+1}\right)|\overline{A}_k=\overline{a}_k, \overline{L}_k=\cdot\right\}$ that use the

true, but unknown, outcomes $\eta_{k+1}\left(\overline{L}_{k+1,i}\right)$, for units i in $\mathcal{N}^k$. One can then use known results for rates of convergence of the estimation error of specific -linear- non-parametric regression estimators of a conditional mean function under smoothness assumptions on it to bound $\eta_{k,DR}-\eta_k$. We follow precisely this strategy in the next subsection to derive bounds for the drift of $\widehat{\theta}^u$. On the other hand, the formula in part (4) of Theorem 2 expresses the drift of $\widehat{\theta}^u_{MR}$ in terms of

$$\begin{aligned}\eta_{k,MR}-\eta_k &= \eta_{k,DR}-\eta_k \\ &+\Pi^k\left[q_{k+1}\left(\cdot,\cdot;\underline{h}^{\dagger}_{k+1},\widetilde{\underline{\eta}}_{k+1,MR}\right)-E_p\left\{Q_{k+1}\left(\underline{h}^{\dagger}_{k+1},\widetilde{\underline{\eta}}_{k+1,MR}\right)\middle|\overline{A}_k=\overline{a}^*_k,\overline{L}_{k+1}=\cdot\right\}\right]\end{aligned}$$

where $\eta_{k,DR}-\eta_k$ is the same estimation error as the one appearing in the expression for the drift of $\widehat{\theta}^u$. In the next subsection we show that, for series estimation, this estimation error dominates the size of $\eta_{k,MR}-\eta_k$.

2.7.2 Application of the formulae to the analysis of bounds for the drifts when η_k is estimated via series estimation

In this section we will derive bounds, under the assumption that each η_k belongs to a Hölder ball with known smoothness order, for the rates of convergence of the drifts of $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ in the special case in which both estimators use series estimators of η_k. We will present a rigorous analysis for the cases $K=2$ and $K=3$. Our rigorous analysis can be generalized to arbitrary K but at the cost of complicating significantly the notation. Thus, for an arbitrary K we will state without proof a conjecture on what the convergence rates should be for the bounds of the drifts of $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$.

To derive the rate of convergence of upper bounds for the drift of $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$, we will use the formulae stated in parts (2) and (4) of Theorem 2 for the special linear operator $\Pi^k[\cdot]$ defined in (2.42) -whose very definition depends on the data in subsample $\mathcal{N}^k-$. Because, as pointed out earlier, these formulae depend on the estimation errors $\eta_{k,DR}-\eta_k$, our results will depend on the rates of convergence to 0 of these estimation errors (in an appropriate norm) as the size of $\mathcal{N}^k$ grows to ∞. However, because the ratio between the size of $\mathcal{N}^k$ and the size n of the entire sample $\mathcal{D}$ is bounded below and above by non-zero constants, then the rates of convergence of the estimation errors can be expressed in terms of n rather than it terms of the size of $\mathcal{N}^k$. In our calculations we will therefore express the rates of convergence for our bounds on the drifts in terms of n rather than in terms of the sizes of each $\mathcal{N}^k$, as in doing so we will avoid unnecessary complications of notation.

Throughout this section we therefore assume that $\Pi^k[\cdot], k\in[K]$ is defined as in (2.42) . Letting $\mathcal{N}$ denote the nuisance estimation sample $\mathcal{D}-\mathcal{D}^u$, $\widehat{h}_k$ be the estimator of h_k used by the estimators $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$, and $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ the series estimators of η_k used by $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ respectively, define for $k=1,2,3$,

$$\begin{aligned}\delta^{DR}_k &\equiv E_p\left\{\frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}}\left(\frac{I_k}{h_k}-\frac{I_k}{\widehat{h}_k}\right)(\eta_{k,DR}-\eta_k)\middle|\mathcal{N}\right\},\\ \delta^{MR}_k &\equiv E_p\left\{\frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}}\left(\frac{I_k}{h_k}-\frac{I_k}{\widehat{h}_k}\right)(\eta_{k,MR}-\eta_k)\middle|\mathcal{N}\right\},\end{aligned}$$

with $\widehat{\pi}^{k-1} = \Pi_{r=1}^{k-1} \widehat{h}_r$. Define also,

$$\xi_{1,2}^{DR} \equiv E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[\eta_{2,DR} - \eta_2 \right] \middle| \mathcal{N} \right\},$$
$$\xi_{1,2}^{MR} \equiv E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[E_p \left\{ \left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2} \right) \left(\eta_{2,MR} - \eta_2 \right) \middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\}$$

$$\xi_{2,3}^{DR} \equiv E_p \left\{ \frac{I_1}{\widehat{h}_1} \left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2} \right) \Pi^2 \left[\eta_{3,DR} - \eta_3 \right] \middle| \mathcal{N} \right\}$$
$$\xi_{2,3}^{MR} \equiv E_p \left\{ \frac{I_1}{\widehat{h}_1} \left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2} \right) \Pi^2 \left[E_p \left\{ \left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3} \right) \left(\eta_{3,MR} - \eta_3 \right) \middle| \overline{A}_2 = \overline{a}_2^*, \overline{L}_3 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\}$$

$$\xi_{1,3}^{DR} \equiv E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[\Pi^2 \left[\eta_{3,DR} - \eta_3 \right] \right] \middle| \mathcal{N} \right\}$$
$$\xi_{1,3}^{MR} \equiv E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[E_p \left\{ \frac{I_2}{\widehat{h}_2} \left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3} \right) \left(\eta_{3,MR} - \eta_3 \right) \middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\}$$

and $\varkappa_{1,2,3}^{MR} \equiv$
$$\equiv E_p \left\{ \left(\tfrac{I_1}{h_1} - \tfrac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[E_p \left\{ \left(\tfrac{I_2}{h_2} - \tfrac{I_2}{\widehat{h}_2} \right) \Pi^2 \left[E_p \left\{ \left(\tfrac{I_3}{h_3} - \tfrac{I_3}{\widehat{h}_3} \right) \left(\eta_{3,MR} - \eta_3 \right) \middle| \overline{a}_2^*, \overline{L}_3 = \cdot, \mathcal{N} \right\} \right] \middle| a_1^*, \overline{L}_2 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\}.$$

Equations 2.46 and 2.47 with $h^\dagger = \widehat{h}$, and with $\widetilde{\eta}_{k,DR}$ and $\widetilde{\eta}_{k,MR}$ being the series estimators $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ of η_k used by $\widehat{\theta}^u$ and $\widehat{\theta}_{MR}^u$ respectively, give that for $K = 2$

$$E_p \left\{ Q_1 \left(\widehat{h}, \widehat{\eta} \right) \middle| \mathcal{N} \right\} - \theta \left(\eta \right) \equiv \delta_1^{DR} + \delta_2^{DR} + \xi_{12}^{DR}$$

and

$$E_p \left\{ Q_1 \left(\widehat{h}, \widehat{\eta}_{MR} \right) \middle| \mathcal{N} \right\} - \theta \left(\eta \right) \equiv \delta_1^{MR} + \delta_2^{MR} + \xi_{12}^{MR},$$

while equations 2.48 and 2.49 give that for $K = 3$,

$$E_p \left\{ Q_1 \left(\widehat{h}, \widehat{\eta} \right) \middle| \mathcal{N} \right\} - \theta \left(\eta \right) = \delta_1^{DR} + \delta_2^{DR} + \delta_3^{DR} + \xi_{1,2}^{DR} + \xi_{1,3}^{DR} + \xi_{2,3}^{DR} \tag{2.50}$$

and

$$E_p \left\{ Q_1 \left(\widehat{h}, \widehat{\eta}_{MR} \right) \middle| \mathcal{N} \right\} - \theta \left(\eta \right) = \delta_1^{MR} + \delta_2^{MR} + \delta_3^{MR} + \xi_{1,2}^{MR} + \xi_{1,3}^{MR} + \xi_{2,3}^{MR} + \varkappa_{1,2,3}^{MR} \tag{2.51}$$

We will now compute bounds for the rates of convergence to 0 of $a^p \left(\widehat{h}, \widehat{\eta} \right)$ and $a^p \left(\widehat{h}, \widehat{\eta}_{MR} \right)$ under the assumption that each $\eta_k \left(\cdot \right)$ belongs to a Hölder ball of finite radius and known smoothness order and when a specific subset of the following conditions below hold for each k - the subset depending on the estimator $\widehat{\eta}$ or $\widehat{\eta}_{MR}$ and on whether $K = 2$ or $K = 3$-.

Condition Hölder(k) $\eta_k \left(\cdot \right)$ lies in a Hölder ball $\mathcal{H} \left(\overline{\mathcal{L}}_k; s_k, \rho_k \right)$ with $\rho_k < \infty$ and known smoothness order $s_k > 0$.

Condition R(k) Assumptions 2 - 5 of Lemma 17 in Appendix B.5 are satisfied with $\overline{L}_k$ in the place of X, $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ in the place of Y, $\eta_k(\cdot)$ in the place of $g(\cdot)$, the distribution of $\overline{L}_{k+1}\big|\overline{A}_k = \overline{a}_k^*$ in the place of the distribution of (Y, X'), $\phi_k\left(\overline{l}_k\right)$ in the place of $p(x)$, $\mathcal{H}\left(\overline{\mathcal{L}}_k; s_k, \rho_k\right)$ in the place of $\mathcal{G}$ and $\gamma_k \equiv s_k / \dim\left(\overline{L}_k\right)$ in the place of γ.

Condition B(k) there exists $\xi > 0$ such that $h_k\left(\overline{l}_k\right) > \xi$ for all $\overline{l}_k$.

Condition Hconvergence(k) $\left\|\widehat{h}_k - h_k\right\|_\infty = o_p(1)$.

Condition HrateInf(k) $\left\|\widehat{h}_k - h_k\right\|_\infty = O_p\left(\alpha_{k,n}\right)$ for some sequence $\alpha_{k,n}$ converging to 0 as n goes to ∞.

Condition HrateL2(k) $\sqrt{E_p\left\{\overline{I}_k\left(\widehat{h}_k - h_k\right)^2\middle|\mathcal{N}\right\}} = O_p\left(\beta_{k,n}\right)$ for some sequence $\beta_{k,n}$ converging to 0 as n goes to ∞.

In what follows whenever Condition Hölder(k) holds and d_k denotes the dimension of the vector $\overline{L}_k$, we let

$$\gamma_k \equiv \frac{s_k}{d_k} \text{ and } r_k \equiv \frac{\gamma_k}{2\gamma_k + 1}.$$

In addition, recall that m_k is the dimension of the vector $\phi_k\left(\overline{L}_k\right)$ used to construct the series estimators $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ (see definition (2.42)).

Condition B(k) is the standard positivity assumption routinely invoked in causal inference.

Conditions R(k) and HrateL2(k) are used in the calculations on the bonds of the drifts of both $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ as we repeatedly invoke the Cauchy-Swartz inequality. Condition R(k) lists standard regularity conditions in the literature on series estimation, including conditions on the approximation properties of the chosen dictionary and conditions on the eigenvalues of covariance of the covariate vector, under which

$$\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,DR} - \eta_k\right)^2\middle|\mathcal{N}\right]} = O_p\left(n^{-\frac{\gamma_k}{2\gamma_k+1}}\right). \tag{2.52}$$

For completeness, in Lemma 17 of the Appendix B.5 we show a general result on the convergence of series estimators from where (2.52) follows under Condition R(k). Condition HrateInf(k) is additionally used in the derivation of our bounds for the drift of $\widehat{\theta}^u_{MR}$, as we invoke Hölder's inequality to bound the components $\xi^{MR}_{j,k}, j < k$, of the drift.

Conditions HrateInf(k) and HrateL2(k) are expressed in terms of generic estimators $\widehat{h}_k$ because in our calculations of the bounds we use the expression of the drift established in Theorem 2 which holds without any restrictions on the form of the estimators of h_k. Nevertheless, for completeness, in Subsection 2.7.2 below we provide regularity assumptions under which Conditions HrateInf(k) and HrateL2(k) holds if $\widehat{h}_k$ is a series estimator and h_k is assumed to belong to a Hölder ball with known smoothness order. In Appendix B.5 we provide further discussion of all the preceding conditions.

The following results are proved in Appendix B.6.

Theorem 3 *Assume that* $K = 2$ *and that Conditions Hölder*(k), $R(k)$, $B(k)$, *Hconvergence*(k) *and HrateL2*(k) *hold for* $k = 1, 2$. *If* $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ *for* $k = 1, 2$, *then*

$$\delta_1^{DR} = O_p\left(\beta_{1,n} n^{-r_1}\right), \ \delta_2^{DR} = O_p\left(\beta_{2,n} n^{-r_2}\right) \text{ and } \xi_{12}^{DR} = O_p\left(\beta_{1,n} n^{-r_2}\right)$$

Consequently,

$$E_p\left\{ Q_1\left(\widehat{h}, \widehat{\eta}\right) \middle| \mathcal{N} \right\} - \theta(\eta) = O_p\left(\mu_{DR,n}\right)$$

where

$$\mu_{DR,n} \equiv \max\left\{\beta_{1,n} n^{-r_1}, \beta_{2,n} n^{-r_2}, \beta_{1,n} n^{-r_2}\right\}$$

Theorem 4 *Assume that* $K = 2$, *that Conditions Hölder*(k), $R(k)$, $B(k)$ *and HrateL2*(k) *hold for* $k = 1, 2$, *that Condition Hconvergence*(k) *holds for* $k = 1$ *and that Condition HrateInf*(k) *holds for* $k = 2$. *If* $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ *for* $k = 1, 2$, *then*

$$\delta_1^{MR} = O_p\left(\beta_{1,n} n^{-r_1}\right), \ \delta_2^{MR} = O_p\left(\beta_{2,n} n^{-r_2}\right) \text{ and } \xi_{12}^{MR} = O_p\left(\beta_{1,n}\alpha_{2,n} n^{-r_2}\right).$$

Consequently,

$$E_p\left\{ Q_1\left(\widehat{h}, \widehat{\eta}_{MR}\right) \middle| \mathcal{N} \right\} - \theta(\eta) = O_p\left(\mu_{MR,n}\right)$$

where

$$\mu_{MR,n} \equiv \max\left\{\beta_{1,n} n^{-r_1}, \beta_{2,n} n^{-r_2}, \beta_{1,n}\alpha_{2,n} n^{-r_2}\right\}.$$

Suppose that in the two preceding theorems the smoothness s_1 and s_2 were the same. Then, because $d_1 < d_2$ (since $\overline{L}_2$ is a superset of L_1), we would have $\gamma_1 > \gamma_2$ and consequently $r_1 > r_2$. If, in addition, it were the case that $\beta_{1,n} = o\left(\beta_{2,n}\right)$, as would be the case if h_1 and h_2 also belonged to Hölder balls with the same, sufficiently large, smoothness order, and $\widehat{h}_1$ and $\widehat{h}_2$ were appropriate series estimators (see Subsection 2.7.2 below), then we would conclude that $\mu_{MR,n} = \mu_{DR,n} = \beta_{2,n} n^{-r_2}$. Thus, for such setting the preceding Theorems indicate that both $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ would be $\sqrt{n}$-consistent, i.e. $n^{1/2}\left\{\widehat{\theta}^u - \theta(\eta)\right\} = O_p(1)$ and $n^{1/2}\left\{\widehat{\theta}^u_{MR} - \theta(\eta)\right\} = O_p(1)$, so long as $\beta_{2,n} n^{1/2-r_2} = O(1)$. Then, for equally smooth functions η_1 and η_2, and equally (sufficiently) smooth functions h_1 and h_2, the preceding Theorems suggest that for the purposes of ensuring $\sqrt{n}-$consistency, there is no gain in using $\widehat{\theta}^u_{MR}$ instead of $\widehat{\theta}^u$, when both are computed using series estimators of the unknown h_k and the unknown η_k. However, if it happens to be the case that $\mu_{DR,n} = \beta_{1,n} n^{-r_2}$ because at the particular data generating law h_2 is so much smoother than h_1 that the estimation error of, say an appropriate series, estimator of h_2 converges at a rate $\beta_{2,n}$ much faster to 0 than the rate $\beta_{1,n}$ of the estimation error of, say the appropriate series, estimator of h_1 then we would have that $\mu_{MR,n} = o\left(\mu_{DR,n}\right)$. In such case, it could then happen that $n^{1/2}\left\{\widehat{\theta}^u_{MR} - \theta(\eta)\right\} = O_p(1)$ even though $n^{1/2}\left\{\widehat{\theta}^u - \theta(\eta)\right\}$ diverges. Thus, the preceding Theorems suggest that, as far as ensuring $\sqrt{n}$-consistency, it never hurts to use $\widehat{\theta}^u_{MR}$ instead of $\widehat{\theta}^u$ and on some exceptional circumstances, it may help. We emphasize that we use the verb "suggest" instead of "imply" because the rates of convergence $\mu_{DR,n}$ and $\mu_{MR,n}$ established in Theorems 3 and 4 are upper bounds on the rates of convergence of $E_p\left\{ Q_1\left(\widehat{h}, \widehat{\eta}_{DR}\right) \middle| \mathcal{N} \right\} - \theta(\eta)$ and $E_p\left\{ Q_1\left(\widehat{h}, \widehat{\eta}_{MR}\right) \middle| \mathcal{N} \right\} - \theta(\eta)$, which may not be sharp.

Theorem 5 *Assume that* $K = 3$ *and that Conditions Hölder*(k), $R(k)$, $B(k)$, *Hconvergence*(k) *and HrateL2*(k) *hold for* $k = 1, 2, 3$. *If* $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ *for* $k = 1, 2, 3$ *then*

$$\delta_k^{DR} = O_p\left(\beta_{k,n} n^{-r_k}\right), k = 1, 2, 3$$

and

$$\xi_{k,j}^{DR} = O_p\left(\beta_{k,n} n^{-r_j}\right), 1 \leq k < j \leq 3.$$

Consequently,

$$E_p\left\{ Q_1\left(\widehat{h}, \widehat{\eta}\right) \middle| \mathcal{N} \right\} - \theta(\eta) = O_p\left(\mu_{DR,n}\right)$$

where

$$\mu_{DR,n} \equiv \max\left\{\beta_{1,n} n^{-r_1}, \beta_{2,n} n^{-r_2}, \beta_{3,n} n^{-r_3}, \beta_{1,n} n^{-r_2}, \beta_{1,n} n^{-r_3}, \beta_{2,n} n^{-r_3}\right\}.$$

Theorem 6 *Assume that* $K = 3$, *that Conditions Hölder*(k), $R(k)$, $B(k)$, *and HrateL2*(k) *hold for* $k = 1, 2, 3$, *that Condition Hconvergence*(k) *holds for* $k = 1$ *and that Condition HrateInf*(k) *holds for* $k = 2, 3$. *If* $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ *for* $k = 1, 2, 3$ *then*

$$\delta_k^{MR} = O_p\left(\beta_{k,n} n^{-r_k}\right), k = 1, 2, 3,$$

$$\xi_{k,j}^{MR} = O_p\left(\beta_{k,n} \alpha_{j,n} n^{-r_j}\right), 1 \leq k < j \leq 3$$

and

$$\varkappa_{1,2,3}^{MR} = O_p\left(\beta_{1,n} \alpha_{2,n} \alpha_{3,n} n^{-r_3}\right).$$

Consequently,

$$E_p\left\{ Q_1\left(\widehat{h}, \widehat{\eta}_{MR}\right) \middle| \mathcal{N} \right\} - \theta(\eta) = O_p\left(\mu_{MR,n}\right)$$

where

$$\mu_{MR,n} \equiv \max\left\{\beta_{1,n} n^{-r_1}, \beta_{2,n} n^{-r_2}, \beta_{3,n} n^{-r_3}, \beta_{1,n} \alpha_{2,n} n^{-r_2}, \beta_{1,n} \alpha_{3,n} n^{-r_3}, \beta_{2,n} \alpha_{3,n} n^{-r_3}\right\},$$

An important first lesson from the results of Theorems 3 to 6, is that the structure of the bounds for the drifts for $K = 2$ and $K = 3$ shares a common pattern. Specifically

i The bounds $\mu_{DR,n}$ and $\mu_{MR,n}$ for $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ are each equal to the dominating rate in a distinct set of convergence rates.

ii Both sets include the products of the $L_2(P)$ convergence rates of the errors for estimating h_k and η_k for all k.

iii The set corresponding to $\widehat{\theta}^u$ additionally contains the products of the $L_2(P)$ convergence rates of the errors for estimating h_k and η_j for $k < j$.

iv In contrast, the set corresponding to $\widehat{\theta}^u_{MR}$ additionally contains the product of the $L_2(P)$ convergence rates of the errors for estimating h_k and η_j times the L_∞ convergence rate of the error for estimating h_j, for $k < j$.

v Thus, for each $k < j$, the product in the set corresponding to $\widehat{\theta}^u_{MR}$ is of smaller order than the corresponding product in the set associated with $\widehat{\theta}^u$.

vi Note also that the bound on the rate of convergence of $\varkappa_{1,2,3}^{MR}$ is irrelevant and does not appears in the definition of $\mu_{MR,n}$ because it is of smaller order than the bound on the rate of convergence of $\xi_{1,3}^{MR}$.

It can be shown that the features (i)-(v) generalizes to an arbitrary K. Furthermore, point (vi) also generalizes to arbitrary K in that the drift of $\widehat{\theta}_{MR}^{u}$ is a sum over more terms than that of the drift of $\widehat{\theta}^{u}$, but the additional terms in the drift of $\widehat{\theta}_{MR}^{u}$ are irrelevant because their rate of convergence to 0 never dominates the convergence rate of the drift of $\widehat{\theta}_{MR}^{u}$

We can then make analogous general qualitative remarks as those made for the case $K=2$. In general, when the smoothness orders s_k of η_k are the same for all k, and the smoothness of h_k are the same for all k and both h_k and η_k are estimated by appropriate series estimation, both $\widehat{\theta}^{u}$ and $\widehat{\theta}_{MR}^{u}$ will be $\sqrt{n}$-consistent so long as $\beta_{K,n}n^{1/2-r_K}=O(1)$. If it just happens that that $\mu_{DR,n}=\beta_{k,n}n^{-r_j}$ for some $k<j$, because at the particular data generating law h_k (η_j) are so much wigglier than the remaining $h_t's$ $(\eta_t's)$, then we would have that $\mu_{MR,n}=o(\mu_{DR,n})$ and it could then happen that $n^{1/2}\left\{\widehat{\theta}_{MR}^{u}-\theta(\eta)\right\}=O_p(1)$ even though $n^{1/2}\left\{\widehat{\theta}^{u}-\theta(\eta)\right\}$ diverges. Thus, as in the case $K=2$, Theorems 5 and 6 suggest that, as far as ensuring $\sqrt{n}$-consistency, it never hurts to use $\widehat{\theta}_{MR}^{u}$ instead of $\widehat{\theta}^{u}$ and on some exceptional circumstances, it may help.

Remark 1 *to arrive at the bounds for all the terms in the expression for the drift of* $\widehat{\theta}_{DR}^{u}$ *and* $\widehat{\theta}_{MR}^{u}$ *in the proofs of Theorems 3 to 6 it was crucial that the estimation of the* $\eta_k's$ *was separately conducted from independent samples* $\mathcal{N}^k$*. To see this, consider for example*

$$\begin{aligned}\left|\xi_{1,2}^{DR}\right| &\equiv \left|E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{\widehat{h}_1}\right)\Pi^1\left[\eta_{2,DR}-\eta_2\right]\middle|\mathcal{N}\right\}\right| \\ &\leq \sqrt{E_p\left\{\left(\frac{I_1}{h_1}-\frac{I_1}{\widehat{h}_1}\right)^2\middle|\mathcal{N}\right\}}\sqrt{E_p\left\{I_1\,\Pi^1\left[\eta_{2,DR}-\eta_2\right]^2\middle|\mathcal{N}\right\}}\end{aligned}$$

In $E_p\left\{I_1\,\Pi^1\left[\eta_{2,DR}-\eta_2\right]^2\middle|\mathcal{N}\right\}$, $\Pi^1\left[\eta_{2,DR}-\eta_2\right]$ *is the least squares projection of the data dependent outcomes* $\eta_{2,DR}-\eta_2$*. When* $\eta_{2,DR}-\eta_2$ *depends only on data in* $\mathcal{N}^2$ *and* Π^1 *is computed from data in* $\mathcal{N}^1$ *we can treat the outcome* $\eta_{2,DR}-\eta_2$ *as i.i.d. and thus apply results for the* L_2 *norm of least squares projections. See Remark 2 of Appendix B.6 for the specific details on this point.*

Sample splitting is additionally needed to handle the drift of $\widehat{\theta}_{MR}^{u}$*. To see this, recall that*

$$\begin{aligned}\eta_{k,MR}-\eta_k &= \eta_{k,DR}-\eta_k \\ &+\Pi^k\left[q_{k+1}\left(\cdot,\cdot;\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right)-E_p\left\{Q_{k+1}\left(\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1}=\cdot,\underline{\mathcal{N}}^{k+1}\right\}\right]\end{aligned}$$

Because $\widehat{\underline{h}}_{k+1}$ *and* $\widehat{\underline{\eta}}_{k+1,MR}$ *depend on* $\underline{\mathcal{N}}^{k+1}=\mathcal{N}^{k+1}\cup...\cup\mathcal{N}^K$ *which is independent from the sample* $\mathcal{N}^k$ *which is used to compute the projection* Π^k*, then conditional of* $\underline{\mathcal{N}}^{k+1}$*,* $Q_{k+1}\left(\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right)-E_p\left\{Q_{k+1}\left(\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}$ *is a mean zero random variable. Thus, since* Π^k *is applied to conditionally i.i.d. mean zero random variables it follows that* $\Pi^k\left[q_{k+1}\left(\cdot,\cdot;\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right)-E_p\left\{Q_{k+1}\left(\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1}=\cdot\right\}\right]^2$ *is of order* $O_p(m_k/n)$ *where, recall,* m_k *is the dimension of the covariate vector* $\phi_k\left(\overline{L}_k\right)$*. Without sample splitting the nuisance sample* $\mathcal{N}$ *we would have no way of controlling this term.*

The drifts when the $h_k's$ are also estimated by series estimation

In this subsection we consider the special case in which the $h_k's$ are also estimated by series estimation under the assumption that they belong to Hölder balls $\mathcal{H}\left(\overline{\mathcal{L}}_k;\nu_k,\sigma_k\right)$. To apply the results of the preceding subsection we will find the sequences $\alpha_{k,n}$ and $\beta_{k,n}$ such that the conditions HrateInf(k) and HrateL2(k) hold. We will assume that the series are computed using $M_k \asymp n^{\frac{d_k}{2\nu_k+d_k}}$ elements of an appropriate dictionary, so that, under regularity conditions stated in Lemma 9 below, condition HrateL2(k) holds for the optimal rate $\beta_{k,n} = n^{-\frac{\nu_k}{2\nu_k+d_k}}$. The key technical challenge we must overcome to finalize our calculation is to find what is the L_∞ rate of convergence $\alpha_{k,n}$ of the series estimator of h_k when the estimator is computed using the number of dictionary elements M_k which yields the optimal L_2 rate of convergence. Using results in the article of Belloni et al.([3]), we will derive rates $\alpha_{k,n}$ when $\lambda_k \equiv \nu_k/d_k > 1/6$ where $d_k = \dim\left(\overline{L}_k\right)$. We have not succeeded in finding results in the literature that will aid us in the computation of these rates when $\lambda_k \leq 1/6$. Even more, we were unable to find results in the literature that would even ensure consistency in L_∞, i.e. that $\left\|\widehat{h}_k - h_k\right\|_\infty = o_p(1)$ when $\lambda_k \leq 1/6$. In fact, we suspect that the rates we find for the case $1/6 < \lambda_k \leq 1/4$ are new.

Henceforth, assume that the estimator $\widehat{h}_k(\cdot)$ used by $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ is equal to

$$\widehat{h}_k(\cdot) \equiv \widehat{\tau}_k'\varphi_k(\cdot) \tag{2.53}$$

where

$$\widehat{\tau}_k \equiv \left[\sum_{i:O_i\in\mathcal{N}^k} \overline{I}_{k-1,i}\varphi_k\left(\overline{L}_{k,i}\right)\varphi_k\left(\overline{L}_{k,i}\right)'\right]^{-1}\left[\sum_{i:O_i\in\mathcal{N}^k} \overline{I}_{k-1,i}\varphi_k\left(\overline{L}_{k,i}\right)I_{k,i}\right]$$

and

$$\varphi_k\left(\overline{L}_k\right) \equiv \left(\varphi_{k,1}\left(\overline{L}_k\right),\ldots,\varphi_{k,M_k}\left(\overline{L}_k\right)\right)'$$

are the first M_k elements of an appropriate dictionary. Note that the range of the functions $\widehat{h}_k(\cdot)$ may not fall inside the interval $[0,1]$ even though the range of $h_k(\cdot)$ does. This is inconsequential for our calculations since they are based on results on the L_2 and L_∞ convergence rates of the estimation error $\widehat{h}_k - h_k$ that are valid without requiring that $\widehat{h}_k(\cdot)$ have the same range as $h_k(\cdot)$.

The following lemma follows immediately from Lemma 17 in Appendix B.5. We state it here for ease of reference.

Lemma 9 *Suppose that, for $k \in [K]$*

1. *$h_k(\cdot)$ lies in a Hölder ball $\mathcal{H}\left(\overline{\mathcal{L}}_k;\nu_k,\sigma_k\right)$ with $\sigma_k < \infty$ and known smoothness order $\nu_k > 0$, and*
2. *the assumptions 2 - 5 of Lemma 17 in Appendix B.5 are satisfied with $\overline{L}_k$ in the place of X, I_k in the place of Y, $h_k(\cdot)$ in the place of $g(\cdot)$, the distribution of $\left(I_k,\overline{L}_k\right)\left|\,\overline{A}_{k-1} = \overline{a}^*_{k-1}\right.$ in the place of the distribution of (Y,X'), $\varphi_k\left(\overline{l}_k\right)$ in the place of $p(x)$, $\mathcal{H}\left(\overline{\mathcal{L}}_k;\nu_k,\sigma_k\right)$ in the place of $\mathcal{G}$ and $\lambda_k = \frac{\nu_k}{d_k}$ in the place of γ.*

Then, for $k \in [K]$, if $M_k \asymp n^{\frac{1}{2\lambda_k+1}}$, it holds that

$$\sqrt{E_p\left\{\overline{I}_k\left(\widehat{h}_k - h_k\right)^2\middle|\mathcal{N}\right\}} = O_p\left(n^{-\frac{\lambda_k}{2\lambda_k+1}}\right).$$

In the next lemma, we establish bounds on the convergence rates of $\left\|\widehat{h}_k - h_k\right\|_\infty$ when $\lambda_k > \frac{1}{6}$. This lemma is an immediate corollary of Lemma 18 in Appendix B.5.

Lemma 10 *Suppose that, for each* $k \in [K]$,

1. $h_k(\cdot)$ *lies in a Hölder ball* $\mathcal{H}\left(\overline{\mathcal{L}}_k; v_k, \sigma_k\right)$ *with* $\sigma_k < \infty$ *and known smoothness order* v_k *such that* $\lambda_k = \frac{v_k}{d_k} > \frac{1}{6}$, *and*

2. *the assumptions 2 - 6 of Lemma 18 in Appendix B.5 hold with* $\overline{L}_k$ *in the place of* X, I_k *in the place of* Y, $h_k(\cdot)$ *in the place of* $g(\cdot)$, *the distribution of* $\left(I_k, \overline{L}_k\right) \big| \overline{A}_{k-1} = \overline{a}^*_{k-1}$ *in the place of the distribution of* (Y, X'), $\varphi_k\left(\overline{l}_k\right)$ *in the place of* $p(x)$, $\mathcal{H}\left(\overline{\mathcal{L}}_k; v_k, \sigma_k\right)$ *in the place of* $\mathcal{G}$ *and* $\lambda_k = \frac{v_k}{d_k}$ *in the place of* γ.

Then, for $k \in [K]$, *if* $M_k \asymp n^{\frac{1}{2\lambda_k+1}}$, *it holds that*

a) $\left\|\widehat{h}_k - h_k\right\|_\infty = O_p\left(n^{-\frac{6\lambda_k-1}{4\lambda_k+2}}\sqrt{\log n}\right)$, *if* $\frac{1}{6} < \lambda_k < \frac{1}{4}$, *and*

b) $\left\|\widehat{h}_k - h_k\right\|_\infty = O_p\left(n^{-\frac{\lambda_k}{2\lambda_k+1}}\log n\right)$, *if* $\lambda_k \geq \frac{1}{4}$.

In particular,

c) $\left\|\widehat{h}_k - h_k\right\|_\infty = o_p(1)$ *for all* $\lambda_k > \frac{1}{6}$.

Applying Lemmas 9 and 10, we conclude that under the assumptions of Theorem 3 and Lemmas 9 and 10 for $K = 2$, we have that

$$\mu_{DR,n} = O_p\left(\max_{1\leq k\leq j\leq 2} n^{-\left(\frac{\lambda_k}{2\lambda_k+1}+\frac{\gamma_j}{2\gamma_j+1}\right)}\right)$$

and if additionally, the conditions of Theorem 4 hold,

$$\mu_{MR,n} = \begin{cases} O_p\left(\max\left\{\max_{1\leq k\leq 2} n^{-\left(\frac{\lambda_k}{2\lambda_k+1}+\frac{\gamma_k}{2\gamma_k+1}\right)}, n^{-\left(\frac{\lambda_1}{2\lambda_1+1}+\frac{6\lambda_2-1}{4\lambda_2+2}+\frac{\gamma_2}{2\gamma_2+1}\right)}\sqrt{\log n}\right\}\right) & \text{if } \frac{1}{6} < \lambda_2 < \frac{1}{4} \\ O_p\left(\max\left\{\max_{1\leq k\leq 2} n^{-\left(\frac{\lambda_k}{2\lambda_k+1}+\frac{\gamma_k}{2\gamma_k+1}\right)}, n^{-\left(\frac{\lambda_1}{2\lambda_1+1}+\frac{\lambda_2}{2\lambda_2+1}+\frac{\gamma_2}{2\gamma_2+1}\right)}\log n\right\}\right) & \text{if } \lambda_2 \geq \frac{1}{4} \end{cases}$$

This result formalizes the discussion in the preceding section regarding the benefits offered by $\widehat{\theta}^u_{MR}$ relative to $\widehat{\theta}^u$. Specifically, if h_1 and h_2 are equally smooth and η_1 and η_2 are equally smooth, then $\lambda_1 > \lambda_2$ and $\gamma_1 > \gamma_2$. In such case, the preceding conditions for $\mu_{DR,n} = o\left(n^{-1/2}\right)$ and $\mu_{MR,n} = o\left(n^{-1/2}\right)$ agree and reduce to the same requirement that $\frac{\lambda_2}{2\lambda_2+1} + \frac{\gamma_2}{2\gamma_2+1} > 1/2$, so our results suggest that $\widehat{\theta}^u_{MR}$ does not offer gains relative to $\widehat{\theta}^u$ insofar ensuring $\sqrt{n}-$consistent estimation of $\theta(\eta)$. However, if it just happens to be the case that h_1 is so much less smooth than h_2 that $\lambda_1 < \lambda_2$, then it may happen that the dominating term in $\mu_{DR,n}$ is $n^{-\left(\frac{\lambda_1}{2\lambda_1+1}+\frac{\gamma_2}{2\gamma_2+1}\right)}$ whereas the dominating term in $\mu_{MR,n}$ is $n^{-\left(\frac{\lambda_1}{2\lambda_1+1}+\frac{6\lambda_2-1}{4\lambda_2+2}+\frac{\gamma_2}{2\gamma_2+1}\right)}\sqrt{\log n}$ if $\frac{1}{6} < \lambda_2 < \frac{1}{4}$ or $n^{-\left(\frac{\lambda_1}{2\lambda_1+1}+\frac{\lambda_2}{2\lambda_2+1}+\frac{\gamma_2}{2\gamma_2+1}\right)}\log n$ if $\lambda_2 \geq \frac{1}{4}$. In such case, $\mu_{MR,n} = o\left(\mu_{DR,n}\right)$, therefore implying the possibility that $\mu_{MR,n}$ is $o\left(n^{-1/2}\right)$

- and consequently that $\widehat{\theta}^u_{MR} = O_p\left(n^{-1/2}\right)$ – even though $\mu_{DR,n}$ is not even $O\left(n^{-1/2}\right)$ -and thus raising the possibility that $\widehat{\theta}^u$ is not $O_p\left(n^{-1/2}\right)$ – .

A similar analysis holds when $K = 3$. Specifically, if the assumptions of Theorem 5 and Lemmas 9 and 10 hold, then

$$\mu_{DR,n} = O_p\left(\max_{1\leq k\leq j\leq 3} n^{-\left(\frac{\lambda_k}{2\lambda_k+1}+\frac{\gamma_j}{2\gamma_j+1}\right)}\right).$$

Also, if the assumptions of Theorem 6 and Lemmas 9 and 10 hold, then

$$\mu_{MR,n} = O_p\left(\max\left[\max_{1\leq k\leq 3}\left\{n^{-\left(\frac{\lambda_k}{2\lambda_k+1}+\frac{\gamma_k}{2\gamma_k+1}\right)}\right\}, \max_{1\leq k<j\leq 3}\left\{n^{-\left(\frac{\lambda_k}{2\lambda_k+1}+\frac{\gamma_j}{2\gamma_j+1}+\varepsilon_j\right)}\log n\right\}\right]\right)$$

where, for $j = 2, 3$,

$$\varepsilon_j = \begin{cases} \frac{6\lambda_j-1}{4\lambda_j+2} & \text{if } \frac{1}{6} < \lambda_j < \frac{1}{4} \\ \frac{\lambda_j}{2\lambda_j+1} & \text{if } \lambda_j \geq \frac{1}{4} \end{cases}$$

From the expressions for $\mu_{DR,n}$ and $\mu_{MR,n}$ we can deduce similar qualitative conclusions. As far as ensuring $\sqrt{n}-$ consistency is concerned, it never hurts and on some exceptional occasions it may help to use $\widehat{\theta}^u_{MR}$ as opposed to $\widehat{\theta}^u$.

2.8 Analysis of the centered empirical process when h_k and η_k are estimated by series estimation.

In this section we argue that, under the assumptions of Theorems 3 and 4 (if $K=2$) or Theorems 5 and 6 (if $K=3$), the condition

$$E_p\left[\left\{Q\left(h^\dagger,\eta^\dagger\right)-Q\left(h,\eta\right)\right\}^2|\mathcal{N}\right]\underset{N_u\to\infty}{\overset{P}{\to}}0 \tag{2.54}$$

holds, when $\left(h^\dagger,\eta^\dagger\right)$ are replaced by the estimators $\left(\widehat{h},\widehat{\eta}\right)$ used to compute $\widehat{\theta}^u$ when the $\widehat{\eta}_k's$ are series estimators, or when $\left(h^\dagger,\eta^\dagger\right)$ are replaced by the estimators $\left(\widehat{h},\widehat{\eta}_{MR}\right)$ used to compute $\widehat{\theta}^u_{MR}$ when the $\widehat{\eta}_{k,MR}'s$ are series estimators. As indicated in Section 2.5, and proved in the Appendix B.3, when (2.54) holds, then the centered empirical process $\mathbb{G}_{N_u}\left\{Q\left(h^\dagger,\eta^\dagger\right)-Q\left(h,\eta\right)\right\}$ converges to 0 as $N_u\to\infty$.

This result and the expansion

$$\begin{aligned}\sqrt{N_u}\left\{\widehat{\theta}^{\dagger,u}-\theta\left(\eta\right)\right\}&=\mathbb{G}_{N_u}\left\{Q\left(h,\eta\right)\right\}+\mathbb{G}_{N_u}\left\{Q\left(h^\dagger,\eta^\dagger\right)-Q\left(h,\eta\right)\right\}\\&\quad+\sqrt{N_u}\left[E_p\left\{Q\left(h^\dagger,\eta^\dagger\right)\right\}-\theta\left(\eta\right)\right]\end{aligned}$$

imply that, under the conditions of the aforementioned theorems, and with $\mu_{DR,n}$ and $\mu_{MR,n}$ as defined in these theorems, the estimator $\widehat{\theta}^u$ ($\widehat{\theta}^u_{MR}$) will be $\sqrt{N_u}$-consistent (and thus $\sqrt{n}$-consistent) and asymptotically normal so long as $\mu_{DR,n}=o\left(n^{-1/2}\right)$ ($\mu_{MR,n}=o\left(n^{-1/2}\right)$). In what follows, $\widehat{h}$ stands for the vector whose components are the estimators $\widehat{h}_k$ used to compute $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$, which need not be series estimators. On the other hand, $\widehat{\eta}$ stands for the vector of *series estimators* $\widehat{\eta}_k$ used to compute $\widehat{\theta}^u$, and $\widehat{\eta}_{MR}$ stands for the vector of *series estimators* $\widehat{\eta}_{k,MR}$ used to compute $\widehat{\theta}^u_{MR}$.

In Appendix B.7 we prove the following result for the special cases $K=2$ and $K=3$..

Lemma 11 *Suppose that for each $k\in[K]$, the conditions Hölder(k), R(k), B(k) and Hconvergence(k) hold and $m_k\asymp n^{\frac{1}{2\gamma_k+1}}$. Then (2.54) holds with $\left(h^\dagger,\eta^\dagger\right)$ replaced by $\left(\widehat{h},\widehat{\eta}\right)$ and it also holds with $\left(h^\dagger,\eta^\dagger\right)$ replaced by $\left(\widehat{h},\widehat{\eta}_{MR}\right)$.*

2.9 Series estimation with the number of dictionary elements chosen by cross-validation

As indicated in Subsection 2.7.2, our comparisons of the asymptotic behavior of $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ when these use series estimators of the $\eta_k's$ heavily rely on Theorems 3 - 6. These theorems assume that the number m_k of dictionary functions is chosen so that $m_k \asymp n^{\frac{d_k}{2s_k+d_k}}$. Thus, our calculations assume that one knows the smoothness order s_k of the Hölder ball where η_k lies. However, in practice s_k is typically unknown, and m_k is selected using a data adaptive procedure, such as V-fold cross-validation. Unfortunately, when m_k is data driven, the results of Subsection 2.7.2 do not immediately apply for a couple of reasons. To explain them, we will focus attention to the case in which m_k is selected via V-fold cross-validation. So we first briefly review this procedure.

As in Subsection 2.5.1, suppose that one has a sample $\mathcal{D}$ comprised of n i.i.d. copies (Z_i, X_i) of (Z, X) where $\dim(X) = d$. Randomly split $\mathcal{D}$ into V equal or nearly equal size sub-samples $\mathcal{D}^1, ..., \mathcal{D}^V$. For a given dictionary $\{p_j(x)\}_{j\geq 1}$, and for each $r = 1, \ldots, V$, let $\widehat{g}^{r,(m)}(\cdot) \equiv \widehat{\beta}^{r,(m)\prime} p^{(m)}(\cdot)$ where $p^{(m)}(\cdot) \equiv (p_1(\cdot), \ldots, p_m(\cdot))'$ and

$$\widehat{\beta}^{r,(m)} \equiv \left[\sum_{i:(Z_i,X_i)\in\mathcal{D}-\mathcal{D}^r} p^{(m)}(X_i)\, p^{(m)}(X_i)'\right]^{-1} \left[\sum_{i:(Z_i,X_i)\in\mathcal{D}-\mathcal{D}^r} p^{(m)}(X_i)\, Z_i\right].$$

The V-fold cross validated number of dictionary elements over a set $\mathbb{M}$ is defined as

$$\widehat{m} \equiv \arg\min_{m\in\mathbb{M}} \sum_{v=1}^{V} \sum_{i:(Z_i,X_i)\in\mathcal{D}^r} \left[Z_i - \widehat{g}^{r,(m)}(X_i)\right]^2.$$

Finally, the cross validated estimator of $g(x)$ is given by $\widehat{g}^{(\widehat{m})}(\cdot) \equiv \widehat{\beta}^{(\widehat{m})\prime} p^{(\widehat{m})}(\cdot)$ where for any m

$$\widehat{\beta}^{(m)} \equiv \left[\sum_{i:(Z_i,X_i)\in\mathcal{D}} p^{(m)}(X_i)\, p^{(m)}(X_i)'\right]^{-1} \left[\sum_{i:(Z_i,X_i)\in\mathcal{D}} p^{(m)}(X_i)\, Z_i\right].$$

An important subtle point is that whereas, for a fixed m, $\widehat{g}^{(m)}(\cdot)$ depends linearly on the outcome vector $Z = (Z_1, ..., Z_n)$, the cross-validated estimator $\widehat{g}^{(\widehat{m})}(\cdot)$ does not, because of its non-linear dependence on Z through $\widehat{m}$.

The cross-validated $\widehat{m}$ attempts to approximate the ideal number of dictionary functions

$$\widetilde{m} \equiv \arg\min_{m\in\mathbb{M}} \int \left(z - \widehat{g}^{(m)}(x)\right)^2 dF(z,x)$$

that would select an oracle that could compute the true risk of each the $M = \#\mathbb{M}$ estimators $\widehat{g}^{(m)}(x)$. Dudoit and van der Laan, ([10], Section 3) showed that the V-fold cross-validated estimator $\widehat{g}^{(\widehat{m})}$ performs asymptotically as well as the oracle estimator $\widehat{g}^{(\widetilde{m})}$ in that if $\log(M) = o\left(nE_F\left[\int \left(z - \widehat{g}^{(\widetilde{m})}(x)\right)^2 dF(z,x)\right]\right)$ then

$$\frac{\int \left\{g(x) - \widehat{g}^{(\widehat{m})}(x)\right\}^2 dF(x)}{\int \left\{g(x) - \widehat{g}^{(\widetilde{m})}(x)\right\}^2 dF(x)} \xrightarrow[n\to\infty]{P} 1.$$

If g belongs to $\mathcal{H}(\mathcal{X}; s, \rho)$ and $\mathbb{M}$ is chosen large enough, one would expect $\widetilde{m} \asymp n^{\frac{d}{2s+d}}$, since as we have indicated in Subsection 2.7.2, the optimal number of dictionary functions grows as $n^{\frac{d}{2s+d}}$. Consequently, one would expect that $\int \left(g(x) - \widehat{g}^{(\widehat{m})}(x)\right)^2 dF(x) \asymp O_p\left(n^{-\left(\frac{s}{2s+d}\right)}\right)$.

The result of Dudoit and van der Laan suggest that if one uses V-fold cross-validation within each subsample $\mathcal{N}^k$ to compute the number $\widehat{m}_k$ of dictionary functions, the conclusions about how $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$ are compared should remain the same as when m_k is non-data driven and optimal. However, this assertion is tempered by two facts. First, our calculations in Subsection 2.7.2, heavily rely on the series estimator being a linear operator, i.e. depending linearly on the outcome vector, in order to use the alternative expressions for the drift derived in Subsection 2.7.1. However, as indicated earlier, the series estimator with number of dictionary functions selected by V-fold cross-validated does not depend linearly on the outcome vector. Second, $\widehat{m}_k$ would be computed using the pseudo-outcomes $\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)$ or $\widehat{Q}_{k+1}$ in place of the unknown true outcome $\eta_{k+1}\left(\overline{L}_{k+1}\right)$. The impact that these replacements have on the behavior of $\int \left\{\eta_{k,DR}\left(\bar{l}_k\right) - \eta_k\left(\bar{l}_k\right)\right\}^2 dF\left(\bar{l}_k | \overline{A}_k = \overline{a}^*_k\right)$ is unclear.

2.10 Resumen

El objetivo de este capítulo es investigar y contrastar las propiedades asintóticas de los estimadores doble y múltiple robustos del parámetro de la g-fórmula de cálculo longitudinal (también conocida como g-fórmula) de Robins ([29]), a partir de n replicaciones i.i.d. de un vector $O = (O_1, \ldots, O_K, L_{K+1})$ donde $O_k = (A_k, L_k), k = 1, ..., K, A_k$ es una variable discreta (que representa el tratamiento recibido en el momento t_k) y L_k es un vector aleatorio, posiblemente multivariado (que contiene a los datos registrados en el sujeto justo un instante previo a recibir el tratamiento A_k).

Sea p la densidad de la ley P de O, con respecto a alguna medida dominante. Definimos

$$p(o) = \prod_{k=0}^{K} g_k\left(l_{k+1}|\bar{l}_k, \bar{a}_k\right) \prod_{k=1}^{K} h_k\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right),$$

o abreviadamente $p = gh$, donde $h_k\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right) \equiv P\left(A_k = a_k|\overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = \bar{a}_{k-1}\right)$ y $g_k\left(l_{k+1}|\bar{l}_k, \bar{a}_k\right)$ es (una versión de) la densidad condicional de L_{k+1}. Aquí y en lo que sigue, para $1 \leq k \leq K$ y cualquier $\{v_j\}_{1 \leq j \leq K}$, denotamos $\overline{v}_k \equiv (v_1, \ldots, v_k)$.

La g-fórmula ([29]) se define como

$$\theta(p) \equiv E_{gh^*}\left\{\kappa\left(\overline{L}_{K+1}\right)\right\}$$

donde κ es una función a valores reales dada y, para cada $k = 1, ..., K, h_k^*\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right)$ es una función de probabilidad puntual dada, es decir conocida, tal que $p^* = gh^*$ es absolutamente continua con respecto a $p = gh$ y $E_{gh^*}(\cdot)$ denota esperanza bajo $p^* = gh^*$. Explícitamente,

$$\theta(p) = \int \varphi(o) \prod_{k=0}^{K} g_k\left(l_{k+1}|\bar{l}_k, \bar{a}_k\right) d\mu(o) \tag{2.55}$$

donde $\left(\bar{l}_0, \bar{a}_0\right) \equiv$nill y

$$\varphi(o) \equiv \left\{\prod_{k=1}^{K} h_k^*\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right)\right\} \kappa\left(\bar{l}_{K+1}\right). \tag{2.56}$$

es una función conocida, es decir, especificada, de o.

Bajo elecciones particulares de h_k^*, arribamos a expresiones de $\theta(p)$ equivalentes a parámetros de interés en la inferencia causal y en el análisis de datos faltantes. Estas elecciones son revisadas en el Apéndice B.1. Un ejemplo importante corresponde al caso en que

$$h_k^*\left(a_k|\bar{l}_k, \bar{a}_{k-1}\right) = I_{\left\{a_k^*\right\}}(a_k) \tag{2.57}$$

que, bajo el supuesto de que no existencia de confusores no medidos, conlleva a un $\theta(p)$ igual a la esperanza de la variable de respuesta contrafactual cuando se fuerza a toda la población a seguir un determinado régimen de tratamiento fijo, es decir, no dinámico, $A_k = a_k^*, k = 1, ..., K$ ([29], [32] and [33]). Para evitar tecnicismos y simplificar la notación, en este capítulo nos centraremos en este caso particular, es decir, asumiremos que h_k^* es la función de probabilidad puntual (2.57). Nuestros resultados se generalizan fácilmente a $h_k^{*\prime}s$ arbitrarias.

Para la función de probabilidad puntual h_k^* de (2.57) la g-fórmula se reduce a

$$\theta(p) \equiv E_{g_0}\left[E_{g_1}\left[\ldots E_{g_{K-1}}\left[E_{g_K}\left\{\kappa\left(\overline{L}_{K+1}\right)|\overline{A}_K=\overline{a}_K^*,\overline{L}_K\right\}\middle|\overline{A}_{K-1}=\overline{a}_{K-1}^*,\overline{L}_{K-1}\right]\ldots|A_1=a_1^*,L_1\right]\right]$$

donde $E_{g_k}(\cdot|\cdot)$ denota la esperanza condicional bajo g_k. Esta expresión deja en claro que $\theta(p)$ depende de p solo a través de g; más precisamente sólo a través de la ley marginal g_0 de L_1, a través de la esperanza condicional

$$\eta_K\left(\overline{L}_K\right) \equiv E_{g_K}\left\{\kappa\left(\overline{L}_{K+1}\right)\middle|\overline{A}_K=\overline{a}_K^*,\overline{L}_K\right\}$$

y a través de las esperanzas condicionales iteradas definidas secuencialmente para $k=K-1,\ldots,1$, como

$$\eta_k\left(\overline{L}_k\right) \equiv E_{g_k}\left\{\eta_{k+1}\left(\overline{L}_{k+1}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}. \tag{2.58}$$

Por lo tanto, de ahora en más denotaremos a $\theta(p)$ como $\theta(\eta)$ con $\eta \equiv (g_0,\eta_1,...,\eta_k)$.

Una primera elección natural para estimar $\theta(\eta)$ es mediante el estimador $\theta(\widehat{\eta})$ llamado "plug-in" donde $\widehat{\eta} \equiv (\widehat{g}_0,\widehat{\eta}_1,...,\widehat{\eta}_K)$ y

1. $\widehat{g}_0 = d\widehat{G}_0$ donde $\widehat{G}_0$ es la distribución empírica de L_1,
2. $\widehat{\eta}_K$ es un estimador preferido de la media condicional de $\kappa\left(\overline{L}_{K+1}\right)$ dado $\overline{L}_K$ entre los individuos con $\overline{A}_K=\overline{a}_K^*$, y,
3. secuencialmente para $k=K-1,\ldots,1$, $\widehat{\eta}_k\left(\overline{L}_k\right)$ es un estimador preferido de la media condicional de $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ dado $\overline{L}_k$ entre los individuos con $\overline{A}_k=\overline{a}_k^*$, obtenido simulando que la "variable de respuesta" desconocida $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ es igual a su estimador $\widehat{\eta}_{k+1}\left(\overline{L}_{k+1}\right)$.

Cuando los estimadores $\widehat{\eta}_k\left(\overline{L}_k\right), k=K+1,...,1$, son calculados bajo modelos de regresión paramétricos, al estimador "plug-in" se lo conoce como el estimador paramétrico de la g-fórmula ([37]). Bajo condiciones de regularidad, el estimador paramétrico de la g-fórmula $\theta(\widehat{\eta})$ satisface que $\sqrt{n}\{\theta(\widehat{\eta})-\theta(\eta)\}$ converge a una distribución normal con media cero si los modelos paramétricos de regresión asumidos para cada $\eta_k, k=K,...,1$ son todos correctos. Sin embargo, $\theta(\widehat{\eta})$ ni siquiera es consistente si uno de estos modelos está mal especificado. Una estrategia muy conocida que produce un estimador que confiere cierta protección contra la especificación incorrecta de los modelos para las $\eta_k's$ consiste en sumar a $\theta(\widehat{\eta})$ la cantidad $\mathbb{P}_n\left[M\left(\widehat{h},\widehat{\eta}\right)\right]$ donde $\widehat{h} \equiv \left(\widehat{h}_1,...,\widehat{h}_K\right)$ es un vector de estimadores preferidos de las $h_k's$ y para cualquier $h^\dagger \equiv \left(h_1^\dagger,...,h_K^\dagger\right)$ y $\eta^\dagger \equiv \left(g_0^\dagger,\eta_1^\dagger,...,\eta_K^\dagger\right)$,

$$\begin{aligned} M\left(h^\dagger,\eta^\dagger\right) &\equiv m\left(O;h^\dagger,\eta^\dagger\right) \\ &\equiv \sum_{k=1}^{K}\left\{\prod_{j=1}^{k}\frac{I_{\{a_j^*\}}(A_j)}{h_j^\dagger\left(A_j|\overline{A}_{j-1},\overline{L}_j\right)}\right\}\left\{\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right)-\eta_k^\dagger\left(\overline{L}_k\right)\right\} \end{aligned} \tag{2.59}$$

con $\eta_{K+1}^\dagger\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$ ([25], [1]). Notemos que $M\left(h^\dagger,\eta^\dagger\right)$ no depende de $g_0^\dagger$ y que cuando $h^\dagger=h$ y $\eta_k^\dagger=\eta_k$ para todo $k=1,...,K$, es decir cuando $\eta_k^\dagger$ es igual a la verdadera esperanza condicional iterada bajo $p=gh$, entonces $E_p\{M(h,\eta)\}=0$ donde, de ahora en más, $E_p(\cdot)$ denota esperanza bajo la ley $p=gh$.

Siguiendo esta estrategia conseguimos el estimador

$$\widehat{\theta} \equiv \theta\left(\widehat{\eta}\right) + \mathbb{P}_n\left[M\left(\widehat{h}, \widehat{\eta}\right)\right].$$

Es bien sabido que la variable aleatoria

$$IF\left(h, \eta\right) \equiv M\left(h, \eta\right) + \eta_1\left(L_1\right) - \theta\left(\eta\right)$$

es la única función de influencia del parámetro $\theta\left(\eta\right)$ bajo un modelo no paramétrico para la ley P de O ([41], [40]). Por lo tanto, dado que $\mathbb{P}_n\left[\widehat{\eta}_1\left(L_1\right)\right] = \theta\left(\widehat{\eta}\right)$, resulta que el estimador $\widehat{\theta}$ es igual al estimador semiparamétrico eficiente a un paso bajo un modelo no paramétrico, es decir

$$\widehat{\theta} \equiv \theta\left(\widehat{\eta}\right) + \mathbb{P}_n\left[IF\left(\widehat{h}, \widehat{\eta}\right)\right].$$

Otra identidad algebraica importante muestra que el estimador a un paso $\widehat{\theta}$ es, de hecho, el así llamado estimador "Augmented Inverse Probability Weighted" (AIPW), familiar en la literatura de datos faltantes e inferencia causal ([44]). Específicamente, mediante ciertos cálculos algebraicos, se puede ver que

$$\begin{aligned}\eta_1^{\dagger}\left(L_1\right) + M\left(h^{\dagger}, \eta^{\dagger}\right) = & \left\{\prod_{j=1}^{K} \frac{I_{\left\{a_j^*\right\}}\left(A_j\right)}{h_j^{\dagger}\left(A_j|\overline{A}_{j-1}, \overline{L}_j\right)}\right\} \kappa\left(\overline{L}_{K+1}\right) \\ & - \sum_{k=1}^{K}\left[\prod_{j=1}^{k-1} \frac{I_{\left\{a_j^*\right\}}\left(A_j\right)}{h_j^{\dagger}\left(A_j|\overline{A}_{j-1}, \overline{L}_j\right)}\left\{\left[\frac{I_{\left\{a_k^*\right\}}\left(A_k\right)}{h_k^{\dagger}\left(A_k|\overline{A}_{k-1}, \overline{L}_k\right)}\right] - 1\right\} \eta_k^{\dagger}\left(\overline{L}_k\right)\right]\end{aligned}$$

donde $\prod_{j=1}^{0}\left(\cdot\right) \equiv 1$. El estimador AIPW es precisamente la media muestral del lado derecho de la igualdad cuando, para cada k, $h_k^{\dagger}$ y $\eta_k^{\dagger}$ son reemplazados por estimadores $\widehat{h}_k, \widehat{\eta}_k$.

Bajo condiciones de regularidad, si las $\eta_k's$ y las $h_k's$ son estimadas bajo modelos paramétricos, lineales generalizados, de regresión (por ejemplo, las $h_k's$ son estimadas bajo modelos de regresión logística si las $A_k's$ son binarias), entonces $\sqrt{n}\left\{\widehat{\theta} - \theta\left(\eta\right)\right\}$ converge a distribución normal con media cero si o bien (i) los modelos de regresión paramétricos asumidos para cada $\eta_k, k = K, ..., 1$ son todos correctos, o bien (ii) los modelos paramétricos asumidos para cada h_k $k = K, ..., 1$ son todos correctos, pero no necesariamente (i) y (ii) se satisfacen simultáneamente. Esta propiedad, cuya demostración heurística se detalla en la Sección 2.4, es conocida como doble robustez, dado que $\widehat{\theta}$ otorga al analista dos oportunidades de obtener inferencias correctas sobre $\theta\left(\eta\right)$, una modelando las $\eta_k's$ correctamente y otra modelando las $h_k's$ correctamente.

El estimador a un paso resulta especialmente atractivo cuando las $\widehat{\eta}_k, k = 1, ..., K$, en los pasos 2 y 3 anteriores se estiman bajo un modelo no paramétrico definido únicamente por suposiciones de suavidad o raleza. En dicho caso, el estimador "plug-in" $\theta\left(\widehat{\eta}\right)$ puede no ser una opción útil de estimación ya que es posible que ni siquiera converja a tasa $\sqrt{n}$ ([31]). En cambio, si se utiliza una estrategia apropiada de división de la muestra, explicada en detalle en la Sección 2.5, el estimador a un paso que utiliza estimadores no paramétricos $\widehat{\eta}_k$ y $\widehat{h}_k$ es $\sqrt{n}-$ consistente para $\theta\left(\eta\right)$ y asintóticamente normal siempre que $\widehat{\eta}_k$ y $\widehat{h}_k$ converjan lo suficientemente rápido a η_k y h_k ([39], [8],

[11], [57]). Más aún, cuando $K = 1$, se ha demostrado que es posible obtener convergencia a tasa $\sqrt{n}$ en el estimador a un paso, compensando tasas de convergencia más lentas para el estimador de una de las funciones de ruido, η_1 o h_1, con tasas más rápidas para el estimador de la otra función de ruido ([39], [31], [8], [11], [57]). Por el contrario, poco se ha reportado en la literatura sobre las compensaciones específicas en las tasas de convergencia para la estimación de las funciones de ruido η y h otorgado por el estimador a un paso cuando $K > 1$, a excepción de [31]. Uno de los objetivos de este capítulo es estudiar qué compensaciones ofrece $\widehat{\theta}$, si es que ofrece alguna, en las tasas de convergencia de los estimadores de las funciones de ruido, cuando $K > 1$.

Recientemente, varios artículos han señalado ([53], [24]) que cuando las $\eta_k's$ son estimadas bajo modelos paramétricos, es posible obtener estimadores que otorgan incluso mayor protección contra la incorrecta especificación de los modelos que el estimador a un paso anterior. La siguiente modificación del paso 3 anterior produce uno de dichos estimadores.

3_MR. secuencialmente, para $k = K-1, \ldots, 1$, calcular $\widehat{\eta}_{k,MR}\left(\overline{L}_k\right)$ un estimador preferido de la media condicional de $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ dado $\overline{L}_k$ entre los individuos con $\overline{A}_k = \overline{a}_k^*$, obtenido simulando que la "variable de respuesta" desconocida $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ es igual la pseudo respuesta

$$\widehat{Q}_{k+1} \equiv \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right) + \frac{I_{\left\{a_{k+1}^*\right\}}\left(A_{k+1}\right)}{\widehat{h}_{k+1}\left(a_{k+1}^* | \overline{a}_k^*, \overline{L}_{k+1}\right)}\left\{\widehat{Q}_{k+2} - \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right)\right\}$$

donde $\widehat{Q}_{K+1} \equiv \kappa\left(\overline{L}_{K+1}\right)$.

Los artículos [46] y [47] han definido y defendido el uso de las pseudo respuestas $\widehat{Q}_{k+1}$ para producir estimadores dobles robustos. No fue hasta el artículo de [53] que se notó que el uso de estas pseudo respuestas produce estimadores que otorgan una mayor protección contra la especificación errónea de los modelos.

Notemos al estimador a un paso que utiliza $\widehat{\eta}_{k,MR}$ en vez de $\widehat{\eta}_k$ (como en el paso 3 anterior) con $\widehat{\theta}_{MR}$, es decir

$$\widehat{\theta}_{MR} \equiv \theta\left(\widehat{\eta}_{MR}\right) + \mathbb{P}_n\left[M\left(\widehat{h}, \widehat{\eta}_{MR}\right)\right]$$

donde $\widehat{\eta}_{MR} \equiv \left(\widehat{g}_0, \widehat{\eta}_{1,MR}, ..., \widehat{\eta}_{K-1,MR}, \widehat{\eta}_K\right)$.

Se puede demostrar que, cuando los modelos utilizados para calcular $\widehat{\eta}_{k,MR}$, y aquéllos utilizados para calcular $\widehat{h}_k$, $k = 1, ..., K$, son paramétricos, $\widehat{\theta}_{MR}$ de hecho coincide con el estimador del coeficiente asociado a $\overline{a}_K^*$ en el MMEM del Capítulo 1 de esta tesis, en el caso particular en que las covariables basales (denotadas en ese capítulo como Z) son nulas, el MMEM es saturado en $\overline{a}_K$ y las funciones $g_k\left(\overline{a}_K, \overline{L}_{j-1}; \gamma_k\right), k = 1, ..., K$ (utilizadas en la Subsección 1.7.1 para modelar las $\rho_k's$) son también saturadas en $\overline{a}_K$. Lo anterior es cierto si, además, (1) los modelos para las $\eta_k\left(\overline{L}_k\right)'s$ son aquéllos que se derivan de los modelos compatibles paramétricos para las $\eta_k\left(\overline{a}_K^*, \overline{L}_k\right)'s$ definidas en la Sección 7.1 de ese capítulo y (2) los parámetros que indexan los modelos para las $\eta_k's$ son calculados mediante regresiones pesadas como en ese capítulo. Por lo tanto, de los resultados del Capítulo 1 se deduce que, bajo condiciones de regularidad, $\sqrt{n}\left\{\widehat{\theta}_{MR} - \theta\left(\eta\right)\right\}$ converge a una distribución normal cuando, para cada $k = 1, ..., K$, o bien el modelo paramétrico para h_k utilizado para calcular $\widehat{h}_k$ es correcto o bien el modelo paramétrico para η_k utilizado para calcular $\widehat{\eta}_{k,MR}$ es correcto. A esta propiedad, cuya demostración heurística se detalla en la Sección 2.4.2 para modelos paramétricos arbitrarios y estimadores de los parámetros arbitrarios, no únicamente aquéllos

utilizados en el Capítulo 1, se la conoce como múltiple robustez ([53], [24]) y se la ha llamado doble robustez secuencial en ([22]). Esto implica que $\widehat{\theta}_{MR}$ otorga mayor protección contra la incorrecta especificación de los modelos que $\widehat{\theta}$ ya que garantiza inferencias válidas no sólo cuando todas las $\eta_k's$ son modeladas correctamente, o todas las $h_k's$ son modeladas correctamente, si no también cuando un subconjunto de las $\eta_k's$ son modeladas correctamente siempre que para los $k's$ para los cuales las $\eta_k's$ son modeladas incorrectamente, las $h_k's$ sean correctamente modeladas.

Si bien los ventajas en cuanto a robustez de $\widehat{\theta}_{MR}$ sobre $\widehat{\theta}$ parecen estar bien documentados y comprendidas en la literatura cuando las funciones de ruido h y η se estiman bajo modelos paramétricos, lo mismo no es cierto para el caso en el que h y η se estiman bajo modelos no paramétricos. Por lo tanto, un segundo objetivo de este capítulo es investigar, cuando h y η se estiman bajo modelos no paramétricos definidos únicamente por suposiciones de suavidad o raleza, si las $\widehat{\eta}'_{k,MR}s$ otorgan compensaciones adicionales en los requisitos sobre las tasas de convergencia de los estimadores de los parámetros de ruido sobre los que otorga el estimador a un paso que usa las $\widehat{\eta}_k$.

Para ser concretos sobre las contribuciones de este capítulo, para garantizar una estimación consistente a tasa $\sqrt{n}$, comenzamos escribiendo una descomposición de la diferencia centrada entre el estimador a un paso y el verdadero parámetro que se utiliza normalmente para analizar las propiedades asintóticas del estimador a un paso. De ahora en más, $h^{\dagger} \equiv \left(h_1^{\dagger}, ..., h_K^{\dagger}\right)$ y $\eta^{\dagger} \equiv \left(g_0^{\dagger}, \eta_1^{\dagger}, ..., \eta_K^{\dagger}\right)$ representan estimadores arbitrarios de h y η y

$$\widehat{\theta}^{\dagger} \equiv \theta\left(\eta^{\dagger}\right) + \mathbb{P}_n\left[M\left(h^{\dagger}, \eta^{\dagger}\right)\right].$$

Notemos que, cuando $g_0^{\dagger} = \widehat{g}_0$, $\theta\left(\eta^{\dagger}\right) = \mathbb{P}_n\left\{\eta_1^{\dagger}\left(L_1\right)\right\}$, de modo que

$$\begin{aligned}\widehat{\theta}^{\dagger} &\equiv \theta\left(\eta^{\dagger}\right) + \mathbb{P}_n\left[M\left(h^{\dagger}, \eta^{\dagger}\right)\right] \\ &= \mathbb{P}_n\left[\eta_1^{\dagger}\left(L_1\right) + M\left(h^{\dagger}, \eta^{\dagger}\right)\right] \\ &\equiv \mathbb{P}_n\left[Q\left(h^{\dagger}, \eta^{\dagger}\right)\right]\end{aligned}$$

donde

$$\begin{aligned}Q\left(h^{\dagger}, \eta^{\dagger}\right) &\equiv q\left(O; h^{\dagger}, \eta^{\dagger}\right) \\ &\equiv \eta_1^{\dagger}\left(L_1\right) + M\left(h^{\dagger}, \eta^{\dagger}\right)\end{aligned}$$

Notemos que $Q\left(h^{\dagger}, \eta^{\dagger}\right)$ no depende de $g_0^{\dagger}$.

Cuando $h^{\dagger} = \widehat{h}$ y $\eta^{\dagger} = \widehat{\eta}$, $\widehat{\theta}^{\dagger}$ coincide con el estimador doble robusto $\widehat{\theta}$ y cuando $h^{\dagger} = \widehat{h}$ y $\eta^{\dagger} = \widehat{\eta}_{MR}$, $\widehat{\theta}^{\dagger}$ es igual al estimador múltiple robusto $\widehat{\theta}_{MR}$. Escribamos

$$\begin{aligned}\sqrt{n}\left\{\widehat{\theta}^{\dagger} - \theta\left(\eta\right)\right\} &= \mathbb{G}_n\left\{Q\left(h, \eta\right)\right\} + \mathbb{G}_n\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\} \\ &\quad + \sqrt{n}\left[E_p\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right)\right\} - \theta\left(\eta\right)\right] \\ &= \Upsilon_{1,n} + \Upsilon_{2,n} + \Upsilon_{3,n}\end{aligned}$$

donde

$$E_p\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right)\right\} \equiv \int q\left(o; h^{\dagger}, \eta^{\dagger}\right) dP\left(o\right)$$

$$\mathbb{G}_n(\cdot) \equiv \sqrt{n}\mathbb{P}_n\{\cdot - E_p(\cdot)\}$$

es el proceso empírico centrado, $\Upsilon_{1,n} \equiv \mathbb{G}_n\{Q(h,\eta)\}$, $\Upsilon_{2,n} \equiv \mathbb{G}_n\left\{Q\left(h^\dagger,\eta^\dagger\right) - Q(h,\eta)\right\}$ y $\Upsilon_{3,n} \equiv \sqrt{n}\left\{E_p\left\{Q\left(h^\dagger,\eta^\dagger\right)\right\} - \theta(\eta)\right\}$.

Por el Teorema Central del Límite, el término $\Upsilon_{1,n}$ converge a una distribución normal con media cero siempre que $Var_{gh}[Q(h,\eta)] < \infty$.

El término $\Upsilon_{2,n}$ es la diferencia entre dos procesos empíricos centrados, uno evaluado en (h,η) y el otro evaluado en su estimador $\left(h^\dagger,\eta^\dagger\right)$. Si las funciones (h,η) son lo suficientemente suaves o ralas, entonces debería ser posible construir estimadores $\left(h^\dagger,\eta^\dagger\right)$ que converjan a (h,η) a una tasa lo suficientemente rápida, como para que $\Upsilon_{2,n}$ sea $o_p(1)$. Es posible hacer este término $o_p(1)$ bajo condiciones de regularidad muy moderadas, incluso sin restricciones de suavidad o raleza, empleando la siguiente estrategia conocida como "cross-fitting". En primer lugar, se divide a la muestra en un número finito $\mathbf{U}$ de submuestras de igual o casi igual tamaño, designando a una de ellas como la "muestra principal de estimación" y al resto como las "muestras de estimación de ruido". Luego, se calculan $\eta^\dagger$ y $h^\dagger$ utilizando los datos de la unión de las submuestras de estimación de ruido y se calcula el estimador a un paso a partir de la submuestra principal de estimación reemplazando las η y h desconocidas por sus estimadores calculados a partir de la unión de las submuestras de estimación de ruido. A continuación, se repite el procedimiento $\mathbf{U}-1$ veces, cada vez designando una submuestra distinta como la muestra principal de estimación. Finalmente, se calcula el estimador $\widehat{\theta}^\dagger$de $\theta(\eta)$ como el promedio de los $\mathbf{U}$ estimadores a un paso. La utilidad de la estrategia "cross-fitting" para evitar imponer condiciones en la complejidad del modelo se ha observado hace mucho tiempo ([49], Capítulo 25 de [58]) pero sólo recientemente ha sido enfatizado y recomendado (ver, por ejemplo, [45], [63] y [8]). En la Sección 2.5 se describen con precisión los pasos de este procedimiento.

Asumiendo que se ha utilizado un procedimiento "cross-fitting", entonces $\sqrt{n}\left\{\widehat{\theta}^\dagger - \theta(\eta)\right\}$ estará acotado en probabilidad siempre que

$$\Upsilon_{3,n} \equiv \sqrt{n}\left[E_p\left\{Q\left(h^\dagger,\eta^\dagger\right)\right\} - \theta(\eta)\right] \tag{2.60}$$

sea $O_p(1)$. Mas aún, $\sqrt{n}\left\{\widehat{\theta}^\dagger - \theta(\eta)\right\}$ será asintóticamente normal con media cero y varianza $Var_{gh}\{Q(h,\eta)\}$ si esta varianza es finita y $\Upsilon_{3,n} = o_p(1)$. Por lo tanto, el término (2.60) que usualmente se conoce como término del "drift" o término del sesgo ([58]), es crucial para determinar la distribución asintótica de $\widehat{\theta}^\dagger$.

La contribución central de este capítulo es la derivación de distintas expresiones para el término del "drift". A nuestro entender, ninguna de las expresiones que se derivarán en este capítulo se han informado anteriormente en la literatura. Cada una de estas expresiones ayuda a visualizar la estructura general de las propiedades de robustez, en términos de compensaciones de las tasas de convergencia de los estimadores no paramétricos de η y h, otorgadas por el estimador doble robusto $\widehat{\theta}$ y el estimador de múltiple robusto $\widehat{\theta}_{MR}$. Para el caso particular en el que los estimadores de cada η_k son lineales en la variable de respuesta, proporcionaremos una expresión adicional para el "drift" que nos permitirá investigar en detalle y comparar el comportamiento asintótico de $\widehat{\theta}$ y $\widehat{\theta}_{MR}$ cuando las η_k son estimadas mediante estimación por series.

Appendix A

Appendix of Chapter 1

A.1 Proof of the variation independence of Robins et al.'s parameterization functionals for the special case $K = 2$

The following proposition establishes that the functionals $f(v|z), f(l_2|l_1,a_1), \eta_0(\overline{a}_2,z), \rho_1(\overline{a}_2,l_1)$ and $\rho_2(\overline{a}_2,\overline{l}_2)$ are variation independent.

Proposition 2 *Let $f^*(v|z)$ and $f^*(l_2|l_1,a_1)$ be arbitrary conditional densities for V given Z, and L_2 given (L_1,A_1) respectively. Let $\eta_0^*(\overline{a}_2,z)$ be an arbitrary function of $(\overline{a}_2,z)$ and let $\rho_1^*(\overline{a}_2,l_1)$ and $\rho_2^*(\overline{a}_2,\overline{l}_2)$ be any functions satisfying $\rho_1^*(\overline{a}_2,z,v=v_0)=0$ and $\rho_2^*(\overline{a}_2,l_1,l_2=l_{2,0})=0$, for fixed given values v_0 and $l_{2,0}$. Then, there exists a distribution for (L_1,A_1,L_2,A_2,Y) with joint density f verifying that $f(v|z)=f^*(v|z)$, $f(l_2|l_1,a_1)=f^*(l_2|l_1,a_1)$, $\eta_0(\overline{a}_2,z)=\eta_0^*(\overline{a}_2,z)$, $\rho_1(\overline{a}_2,l_1)=\rho_1^*(\overline{a}_2,l_1)$ and $\rho_2(\overline{a}_2,\overline{l}_2)=\rho_2^*(\overline{a}_2,\overline{l}_2)$ where η_0,ρ_1 and ρ_2 are the functionals defined in Sections 1.4.2 and 1.5 applied to f.*

Proof. Define

$$\eta_2^*(\overline{a}_2,\overline{l}_2) \equiv \eta_0^*(\overline{a}_2,z) + \{\rho_1^*(\overline{a}_2,l_1) - \Gamma_1^*(\overline{a}_2,z)\} + \left\{\rho_2^*(\overline{a}_2,\overline{l}_2) - \Gamma_2^*(\overline{a}_2,l_1)\right\},$$

where $\Gamma_1^*(\overline{a}_2,z) \equiv E_{f^*(v|z)}\{\rho_1^*(\overline{a}_2,L_1)\,|Z=z\}$ and $\Gamma_2^*(\overline{a}_2,l_1) \equiv E_{f^*(l_2|l_1,a_1)}\left\{\rho_2^*(\overline{a}_2,\overline{L}_2)\,|A_1=a_1,L_1=l_1\right\}$.

Let f be the density corresponding to a distribution function of (L_1,A_1,L_2,A_2,Y) such that

$$f(V|Z)=f^*(V|Z), f(L_2|L_1,A_1)=f^*(L_2|L_1,A_1) \text{ and } E\left(Y|\overline{A}_2,\overline{L}_2\right)=\eta_2^*\left(\overline{A}_2,\overline{L}_2\right). \quad \text{(A.1)}$$

We will show that any f satisfying (A.1) also satisfies that $\eta_0(\overline{a}_2,z)=\eta_0^*(\overline{a}_2,z)$, $\rho_1(\overline{a}_2,l_1)=\rho_1^*(\overline{a}_2,l_1)$ and $\rho_2(\overline{a}_2,\overline{l}_2)=\rho_2^*(\overline{a}_2,\overline{l}_2)$.

The fact that $\rho_2=\rho_2^*$ follows because $\eta_2(\overline{A}_2,\overline{L}_2) \equiv E\left(Y|\overline{A}_2,\overline{L}_2\right)=\eta_2^*\left(\overline{A}_2,\overline{L}_2\right)$ and then,

$$\begin{aligned}
\rho_2(\overline{a}_2,\overline{l}_2) &\equiv \eta_2(\overline{a}_2,\overline{l}_2) - \eta_2(\overline{a}_2,l_1,l_2=l_{2,0}) \\
&= \eta_2^*(\overline{a}_2,\overline{l}_2) - \eta_2^*(\overline{a}_2,l_1,l_2=l_{2,0}) \\
&= \rho_2^*(\overline{a}_2,\overline{l}_2) - \rho_2^*(\overline{a}_2,l_1,l_2=l_{2,0}) \\
&= \rho_2^*(\overline{a}_2,\overline{l}_2).
\end{aligned}$$

Analogously, to see that $\rho_1 = \rho_1^*$ it is enough to show that $\eta_1 = \eta_1^*$ where

$$\eta_1^*(\overline{a}_2, l_1) \equiv \eta_0^*(\overline{a}_2, z) + \{\rho_1^*(\overline{a}_2, l_1) - \Gamma_1^*(\overline{a}_2, z)\}.$$

To prove that $\eta_1 = \eta_1^*$, first note that, by definition,

$$\eta_1(\overline{a}_2, l_1) = E_{f(l_2|l_1,a_1)}\left\{\eta_2\left(\overline{a}_2, \overline{L}_2\right)|A_1 = a_1, L_1 = l_1\right\}. \tag{A.2}$$

Now, by (A.1) and the fact that $\eta_2 = \eta_2^*$,

$$\eta_1(\overline{a}_2, l_1) = E_{f^*(l_2|l_1,a_1)}\left\{\eta_2^*\left(\overline{a}_2, \overline{L}_2\right)|A_1 = a_1, L_1 = l_1\right\},$$

and, by definition of η_1^* and η_2^*,

$$\begin{aligned} &E_{f^*(l_2|l_1,a_1)}\left\{\eta_2^*\left(\overline{a}_2, \overline{L}_2\right)|A_1 = a_1, L_1 = l_1\right\} = \\ &\eta_1^*(\overline{a}_2, l_1) + E_{f^*(l_2|l_1,a_1)}\left[\left\{\rho_2^*\left(\overline{a}_2, \overline{L}_2\right) - \Gamma_2^*(\overline{a}_2, L_1)\right\}|A_1 = a_1, L_1 = l_1\right]. \end{aligned} \tag{A.3}$$

The second term in the right hand side in (A.3) is zero by definition of Γ_2^*, which implies that $\eta_1 = \eta_1^*$. Consequently,

$$\begin{aligned} \rho_1(\overline{a}_2, l_1) &\equiv \eta_1(\overline{a}_2, l_1) - \eta_1(\overline{a}_2, z, v = v_0) \\ &= \eta_1^*(\overline{a}_2, l_1) - \eta_1^*(\overline{a}_2, z, v = v_0) \\ &= \rho_1^*(\overline{a}_2, l_1) - \rho_1^*(\overline{a}_2, z, v = v_0) \\ &= \rho_1^*(\overline{a}_2, l_1). \end{aligned}$$

The argument leading to the fact that $\eta_0 = \eta_0^*$ is analogous to the one used to prove that $\eta_1 = \eta_1^*$. Specifically, it holds because

(i) $\eta_0(\overline{a}_2, z) \equiv E_{f(v|z)}\{\eta_1(\overline{a}_2, L_1)|Z = z\}$,
(ii) $f(v|z) = f^*(v|z)$ and $\eta_1 = \eta_1^*$,
(iii) $E_{f^*(v|z)}\{\eta_1^*(\overline{a}_2, L_1)|Z = z\} = \eta_0^*(\overline{a}_2, z) + E_{f^*(v|z)}[\{\rho_1^*(\overline{a}_2, l_1) - \Gamma_1^*(\overline{a}_2, z)\}|Z = z]$, and
(v) $E_{f^*(v|z)}[\{\rho_1^*(\overline{a}_2, L_1) - \Gamma_1^*(\overline{a}_2, Z)\}|Z = z] = 0$. ■

A.2 Heuristics for fact (IV) of Subsection 1.6.4

Throughout this section, we use η_j^* and π_j^* to denote the trues η_j and π_j respectively, $j \in \{1,2\}$. Here, we give an intuitive argument, invoking counterfactuals, that the estimator $\widehat{\eta}_1$, constructed in step 4 of Subsection 1.6.4, is itself doubly robust in the sense that, under regularity conditions, it is consistent for η_1^* under the model that assumes that $\mathcal{R}_1$ holds and that either $\mathcal{R}_2$ or $\mathcal{P}_2$ holds.

Recall that, under the identifying assumptions, $\eta_1^*(\overline{a}_2, L_1) = E(Y_{\overline{a}_2}|A_1 = a_1, L_1)$. Thus, $\eta_1(A_1, a_2, L_1; \psi, \gamma_1, \tau_1)$ can be regarded as a MSMM for the conditional mean of the counterfactual outcome in a point exposure study with treatment variable A_2. Hence, we are effectively replicating our model formulation with $K = 1$, with (A_1, L_1) playing the role of Z and with L_2 playing the role of V. Invoking the theory of Molina et al. ([24]), we conclude that for any given $d_1\left(\overline{A}_2, L_1\right)$, the function

$$U_{d_1}^1\left\{(\psi, \gamma_1, \tau_1), \eta_2, \pi_2\right\} \equiv S_{d_1}^{1,1}(\eta_2, \pi_2) + S_{d_1}^{1,0}\left\{(\psi, \gamma_1, \tau_1), \eta_2\right\},$$

where

$$S_{d_1}^{1,1}(\eta_2, \pi_2) \equiv \frac{d_1\left(\overline{A}_2, L_1\right)}{\pi_2\left(\overline{A}_2, \overline{L}_2\right)}\left\{Y - \eta_2\left(\overline{A}_2, \overline{L}_2\right)\right\} \text{ and}$$

$$S_{d_1}^{1,0}\left\{(\psi, \gamma_1, \tau_1), \eta_2\right\} \equiv \sum_{a_2 \in \mathcal{A}_2} d_1(A_1, a_2, L_1)\left\{\eta_2\left(A_1, a_2, \overline{L}_2\right) - \eta_1(A_1, a_2, L_1; \psi, \gamma_1, \tau_1)\right\},$$

satisfies that, under model $\mathcal{R}_1$, $E\left[U_{d_1}^1\left\{(\psi^*, \gamma_1^*, \tau_1^*), \eta_2, \pi_2\right\}\right] = 0$ if either $\eta_2 = \eta_2^*$ or $\pi_2 = \pi_2^*$. Now consider estimators of (ψ^*, γ_1^*) solving

$$\mathbb{P}_n\left\{S_{\widehat{d}_1}^{1,1}(\widehat{\eta}_2, \widehat{\pi}_2)\right\} + \mathbb{P}_n\left[S_{\widehat{d}_1}^{1,0}\left\{(\psi, \gamma_1, \widehat{\tau}_1), \widehat{\eta}_2\right\}\right] = 0 \tag{A.4}$$

where $\widehat{d}_1$ is a, possibly data dependent, function, and $\widehat{\tau}_1, \widehat{\eta}_2$ and $\widehat{\pi}_2$ are consistent for τ_1^*, η_2^* and π_2^*, under $\mathcal{R}_1$, $\mathcal{R}_2$ and $\mathcal{P}_2$ respectively. Under regularity conditions, such estimators of (ψ^*, γ_1^*) are CAN under the model that assumes that $\mathcal{R}_1$ holds and that either $\mathcal{R}_2$ or $\mathcal{P}_2$ holds. The estimator $\left(\widehat{\psi}^{(1)}, \widehat{\gamma}_1^{(1)}\right)$ computed in step 4 of the algorithm of Subsection 1.6.4 solves an equation of the form (A.4) with $\widehat{d}_1(\cdot,\cdot) = \widehat{\pi}_1^{-1}\overset{\bullet}{\eta}_1(\cdot,\cdot;\widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)})$ and with $\widehat{\pi}_2, \widehat{\tau}_1$ and $\widehat{\eta}_2$ computed in steps 1, 2 and 3 of that estimation algorithm and, hence, consistent for π_2^*, τ_1^* and η_2^* under $\mathcal{P}_2, \mathcal{R}_1$ and $\mathcal{R}_2$ respectively. To see this first note that, by step 4, $\left(\widehat{\psi}^{(1)}, \widehat{\gamma}_1^{(1)}\right)$ solves $\mathbb{P}_n\left[S_{\widehat{d}_1}^{1,0}\left\{(\psi, \gamma_1, \widehat{\tau}_1), \widehat{\eta}_2\right\}\right] = 0$. In addition, by step 3 and the fact that $\overset{\bullet}{\eta}_1$ is a subvector of $\overset{\bullet}{\eta}_2$, $\mathbb{P}_n\left\{S_{\widehat{d}_1}^{1,1}(\widehat{\eta}_2, \widehat{\pi}_2)\right\} = 0$. Note that if, in the equation of step 4, $\overset{\bullet}{\eta}_1$ would have been evaluated at (ψ, γ_1) instead of $\left(\widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)}\right)$, then it had not been true that $\left(\widehat{\psi}^{(1)}, \widehat{\gamma}_1^{(1)}\right)$ solves an equation of the form (A.4), unless m and g_1 are linear in the parameters. This explains why evaluation of $\overset{\bullet}{\eta}_1$ at $\left(\widehat{\psi}^{(2)}, \widehat{\gamma}_1^{(2)}\right)$ in step 4 is essential to ensure the double robustness of $\left(\widehat{\psi}^{(1)}, \widehat{\gamma}_1^{(1)}\right)$ and hence, the multiple robustness of $\widehat{\psi}_{MR}$.

A.3 Proofs of results of Section 1.10

A.3.1 Proof of Lemma 1

First, note that the equations solved in steps 1 and 2 of the estimation algorithm of Section 1.7.2 are $\mathbb{P}_n\left(\phi^k_{\theta_k}\right)=0$ and $\mathbb{P}_n\left(\phi^{K+k}_{\theta_{K+k}}\right)=0$ respectively, $k=1,\ldots,K$. Hence, $\widehat{\theta}_k=\widehat{\alpha}_k$ solves

$$\mathbb{P}_n\left(\phi^k_{\theta_k}\right)=0$$

and $\widehat{\theta}_{K+k}=\widehat{\tau}_k$ solves

$$\mathbb{P}_n\left(\phi^{K+k}_{\theta_{K+k}}\right)=0,$$

$k=1,\ldots,K$.

Then, we would arrive at the desired result if we show that

$$\widehat{\overline{\theta}}_{3K+1-k} \text{ solves } \mathbb{P}_n\left(\phi^{3K+1-k}_{\overline{\theta}_{3K+1-k}}\right)=0, \tag{A.5}$$

for $k=0,\ldots,K$.

To see (A.5) for $k=K$, note that the equation in step 3 of the estimation algorithm is the equation in θ_{2K+1} given by

$$\mathbb{P}_n\left(\phi^{2K+1}_{\widehat{\overline{\theta}}_{2K},\theta_{2K+1}}\right)=0.$$

By assumption of the lemma, this equation has a solution. Then, by definition of $\left(\widehat{\psi}^{(K)},\widehat{\overline{\gamma}}^{(K)}_K\right)$ in step 3 of the estimation algorithm, $\widehat{\theta}_{2K+1}=\left(\widehat{\psi}^{(K)},\widehat{\overline{\gamma}}^{(K)}_K\right)$ is a solution of $\mathbb{P}_n\left(\phi^{2K+1}_{\widehat{\overline{\theta}}_{2K},\theta_{2K+1}}\right)=0$ and, hence, $\widehat{\overline{\theta}}_{2K+1}$ solves $\mathbb{P}_n\left(\phi^{2K+1}_{\overline{\theta}_{2K+1}}\right)=0$.

To see (A.5) for $k=1,\ldots,K-1$, first note that the equations in step 4 of the estimation algorithm are the equations in θ_{3K+1-k} given by

$$\mathbb{P}_n\left(\varphi^{k,k}_{\widehat{\overline{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}}\right)=0,$$

for $k=1,\ldots,K-1$. By assumption of the lemma, these equations have a solution. Then, by definition of $\left(\widehat{\psi}^{(k)},\widehat{\overline{\gamma}}^{(k)}_k\right)$ in step 4 of the estimation algorithm, $\widehat{\theta}_{3K+1-k}=\left(\widehat{\psi}^{(k)},\widehat{\overline{\gamma}}^{(k)}_k\right)$ is a solution of $\mathbb{P}_n\left(\varphi^{k,k}_{\widehat{\overline{\theta}}_{3K-k},\theta_{3K+1-k}}\right)=0$, so that

$$\widehat{\overline{\theta}}_{3K+1-k} \text{ solves } \mathbb{P}_n\left(\varphi^{k,k}_{\overline{\theta}_{3K+1-k}}\right)=0 \text{ for } k=1,\ldots,K-1. \tag{A.6}$$

Now, the fact that (A.5) holds for $k=1,\ldots,K-1$, follows from the facts that

(i) $\phi^{3K+1-k}_{\overline{\theta}_{3K+1-k}}=\sum_{j=k}^{K}\varphi^{k,j}_{\overline{\theta}_{3K+1-j}}$,

(ii) $\mathbb{P}_n\left(\varphi^{k,k}_{\widehat{\overline{\theta}}_{3K+1-k}}\right)=0$ by (A.6),

(iii) for $j = k+1, \ldots, K-1$, $\mathbb{P}_n\left(\varphi^{k,j}_{\widehat{\widehat{\theta}}_{3K+1-j}}\right) = 0$ because $\mathbb{P}_n\left(\varphi^{j,j}_{\widehat{\widehat{\theta}}_{3K+1-j}}\right) = 0$ by (A.6) and $\overset{\bullet}{\eta}_k$ is a subvector of $\overset{\bullet}{\eta}_j$, and

(iv) $\mathbb{P}_n\left(\varphi^{k,K}_{\widehat{\widehat{\theta}}_{2K+1}}\right) = 0$ because $\mathbb{P}_n\left(\phi^{2K+1}_{\widehat{\widehat{\theta}}_{2K+1}}\right) = 0$, by (A.5) for $k = K$, and $\overset{\bullet}{\eta}_k$ is a subvector of $\overset{\bullet}{\eta}_K$.

Finally, to see (A.5) for $k = 0$, first note that the equation in step 5 of the estimation algorithm is the equation in θ_{3K+1},

$$\mathbb{P}_n\left(\varphi^{0,0}_{\widehat{\widehat{\theta}}_{3K},\theta_{3K+1}}\right) = 0.$$

By assumption of the lemma, this equation has a solution. Then, by definition of $\widehat{\psi}_{MR}$ in step 5 of the estimation algorithm, $\widehat{\theta}_{3K+1} = \widehat{\psi}_{MR}$ is a solution of $\mathbb{P}_n\left(\varphi^{0,0}_{\widehat{\widehat{\theta}}_{3K},\theta_{3K+1}}\right) = 0$, so that

$$\widehat{\widehat{\theta}}_{3K+1} \text{ solves } \mathbb{P}_n\left(\varphi^{0,0}_{\theta_{3K+1}}\right) = 0. \tag{A.7}$$

Hence, the fact that (A.5) holds for $k = 0$ is a consequence of the following facts:

(i) $\phi^{3K+1}_{\theta_{3K+1}} = \sum_{j=0}^{K} \varphi^{0,j}_{\theta_{3K+1-j}}$,

(ii) $\mathbb{P}_n\left(\varphi^{0,0}_{\widehat{\widehat{\theta}}_{3K+1}}\right) = 0$ by (A.7),

(iii) for $j = 1, \ldots, K-1$, $\mathbb{P}_n\left(\varphi^{0,j}_{\widehat{\widehat{\theta}}_{3K+1-j}}\right) = 0$ because $\mathbb{P}_n\left(\varphi^{j,j}_{\widehat{\widehat{\theta}}_{3K+1-j}}\right) = 0$ by (A.6) and $\overset{\bullet}{m}$ is a subvector of $\overset{\bullet}{\eta}_j$, and

(iv) $\mathbb{P}_n\left(\varphi^{0,K}_{\widehat{\widehat{\theta}}_{2K+1}}\right) = 0$ because $\mathbb{P}_n\left(\phi^{2K+1}_{\widehat{\widehat{\theta}}_{2K+1}}\right) = 0$ and $\overset{\bullet}{m}$ is a subvector of $\overset{\bullet}{\eta}_K$.

A.3.2 Proof of Proposition 1

We start with the proof of fact (a). First, recall that,

$$U_d(\psi, \widetilde{\eta}, \widetilde{\pi}) = S_d^K\left(\widetilde{\eta}_K, \widetilde{\overline{\pi}}_K\right) + \sum_{k=1}^{K-1} S_d^k\left(\widetilde{\eta}_k, \widetilde{\eta}_{k+1}, \widetilde{\overline{\pi}}_k\right) + S_d^0(\psi, \widetilde{\eta}_1),$$

where

$$S_d^K\left(\widetilde{\eta}_K, \widetilde{\overline{\pi}}_K\right) = \frac{d\left(\overline{A}_K, Z\right)}{\widetilde{\pi}_K^p\left(\overline{A}_K, \overline{L}_K\right)}\left\{Y - \widetilde{\eta}_K\left(\overline{A}_K, \overline{L}_K\right)\right\},$$

for $k = 1, \ldots, K-1$,

$$S_d^k\left(\widetilde{\eta}_k, \widetilde{\eta}_{k+1}, \widetilde{\overline{\pi}}_k\right) = \sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{d(\overline{A}_k, \underline{a}_{k+1}, Z)}{\widetilde{\pi}_k^p\left(\overline{A}_k, \overline{L}_k\right)}\left\{\widetilde{\eta}_{k+1}\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1}\right) - \widetilde{\eta}_k\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k\right)\right\},$$

and

$$S_d^0(\psi, \widetilde{\eta}_1) = \sum_{\underline{a}_1 \in \underline{\mathcal{A}}_1} d(\underline{a}_1, Z)\left\{\widetilde{\eta}_1(\underline{a}_1, L_1) - m(\underline{a}_1, Z; \psi)\right\}.$$

Here, for arbitrary functions $\widetilde{\pi}_j, j = 1, \ldots, K$, with range in $(0,1)$ and domain in the sample space of $\left(\overline{A}_j, \overline{L}_j\right)$, for $k = 1, \ldots, K$, we denote

$$\widetilde{\pi}_k^p\left(\overline{a}_k, \overline{l}_k\right) \equiv \prod_{j=1}^{k} \widetilde{\pi}_j\left(\overline{a}_j, \overline{l}_j\right).$$

Now, define the following functions. Let

$$M_d^{K+1,K}\left(\psi, \widetilde{\overline{\pi}}_K\right) \equiv \frac{d\left(\overline{A}_K, Z\right)}{\widetilde{\pi}_K^p\left(\overline{A}_K, \overline{L}_K\right)}\left\{Y - m\left(\overline{A}_K, Z; \psi\right)\right\},$$

$$M_d^{K,K}\left(\psi, \widetilde{\eta}_K, \widetilde{\overline{\pi}}_K\right) \equiv \frac{d\left(\overline{A}_K, Z\right)}{\widetilde{\pi}_K^p\left(\overline{A}_K, \overline{L}_K\right)}\left\{\widetilde{\eta}_K\left(\overline{A}_K, \overline{L}_K\right) - m\left(\overline{A}_K, Z; \psi\right)\right\},$$

and, for $k = 1, \ldots, K-1$, let

$$M_d^{k+1,k}\left(\psi, \widetilde{\eta}_{k+1}, \widetilde{\overline{\pi}}_k\right) \equiv \sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{d\left(\overline{A}_k, \underline{a}_{k+1}, Z\right)}{\widetilde{\pi}_k^p\left(\overline{A}_k, \overline{L}_k\right)}\left\{\widetilde{\eta}_{k+1}\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_{k+1}\right) - m\left(\overline{A}_k, \underline{a}_{k+1}, Z; \psi\right)\right\}$$

and

$$M_d^{k,k}\left(\psi, \widetilde{\eta}_k, \widetilde{\overline{\pi}}_k\right) \equiv \sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{d\left(\overline{A}_k, \underline{a}_{k+1}, Z\right)}{\widetilde{\pi}_k^p\left(\overline{A}_k, \overline{L}_k\right)}\left\{\widetilde{\eta}_k\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k\right) - m\left(\overline{A}_k, \underline{a}_{k+1}, Z; \psi\right)\right\}.$$

Finally, let

$$M_d^{1,0}\left(\psi, \widetilde{\eta}_1\right) \equiv \sum_{\underline{a}_1 \in \underline{\mathcal{A}}_1} d\left(\underline{a}_1, Z\right)\left\{\widetilde{\eta}_1\left(\underline{a}_1, L_1\right) - m\left(\underline{a}_1, Z; \psi\right)\right\}.$$

Since

$$M_d^{K+1,K}\left(\psi, \widetilde{\overline{\pi}}_K\right) - M_d^{K,K}\left(\psi, \widetilde{\eta}_K, \widetilde{\overline{\pi}}_K\right) = S_d^K\left(\widetilde{\eta}_K, \widetilde{\overline{\pi}}_K\right),$$

for $k = 1, \ldots, K-1$,

$$M_d^{k+1,k}\left(\psi, \widetilde{\eta}_{k+1}, \widetilde{\overline{\pi}}_k\right) - M_d^{k,k}\left(\psi, \widetilde{\eta}_k, \widetilde{\overline{\pi}}_k\right) = S_d^k\left(\widetilde{\eta}_k, \widetilde{\eta}_{k+1}, \widetilde{\overline{\pi}}_k\right)$$

and

$$M_d^{1,0}\left(\psi, \widetilde{\eta}_1\right) = S_d^0\left(\psi, \widetilde{\eta}_1\right),$$

then we can write

$$\begin{aligned} U_d\left(\psi, \widetilde{\eta}, \widetilde{\pi}\right) = {} & M_d^{K+1,K}\left(\psi, \widetilde{\overline{\pi}}_K\right) - M_d^{K,K}\left(\psi, \widetilde{\eta}_K, \widetilde{\overline{\pi}}_K\right) \\ & + \sum_{k=1}^{K-1}\left\{M_d^{k+1,k}\left(\psi, \widetilde{\eta}_{k+1}, \widetilde{\overline{\pi}}_k\right) - M_d^{k,k}\left(\psi, \widetilde{\eta}_k, \widetilde{\overline{\pi}}_k\right)\right\} \\ & + M_d^{1,0}\left(\psi, \widetilde{\eta}_1\right). \end{aligned}$$

Rearranging the terms in the last display, we arrive at

$$U_d(\psi,\widetilde{\eta},\widetilde{\pi}) = M_d^{K+1,K}\left(\psi,\widetilde{\overline{\pi}}_K\right) - \sum_{k=2}^{K}\left\{M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right) - M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)\right\}$$
$$- \left\{M_d^{1,1}(\psi,\widetilde{\eta}_1,\widetilde{\pi}_1) - M_d^{1,0}(\psi,\widetilde{\eta}_1)\right\}$$

Our proof of (a) relies on the following facts:

(1) For $k = 2,\ldots,K$,

$$E_P\left\{M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right) - M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)\right\} = 0$$

for every $P,\psi,\widetilde{\eta}_k$ and $\widetilde{\overline{\pi}}_{k-1}$ if $\widetilde{\pi}_k = \pi_k$, and

$$E_P\left\{M_d^{1,1}(\psi,\widetilde{\eta}_1,\widetilde{\pi}_1) - M_d^{1,0}(\psi,\widetilde{\eta}_1)\right\} = 0$$

for every P,ψ and $\widetilde{\eta}_1$ if $\widetilde{\pi}_1 = \pi_1$.

(2.a) For $k = 1,\ldots,K-1$,

$$E_P\left\{M_d^{k+1,k}\left(\psi,\widetilde{\eta}_{k+1},\widetilde{\overline{\pi}}_k\right) - M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)\right\} = 0$$

for every P,ψ and $\widetilde{\overline{\pi}}_k$ if $\widetilde{\eta}_{k+1} = \eta_{k+1}$ and $\widetilde{\eta}_k = \eta_k$,

(2.b) $E_P\left\{M_d^{K+1,K}\left(\psi,\widetilde{\overline{\pi}}_K\right) - M_d^{K,K}\left(\psi,\widetilde{\eta}_K,\widetilde{\overline{\pi}}_K\right)\right\} = 0$ for every P,ψ and $\widetilde{\overline{\pi}}_K$ if $\widetilde{\eta}_K = \eta_K$, and

(2.c) $E_P\left\{M_d^{1,0}(\psi(P),\widetilde{\eta}_1)\right\} = 0$ for every $P \in \mathcal{M}$ if $\widetilde{\eta}_1 = \eta_1$.

Fact (1) follows from the fact that, for $k = 2,\ldots,K$,

$$E_P\left\{M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)|\overline{A}_{k-1},\overline{L}_k\right\} = M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)$$

for every $P,\psi,\widetilde{\eta}_k$ and $\widetilde{\overline{\pi}}_{k-1}$ if $\widetilde{\pi}_k = \pi_k$ and

$$E_P\left\{M_d^{1,1}(\psi,\widetilde{\eta}_1,\widetilde{\pi}_1)|L_1\right\} = M_d^{1,0}(\psi,\widetilde{\eta}_1)$$

for every P,ψ and $\widetilde{\eta}_1$ if $\widetilde{\pi}_1 = \pi_1$. This is because, for any random random vector (W,X) with W discrete and for any function $g(w,x)$, $E\left\{\frac{g(W,X)}{\Pr(W|X)}|X\right\} = \sum_w g(w,X)$.

Fact (2.a) follows from the fact that, for $k = 1,\ldots,K-1$,

$$E_P\left\{M_d^{k+1,k}\left(\psi,\eta_{k+1},\widetilde{\overline{\pi}}_k\right)|\overline{A}_k,\overline{L}_k\right\} = M_d^{k,k}\left(\psi,\eta_k,\widetilde{\overline{\pi}}_k\right)$$

for every P,ψ and $\widetilde{\overline{\pi}}_k$. This is because

$$E_P\left\{\eta_{k+1}\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_{k+1}\right)|\overline{A}_k,\overline{L}_k\right\} = \eta_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k\right)$$

for every $\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}$.

Likewise, fact (2.b) follows from the fact that

$$E_P\left\{M_d^{K+1,K}\left(\psi,\widetilde{\overline{\pi}}_K\right)|\overline{A}_K,\overline{L}_K\right\}=M_d^{K,K}\left(\psi,\eta_K,\widetilde{\overline{\pi}}_K\right)$$

for every P,ψ and $\widetilde{\overline{\pi}}_K$, because $E_P\left\{Y|\overline{A}_K,\overline{L}_K\right\}=\eta_K\left(\overline{A}_K,\overline{L}_K\right)$.

Fact (2.c) follows from the facts that

$$E_P\left\{\eta_1\left(\underline{a}_1,L_1\right)|Z\right\}=\eta_0\left(\underline{a}_1,Z\right)$$

for every $\underline{a}_1\in\underline{\mathcal{A}}_1$ by definition of η_0, and $\eta_0\left(\underline{a}_1,Z\right)=m\left(\underline{a}_1,Z;\psi\left(P\right)\right)$ for every $P\in\mathcal{M}$.

Now, let $P\in\mathcal{M}$, d any function and let $\widetilde{\eta}=\left(\widetilde{\eta}_1,\ldots,\widetilde{\eta}_K\right)$ and $\widetilde{\pi}=\left(\widetilde{\pi}_1,\ldots,\widetilde{\pi}_K\right)$ such that, for each $k\in\{1,\ldots,K\}$, either $\widetilde{\eta}_k=\eta_k$ or $\widetilde{\pi}_k=\pi_k$. Define

$$C\equiv\left\{k\in[K]:\widetilde{\eta}_k=\eta_k\right\},$$

with $[K]\equiv\{1,\ldots,K\}$ and

$$\overline{C}\equiv[K]\setminus C,$$

Thus, for each $k\in\{1,\ldots,K\}$, if $k\in C$, then $\widetilde{\eta}_k=\eta_k$ and, if $k\in\overline{C}$, then $\widetilde{\pi}_k=\pi_k$. Also, define the function $s:C\cup\{K+1\}\to C\cup\{0\}$ / $s\left(k\right)=\max\left\{r\in C\cup\{0\}:r<k\right\}$. Note that s is a bijection.

For notational convenience, we define $\widetilde{\pi}_0\equiv\widetilde{\eta}_0\equiv\widetilde{\eta}_{K+1}\equiv nill$. Now define $M_d^{K+1,K}\left(\psi,\widetilde{\eta}_{K+1},\widetilde{\overline{\pi}}_K\right)\equiv M_d^{K+1,K}\left(\psi,\widetilde{\overline{\pi}}_K\right)$, $M_d^{1,0}\left(\psi,\widetilde{\eta}_1,\widetilde{\pi}_0\right)\equiv M_d^{1,0}\left(\psi,\widetilde{\eta}_1\right)$ and $M_d^{0,0}\left(\psi,\widetilde{\eta}_0,\widetilde{\pi}_0\right)\equiv 0$. Then, we can write

$$\begin{aligned}U_d\left(\psi,\widetilde{\eta},\widetilde{\pi}\right)&=\sum_{k=0}^{K}\left\{M_d^{k+1,k}\left(\psi,\widetilde{\eta}_{k+1},\widetilde{\overline{\pi}}_k\right)-M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)\right\}\\&=\sum_{k=1}^{K+1}M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-\sum_{k=0}^{K}M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right).\end{aligned}$$

Hence, rearranging the terms in a convenient way, we arrive at

$$\begin{aligned}U_d\left(\psi,\widetilde{\eta},\widetilde{\pi}\right)=&\sum_{k\in C\cup\{K+1\}}\left\{M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{s(k),s(k)}\left(\psi,\widetilde{\eta}_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right)\right\}\\&-\sum_{k\in\overline{C}}\left\{M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)-M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)\right\}.\end{aligned}\tag{A.8}$$

To see this note that, since s is a bijection of $C\cup\{K+1\}$ in $C\cup\{0\}$,

$$\sum_{k\in C\cup\{K+1\}}M_d^{s(k),s(k)}\left(\psi,\widetilde{\eta}_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right)=\sum_{k\in C\cup\{0\}}M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)$$

Therefore, the right hand side of (A.8) is equal to

$$\begin{aligned}
&\sum_{k\in C\cup\{K+1\}} M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-\sum_{k\in C\cup\{0\}} M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)\\
&-\sum_{k\in\overline{C}} M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)+\sum_{k\in\overline{C}} M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)\\
&=\sum_{k=1}^{K+1} M_d^{k,k-1}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-\sum_{k=0}^{K} M_d^{k,k}\left(\psi,\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)\\
&=U_d\left(\psi,\widetilde{\eta},\widetilde{\pi}\right).
\end{aligned}$$

Equation (A.8) implies that, to prove fact (a), it suffices to prove that

(I) $E_P\left\{M_d^{k,k-1}\left(\psi\left(P\right),\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{s(k),s(k)}\left(\psi\left(P\right),\widetilde{\eta}_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right)\right\}=0$

for every $k\in C\cup\{K+1\}$ and $P\in\mathcal{M}$, and

(II) $E_P\left\{M_d^{k,k}\left(\psi\left(P\right),\widetilde{\eta}_k,\widetilde{\overline{\pi}}_k\right)-M_d^{k,k-1}\left(\psi\left(P\right),\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)\right\}=0$

for every $k\in\overline{C}$ and $P\in\mathcal{M}$.

Fact (II) follows from fact (1) above because $\widetilde{\pi}_k=\pi_k$ when $k\in\overline{C}$.

To prove of fact (I), we consider the following six settings:

(i) $k\in C$ and $1\leq s\left(k\right)=k-1$,

(ii) $k\in C$ and $1\leq s\left(k\right)<k-1$,

(iii) $k\in C$ and $s\left(k\right)=0$,

(iv) $k=K+1$ and $s\left(K+1\right)=K$,

(v) $k=K+1$ and $1\leq s\left(K+1\right)<K$, and

(vi) $k=K+1$ and $s\left(K+1\right)=0$.

Under setting (i), $\widetilde{\eta}_k=\eta_k$ because $k\in C$. Also, $s\left(k\right)=k-1$ and $s\left(k\right)\in C$ by definition of $s\left(\cdot\right)$ and because $s\left(k\right)\geq 1$. Then, $\widetilde{\eta}_{k-1}=\eta_{k-1}$ and, hence,

$$\begin{aligned}
&E_P\left\{M_d^{k,k-1}\left(\psi\left(P\right),\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{s(k),s(k)}\left(\psi\left(P\right),\widetilde{\eta}_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right)\right\}\\
&=E_P\left\{M_d^{k,k-1}\left(\psi\left(P\right),\eta_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{k-1,k-1}\left(\psi\left(P\right),\eta_{k-1},\widetilde{\overline{\pi}}_{k-1}\right)\right\}
\end{aligned}$$

which is equal to zero by (2.a).

Under setting (ii), $\widetilde{\eta}_k=\eta_k$ because $k\in C$. Likewise, $\widetilde{\eta}_{s(k)}=\eta_{s(k)}$ since $s\left(k\right)\in C$, by definition of $s\left(\cdot\right)$ and because $s\left(k\right)\geq 1$. Hence,

$$\begin{aligned}
&M_d^{k,k-1}\left(\psi\left(P\right),\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{s(k),s(k)}\left(\psi\left(P\right),\widetilde{\eta}_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right)\\
&=M_d^{k,k-1}\left(\psi\left(P\right),\eta_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{s(k),s(k)}\left(\psi\left(P\right),\eta_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right).
\end{aligned}$$

Now, we apply a telescopic sum and write

$$\begin{aligned}
&M_d^{k,k-1}\left(\psi\left(P\right),\eta_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{s(k),s(k)}\left(\psi\left(P\right),\eta_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right)\\
&=\sum_{j=s(k)+1}^{k}\left\{M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\widetilde{\overline{\pi}}_{j-1}\right)-M_d^{j-1,j-1}\left(\psi\left(P\right),\eta_{j-1},\widetilde{\overline{\pi}}_{j-1}\right)\right\}\\
&+\sum_{j=s(k)+1}^{k-1}\left\{M_d^{j,j}\left(\psi\left(P\right),\eta_j,\widetilde{\overline{\pi}}_{j}\right)-M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\widetilde{\overline{\pi}}_{j-1}\right)\right\}.
\end{aligned}$$

Notice that, for $j=s\left(k\right)+1,\ldots,k$,

$$E_P\left\{M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\widetilde{\overline{\pi}}_{j-1}\right)-M_d^{j-1,j-1}\left(\psi\left(P\right),\eta_{j-1},\widetilde{\overline{\pi}}_{j-1}\right)\right\}=0$$

by (2.a). Also note that, under this setting, for $j=s\left(k\right)+1,\ldots,k-1$, $\widetilde{\pi}_j=\pi_j$ and, hence,

$$E_P\left\{M_d^{j,j}\left(\psi\left(P\right),\eta_j,\widetilde{\overline{\pi}}_{j}\right)-M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\widetilde{\overline{\pi}}_{j-1}\right)\right\}=0$$

by (1). This concludes the proof of fact (I) under (ii).

Under (iii), $\widetilde{\eta}_k=\eta_k$ because $k\in C$. Also, $k=\min C$ because $s\left(k\right)=0$ and, hence, $\widetilde{\pi}_j=\pi_j$ for $j=1,\ldots,k-1$. Then,

$$\begin{aligned}
&M_d^{k,k-1}\left(\psi\left(P\right),\widetilde{\eta}_k,\widetilde{\overline{\pi}}_{k-1}\right)-M_d^{s(k),s(k)}\left(\psi\left(P\right),\widetilde{\eta}_{s(k)},\widetilde{\overline{\pi}}_{s(k)}\right)\\
&=M_d^{k,k-1}\left(\psi\left(P\right),\eta_k,\overline{\pi}_{k-1}\right)-M_d^{0,0}\left(\psi\left(P\right),\widetilde{\eta}_0,\widetilde{\pi}_0\right)\\
&=M_d^{k,k-1}\left(\psi\left(P\right),\eta_k,\overline{\pi}_{k-1}\right).
\end{aligned}$$

Again, we apply a telescopic sum and write

$$\begin{aligned}
&M_d^{k,k-1}\left(\psi\left(P\right),\eta_k,\overline{\pi}_{k-1}\right)\\
&=\sum_{j=2}^{k}\left\{M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\overline{\pi}_{j-1}\right)-M_d^{j-1,j-1}\left(\psi\left(P\right),\eta_{j-1},\overline{\pi}_{j-1}\right)\right\}\\
&+\sum_{j=1}^{k-1}\left\{M_d^{j,j}\left(\psi\left(P\right),\eta_j,\overline{\pi}_{j}\right)-M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\overline{\pi}_{j-1}\right)\right\}+M_d^{1,0}\left(\psi\left(P\right),\eta_1\right).
\end{aligned}$$

Notice that, for $j=2,\ldots,k$,

$$E_P\left\{M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\overline{\pi}_{j-1}\right)-M_d^{j-1,j-1}\left(\psi\left(P\right),\eta_{j-1},\overline{\pi}_{j-1}\right)\right\}=0$$

by (2.a). Also, for $j=1,\ldots,k-1$,

$$E_P\left\{M_d^{j,j}\left(\psi\left(P\right),\eta_j,\overline{\pi}_{j}\right)-M_d^{j,j-1}\left(\psi\left(P\right),\eta_j,\overline{\pi}_{j-1}\right)\right\}=0$$

by (1), and $E_P\left\{M_d^{1,0}\left(\psi\left(P\right),\eta_1\right)\right\}=0$ by (2.c). Hence, fact (I) also holds under this setting.

Under setting (iv), $k = K+1$ and $s(K+1) = K$, hence $\widetilde{\eta}_K = \eta_K$ and

$$\begin{aligned} & E_P\left\{M_d^{k,k-1}\left(\psi(P), \widetilde{\eta}_k, \widetilde{\overline{\pi}}_{k-1}\right) - M_d^{s(k),s(k)}\left(\psi(P), \widetilde{\eta}_{s(k)}, \widetilde{\overline{\pi}}_{s(k)}\right)\right\} \\ & = E_P\left\{M_d^{K+1,K}\left(\psi(P), \widetilde{\overline{\pi}}_K\right) - M_d^{K,K}\left(\psi(P), \eta_K, \widetilde{\overline{\pi}}_K\right)\right\} \end{aligned}$$

which is zero by (2.b).

Under (v), $k = K+1$ and $1 \leq s(K+1) < K$, then $\widetilde{\eta}_{s(K+1)} = \eta_{s(K+1)}$ and

$$\begin{aligned} & M_d^{k,k-1}\left(\psi(P), \widetilde{\eta}_k, \widetilde{\overline{\pi}}_{k-1}\right) - M_d^{s(k),s(k)}\left(\psi(P), \widetilde{\eta}_{s(k)}, \widetilde{\overline{\pi}}_{s(k)}\right) \\ & = M_d^{K+1,K}\left(\psi(P), \widetilde{\overline{\pi}}_K\right) - M_d^{s(K+1),s(K+1)}\left(\psi(P), \eta_{s(K+1)}, \widetilde{\overline{\pi}}_{s(K+1)}\right). \end{aligned}$$

Analogously to setting (ii), applying a telescopic sum, we have that the right hand side of last equation is equal to

$$\begin{aligned} & M_d^{K+1,K}\left(\psi(P), \widetilde{\overline{\pi}}_K\right) - M_d^{K,K}\left(\psi(P), \eta_K, \widetilde{\overline{\pi}}_K\right) \\ & + \sum_{j=s(K+1)+1}^{K}\left\{M_d^{j,j-1}\left(\psi(P), \eta_j, \widetilde{\overline{\pi}}_{j-1}\right) - M_d^{j-1,j-1}\left(\psi(P), \eta_{j-1}, \widetilde{\overline{\pi}}_{j-1}\right)\right\} \\ & + \sum_{j=s(K+1)+1}^{K}\left\{M_d^{j,j}\left(\psi(P), \eta_j, \widetilde{\overline{\pi}}_j\right) - M_d^{j,j-1}\left(\psi(P), \eta_j, \widetilde{\overline{\pi}}_{j-1}\right)\right\}. \end{aligned}$$

Notice that

$$E_P\left\{M_d^{K+1,K}\left(\psi(P), \widetilde{\overline{\pi}}_K\right) - M_d^{K,K}\left(\psi(P), \eta_K, \widetilde{\overline{\pi}}_K\right)\right\} = 0$$

by (2.b) and, for $j = s(K+1)+1, \ldots, K$,

$$E_P\left\{M_d^{j,j-1}\left(\psi(P), \eta_j, \widetilde{\overline{\pi}}_{j-1}\right) - M_d^{j-1,j-1}\left(\psi(P), \eta_{j-1}, \widetilde{\overline{\pi}}_{j-1}\right)\right\} = 0$$

by (2.a). Also note that, under this setting, for $j = s(K+1)+1, \ldots, K$, $\widetilde{\pi}_j = \pi_j$ and, hence,

$$E_P\left\{M_d^{j,j}\left(\psi(P), \eta_j, \widetilde{\overline{\pi}}_j\right) - M_d^{j,j-1}\left(\psi(P), \eta_j, \widetilde{\overline{\pi}}_{j-1}\right)\right\} = 0$$

by (1). This concludes the proof of fact (I) under (v).

Finally, under setting (vi), $s(K+1) = 0$, which implies that $C = \emptyset$ and, hence, that $\widetilde{\pi}_j = \pi_j$ for $j = 1, \ldots, K$. Then,

$$\begin{aligned} & E_P\left\{M_d^{k,k-1}\left(\psi(P), \widetilde{\eta}_k, \widetilde{\overline{\pi}}_{k-1}\right) - M_d^{s(k),s(k)}\left(\psi(P), \widetilde{\eta}_{s(k)}, \widetilde{\overline{\pi}}_{s(k)}\right)\right\} \\ & = E_P\left\{M_d^{K+1,K}\left(\psi(P), \overline{\pi}_K\right) - M_d^{0,0}\left(\psi(P), \widetilde{\eta}_0, \widetilde{\pi}_0\right)\right\} \\ & = E_P\left\{M_d^{K+1,K}\left(\psi(P), \overline{\pi}_K\right)\right\} \end{aligned}$$

which is zero because $E_P\left[\pi_K^p\left(\overline{A}_K, \overline{L}_K\right)^{-1}\left\{Y - m\left(\overline{A}_K, Z; \psi(P)\right)\right\}|\overline{A}_K, Z\right] = 0$ by [35].

Fact (b) follows from fact (a) because, for each $k = 1, \ldots, K-1$, the model for $\eta_k\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k\right)$ implied by $\mathcal{R}_k$ can be regarded as a MSMM with parameter $\left(\psi(P), \overline{\gamma}_k(P), \overline{\tau}_k(P)\right)$ in longitudinal study with $K-k$ time points, outcome Y, treatment variables $A_{k+1}, \ldots, A_K$, and with $\left(\overline{A}_k, \overline{L}_k\right)$ playing the role of Z and L_{k+1} playing the role of V.

A.3.3 Proof of Lemma 2

Fact (a) follows immediately from the assumptions of the lemma. To see fact (b), first note that, by part (iv) of Condition SPob, it suffices to show that $\psi(P) = \theta_{3K+1}(P)$ solves the equation in $\psi = \theta_{3K+1}$,

$$E_P\left\{\phi^{3K+1}_{\left(\overline{\theta}^{\dagger}_{3K}(P),\theta_{3K+1}\right)}\right\} = 0.$$

Now, notice that

$$\phi^{3K+1}_{\left(\overline{\theta}^{\dagger}_{3K}(P),\theta_{3K+1}\right)} = U_d\left(\psi, \eta^{\dagger}, \pi^{\dagger}\right)$$

with U_d the function defined in (1.20) and with

$$d\left(\overline{a}_K, z\right) \equiv \overset{\bullet}{m}\left(\overline{a}_K, z; \psi^{\dagger(K)}(P)\right),$$

$$\eta^{\dagger}_k \equiv \eta_k\left(\overline{a}_K, \overline{l}_k; \psi^{\dagger(k)}(P), \overline{\gamma}^{\dagger(k)}_k(P), \overline{\tau}^{\dagger}_k(P)\right)$$

and

$$\pi^{\dagger}_k \equiv \pi_k\left(\overline{a}_k, \overline{l}_k; \alpha^{\dagger}_k(P)\right),$$

$k = 1, \ldots, K$. Then, to see (b) ti suffices to show that, for every $P \in \mathcal{F}$, $\psi(P)$ solves

$$E_P\left\{U_d\left(\psi, \eta^{\dagger}, \pi^{\dagger}\right)\right\} = 0.$$

Since $\mathcal{F} \subseteq \mathcal{M}$, Proposition 1 implies that it holds if, when $P \in \mathcal{F}$, for each $k \in \{1, \ldots, K\}$, either $\eta^{\dagger}_k = \eta_k$ or $\pi^{\dagger}_k = \pi_k$. Now, the facts that $\mathcal{R}_k \subset \mathcal{R}_{k-1}, k = 2, \ldots, K$, and $\mathcal{R}_1 \subset \mathcal{M}$ imply that model $\mathcal{F}$ can also be written as

$$\mathcal{F} = \mathcal{M} \cap \left\{\bigcap_{k=1}^{K}\left(\mathcal{P}_k \cup \mathcal{R}_k\right)\right\}.$$

Thus, if $P \in \mathcal{F}$ then, for each $k = 1, \ldots, K$, either $P \in \mathcal{P}_k$ or $P \in \left[\mathcal{R}_k \cap \left\{\bigcap_{s=k+1}^{K}\left(\mathcal{P}_s \cup \mathcal{R}_s\right)\right\}\right]$. It implies that, to prove (b), it suffices to show that

(1) for each $k \in \{1, \ldots, K\}$, if $P \in \mathcal{P}_k$ then $\pi^{\dagger}_k = \pi_k$, and

(2) for each $k \in \{1, \ldots, K\}$, if $P \in \mathcal{R}_k \cap \left\{\bigcap_{s=k+1}^{K}\left(\mathcal{P}_s \cup \mathcal{R}_s\right)\right\}$ then $\eta^{\dagger}_k = \eta_k$.

Fact (1) follows from the fact that $\alpha^{\dagger}_k(P) = \alpha_k(P)$ for $P \in \mathcal{P}_k, k = 1, \ldots, K$. This is because (a) if $P \in \mathcal{P}_k$, then $\alpha_k(P)$ solves $E_P\left(\phi^k_{\alpha_k}\right) = 0$ since $\phi^k_{\alpha_k}(o)$ is the score for α_k under model $\mathcal{P}_k$ and Condition D is verified, and (b) the equation in α_k, $E_P\left(\phi^k_{\alpha_k}\right) = 0$, has a unique solution at $\alpha^{\dagger}_k(P)$ by assumption (i).

To see fact (2), first note that, for each $k = 1, \ldots, K$, if $P \in \mathcal{R}_k$, then $\tau^{\dagger}_k(P) = \tau_k(P)$. It follows from the facts that (a) $\tau_k(P)$ solves $E_P\left(\phi^{K+k}_{\tau_k}\right) = 0$ when $P \in \mathcal{R}_k$, and (b) the equation in τ_k, $E_P\left(\phi^{K+k}_{\tau_k}\right) = 0$, has a unique solution at $\tau^{\dagger}_k(P)$ by assumption (ii). Moreover, since $\mathcal{R}_k \subset \mathcal{R}_{k-1}, k = 2, \ldots, K$, then $\overline{\tau}^{\dagger}_k(P) = \overline{\tau}_k(P)$ when $P \in \mathcal{R}_k, k = 1, \ldots, K$. Thus, we would arrive at the desired result if we show that

$$\left(\psi^{\dagger(K)}(P), \overline{\gamma}^{\dagger}_K{}^{(K)}(P)\right) = \left(\psi(P), \overline{\gamma}_K(P)\right) \text{ if } P \in \mathcal{R}_K \tag{A.9}$$

and that, for each $k = 1, \ldots, K-1$,

$$\left(\psi^{\dagger(k)}(P), \overline{\gamma}_k^{\dagger(k)}(P)\right) = \left(\psi(P), \overline{\gamma}_k(P)\right) \text{ if } P \in \mathcal{R}_k \cap \left\{\bigcap_{s=k+1}^{K} (\mathcal{P}_s \cup \mathcal{R}_s)\right\}. \tag{A.10}$$

We will prove (A.9) and (A.10) by backward induction. Fact (A.9) follows from assumption (iii) and from the fact that, when $P \in \mathcal{R}_K$, $\left(\psi(P), \overline{\gamma}_K(P)\right)$ solves the equation in $\left(\psi^{(K)}, \overline{\gamma}_K^{(K)}\right) = \theta_{2K+1}$,

$$E_P\left\{\phi^{2K+1}_{\left(\overline{\theta}^{\dagger}_{2K}(P), \theta_{2K+1}\right)}\right\} = 0.$$

This is because $\mathcal{R}_K$ is a regression model for the outcome Y with covariates $\left(\overline{A}_K, \overline{L}_K\right)$ and $\overline{\tau}_K^{\dagger}(P) = \overline{\tau}_K(P)$ if $P \in \mathcal{R}_K$. Now, given $1 \leq j \leq K-1$, assume that (A.10) holds for $k = j+1, \ldots, K$. We now show that (A.10) holds for $k = j$. By assumption (iii), it is enough to show that, if $P \in \mathcal{R}_j \cap \left\{\bigcap_{s=j+1}^{K} (\mathcal{P}_s \cup \mathcal{R}_s)\right\}$, then $\left(\psi(P), \overline{\gamma}_j(P)\right)$ solves the equation in $\left(\psi^{(j)}, \overline{\gamma}_j^{(j)}\right) = \theta_{3K+1-j}$,

$$E_P\left\{\phi^{3K+1-j}_{\left(\overline{\theta}^{\dagger}_{3K+1-(j+1)}(P), \theta_{3K+1-j}\right)}\right\} = 0.$$

If $P \in \mathcal{R}_j \cap \left\{\bigcap_{s=j+1}^{K} (\mathcal{P}_s \cup \mathcal{R}_s)\right\}$, then $\overline{\tau}_j^{\dagger}(P) = \overline{\tau}_j(P)$ and, hence,

$$\phi^{3K+1-j}_{\left(\overline{\theta}^{\dagger}_{3K+1-(j+1)}(P), \theta_{3K+1-j}\right)} = U_{d_j}^{j}\left\{\left(\psi^{(j)}, \overline{\gamma}_j^{(j)}, \overline{\tau}_j(P)\right), \underline{\eta}_{j+1}^{\dagger}, \underline{\pi}_{j+1}^{\dagger}\right\}$$

with U_{dj}^{j} the function defined in (1.36) and with

$$d_j\left(\overline{a}_K, \overline{l}_j\right) \equiv \frac{\dot{\eta}_j\left(\overline{a}_K, \overline{l}_j; \psi^{\dagger(K)}(P), \overline{\gamma}_j^{\dagger(K)}(P), \overline{\tau}_j(P)\right)}{\prod_{s=1}^{j} \pi_s\left(\overline{A}_s, \overline{L}_s; \alpha_s^{\dagger}(P)\right)},$$

$$\underline{\eta}_{j+1}^{\dagger} \equiv \left(\eta_{j+1}^{\dagger}, \ldots, \eta_K^{\dagger}\right)$$

and

$$\underline{\pi}_{j+1}^{\dagger} \equiv \left(\pi_{j+1}^{\dagger}, \ldots, \pi_K^{\dagger}\right).$$

Proposition 1 implies that, if $P \in \mathcal{R}_j$, then $\left(\psi(P), \overline{\gamma}_j(P), \overline{\tau}_j(P)\right)$ solves the equation in $\left(\psi, \overline{\gamma}_j, \overline{\tau}_j\right)$,

$$E_P\left[U_{d_j}^{j}\left\{\left(\psi, \overline{\gamma}_j, \overline{\tau}_j\right), \underline{\eta}_{j+1}^{\dagger}, \underline{\pi}_{j+1}^{\dagger}\right\}\right] = 0,$$

whenever, for each $s \in \{j+1, \ldots, K\}$, either $\eta_s^{\dagger} = \eta_s$ or $\pi_s^{\dagger} = \pi_s$. Then, to prove fact (A.10) for $k = j$, it is enough to show that, if $P \in \mathcal{R}_j \cap \left\{\bigcap_{s=j+1}^{K} (\mathcal{P}_s \cup \mathcal{R}_s)\right\}$ then, for each $s \in \{j+1, \ldots, K\}$, (i) if $P \in \mathcal{P}_s$ then $\pi_s^{\dagger} = \pi_s$ and (ii) if $P \in \mathcal{R}_s \cap \left\{\bigcap_{r=s+1}^{K} (\mathcal{P}_r \cup \mathcal{R}_r)\right\}$ then $\eta_s^{\dagger} = \eta_s$. Again, fact (i) follows from the fact that, if $P \in \mathcal{P}_s$ then $\alpha_s^{\dagger}(P) = \alpha_s(P), s = j+1, \ldots, K$. Finally, fact (ii) follows from the facts that, for $s = j+1, \ldots, K$, $\overline{\tau}_s^{\dagger}(P) = \overline{\tau}_s(P)$ for every $P \in \mathcal{R}_s$ and $\left(\psi^{\dagger(s)}(P), \overline{\gamma}_s^{\dagger(s)}(P)\right) = \left(\psi(P), \overline{\gamma}_s(P)\right)$ for every $P \in \mathcal{R}_s \cap \left\{\bigcap_{r=s+1}^{K} (\mathcal{P}_r \cup \mathcal{R}_r)\right\}$ by inductive hypothesis. This concludes the proof.

A.3.4 Proof of Lemma 3

To prove Lemma 3, we need to introduce the following proposition, which follows from Theorems 5.41 and 5.42 of [58] on consistency and asymptotic normality of Z-estimators.

Proposition 3 (from Theorems 5.41 and 5.42 of [58]) *Let $\mathcal{B}$ be an open subset of an Euclidean space and let X be a random vector with range in some subset $\mathcal{X}$ of an Euclidean space. Let $\left\{q_\beta(\cdot) : \mathcal{X} \to R^N : \beta \in \mathcal{B}\right\}$ be a collection of Borel measurable functions. Also, assume that*

(i) there exists $\beta_0 \in \mathcal{B}$ such that $E\left(q_{\beta_0}\right) = 0$,

(ii) $q_\beta(x)$ is twice continuously differentiable w.r.t. β for each $x \in \mathcal{X}$,

(iii) $E\left(\left\|q_{\beta_0}\right\|^2\right) < \infty$,

(iv) the matrix $E\left(\dot{q}_{\beta_0}\right)$ exists and is nonsingular, and

(v) the second-order partial derivatives of $q_\beta(\cdot)$ w.r.t. β, $\frac{\partial^2}{\partial\beta_i\beta_j}q_\beta(\cdot), 1 \leq i,j \leq p$, are dominated by a fixed integrable function in a neighborhood of β_0.

Then, for i.i.d. copies $X_1, \ldots, X_n$ of X, we have that

(a) there exists a sequence $\widetilde{\beta}_n$ solving $\mathbb{P}_n\left(q_\beta\right) = 0$ with probability tending to one, which converges to β_0 in probability, and

(b) every estimator sequence $\widehat{\beta}_n$ such that $\mathbb{P}_n\left(q_{\widehat{\beta}_n}\right) = 0$ with probability tending to one, that converges to β_0 in probability, satisfies

$$\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) = -E\left(\dot{q}_{\beta_0}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} q_{\beta_0}(X_i) + o_P(1).$$

that is, is asymptotically linear for β_0 with influence function

$$\varsigma(x) = -E\left(\dot{q}_{\beta_0}\right)^{-1} q_{\beta_0}(x).$$

In particular, the sequence $\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right)$ converges to a mean zero Normal distribution with variance $E\left(\dot{q}_{\beta_0}\right)^{-1} E\left(q_{\beta_0}q_{\beta_0}^T\right)\left\{E\left(\dot{q}_{\beta_0}\right)^{-1}\right\}^T$.

As noticed by Van der Vaart [58], the assertion of the proposition does not guarantee the existence of a consistent and asymptotically normal sequence of estimators. The only claim of Proposition 3 is that a clairvoyant statistician (with preknowledge of β_0) can choose a consistent and asymptotically normal sequence of roots. However, as also noticed by Van der Vaart, if the solution of the estimating equation $\mathbb{P}_n\left(q_\beta\right) = 0$ is unique with probability tending to one, then this unique solution must agree with that of the clairvoyant statistician and, hence, it must be CAN for β_0. Therefore, we can derive the consistency and asymptotic normality of $\widehat{\theta}$ by finding conditions guaranteeing that, for i.i.d. copies $O_1, \ldots, O_n$ of $O \sim P \in \mathcal{F}$, (a) ϕ_θ and $\theta^\dagger(P)$ verify

the assumptions of q_β and β_0 in Proposition 3, (b) the equation in θ, $\mathbb{P}_n(\phi_\theta) = 0$, has at most one solution with probability tending to one under P, and (c) $\widehat{\theta}$ solves that equation with probability tending to one under P. By part (a) of Lemma 2, if $P \in \mathcal{F}$ verify Condition SPob and Condition D holds, then ϕ_θ and $\theta^\dagger(P)$ verify the assumption (i) of q_β and β_0 in Proposition 3. In addition, conditions (b) and (c) hold if ϕ_θ and $\theta^\dagger(P)$ verify the assumptions of q_β and β_0 in Proposition 3 and the Condition S holds. These observations are key to prove Lemma 3, which we do now.

Proof of Lemma 3. First note that, by Lemma 2, $E_P\left\{\phi_{\theta^\dagger(P)}\right\} = 0$. This observation, together with Conditions D(ϕ_θ) to Domination, imply that ϕ_θ and $\theta^\dagger(P)$ verify the assumptions of q_β and β_0 in Proposition 3. Hence, there exists a sequence $\widetilde{\theta}_n$ solving $\mathbb{P}_n(\phi_\theta) = 0$ with probability tending to one under P that is asymptotically linear for $\theta^\dagger(P)$ with influence function $\xi(o) = -E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)^{-1}\phi_{\theta^\dagger(P)}(o)$. Thus, to conclude the proof, it suffices to show that $\widehat{\theta}_n (\equiv \widehat{\theta})$ is equal to $\widetilde{\theta}_n$ with probability tending to one under P. Let $B_n \equiv \left\{\widetilde{\theta}_n \text{ solves } \mathbb{P}_n(\phi_\theta) = 0\right\}$. In what follows, for simplicity, we omit the subscript n in $\widetilde{\theta}_n, \widehat{\theta}_n, C_n$ and B_n. Since $P(C \cap B) \longrightarrow 1$, it suffices to prove that $(C \cap B) \subseteq \bigcap_{s=1}^{3K+1}\left\{\widehat{\theta}_s = \widetilde{\theta}_s\right\}$. Here, for $s = 1, \ldots, 3K+1, \widetilde{\theta}_s$ is the vector whose elements are the components of $\widetilde{\theta}$ having the same subscripts as the components of $\widehat{\theta}$ that make up the vector $\widehat{\theta}_s$. The fact that $(C \cap B) \subseteq \left\{\widehat{\theta}_k = \widetilde{\theta}_k\right\}, k = 1, \ldots, K$, follows from the facts that, for $k = 1, \ldots, K$, (i) $B \subseteq \left\{\widetilde{\theta}_k \text{ solves } \mathbb{P}_n\left(\phi^k_{\theta_k}\right) = 0\right\}$, (ii) $\widehat{\theta}_k$ solves $\mathbb{P}_n\left(\phi^k_{\theta_k}\right) = 0$ by the definition of $\widehat{\theta}_k (= \widehat{\alpha}_k)$ in the estimation algorithm, and (iii) $C \subseteq \left\{\mathbb{P}_n\left(\phi^k_{\theta_k}\right) = 0 \text{ has at most one solution}\right\}$. With an identical argument, we can show that $C \cap B \subseteq \left\{\widehat{\theta}_{K+k} = \widetilde{\theta}_{K+k}\right\}, k = 1, \ldots, K$. Finally, the fact that $C \cap B \subseteq \left\{\widehat{\theta}_{3K+1-k} = \widetilde{\theta}_{3K+1-k}\right\}, k = 0, \ldots, K$, follows by backward induction in k, from the following facts:

(i) $B \subseteq \left\{\widetilde{\theta}_{3K+1-k} \text{ solves the equation in } \theta_{3K+1-k}, \mathbb{P}_n\left(\phi^{3K+1-k}_{\widetilde{\theta}_{3K+1-(k+1)}, \theta_{3K+1-k}}\right) = 0\right\}$,

(ii) if $(C \cap B) \subseteq \left\{\widehat{\widetilde{\theta}}_{3K+1-(k+1)} = \widetilde{\widetilde{\theta}}_{3K+1-(k+1)}\right\}$ then

$(C \cap B) \subseteq \left\{\widehat{\theta}_{3K+1-k} \text{ solves the equation in } \theta_{3K+1-k}, \mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)}, \theta_{3K+1-k}\right)}\right) = 0\right\}$,

(iii) $\left\{\widehat{\widetilde{\theta}}_{3K+1-(k+1)} = \widetilde{\widetilde{\theta}}_{3K+1-(k+1)}\right\} \subseteq \left\{\mathbb{P}_n\left(\phi^{3K+1-k}_{\widetilde{\theta}_{3K+1-(k+1)}, \theta_{3K+1-k}}\right) = \mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)}, \theta_{3K+1-k}\right)}\right)\right\}$,

and

(iv) if $(C \cap B) \subseteq \left\{\widehat{\widetilde{\theta}}_{3K+1-(k+1)} = \widetilde{\widetilde{\theta}}_{3K+1-(k+1)}\right\}$ then

$(C \cap B) \subseteq \left\{\text{the equation in } \theta_{3K+1-k},\ \mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)}, \theta_{3K+1-k}\right)}\right) = 0,\ \text{has at most one solution}\right\}$.

Fact (ii) follows from the facts that

$B \subseteq \left\{\text{the equation in } \theta_{3K+1-k},\ \mathbb{P}_n\left(\phi^{3K+1-k}_{\widetilde{\theta}_{3K+1-(k+1)}, \theta_{3K+1-k}}\right) = 0,\ \text{has a solution}\right\}$ and, hence, $C \cap B \subseteq \left\{\widehat{\widetilde{\theta}}_{3K+1-(k+1)} = \widetilde{\widetilde{\theta}}_{3K+1-(k+1)}\right\}$ implies that

$C \cap B \subseteq \left\{ \text{the equation in } \theta_{3K+1-k},\ \mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}\right)}\right) = 0,\ \text{has a solution} \right\}$. Also, if $C \cap B \subseteq \left\{\widehat{\widetilde{\theta}}_{3K+1-(k+1)} = \widetilde{\widetilde{\theta}}_{3K+1-(k+1)}\right\}$ then $C \cap B \subseteq \left\{\mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)}}\right) = 0\right\}$. Therefore, to conclude the proof of fact (ii), it suffices to show that, for $k = 0, \ldots, K$, if (1) $\mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)}}\right) = 0$ and (2) the equation in θ_{3K+1-k}, $\mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}\right)}\right) = 0$, has a solution, then $\widehat{\theta}_{3K+1-k}$ solves the equation in $\theta_{3K+1-k}, \mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}\right)}\right) = 0$. When $k = K$, it follows by the definition of $\widehat{\theta}_{2K+1}$ in step 3 of the MR estimation algorithm of Section 1.7.2. To see this for $k = 0, \ldots, K-1$, first note that if $\mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)}}\right) = 0$, then $\mathbb{P}_n\left(\sum_{j=k+1}^{K} \varphi^{k,j}_{\widehat{\widetilde{\theta}}_{3K+1-j}}\right) = 0$. This is because $\sum_{j=k+1}^{K} \varphi^{k,j}_{\widehat{\widetilde{\theta}}_{3K+1-j}}$ is a subvector of $\phi^{3K+1-(k+1)}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)}}$ by being $\overset{\bullet}{\eta}_k$ a subvector of $\overset{\bullet}{\eta}_{k+1}, k = 1, \ldots, K-1$, and $\overset{\bullet}{m}$ a subvector of $\overset{\bullet}{\eta}_1$. Hence, $\mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}\right)}\right) = \mathbb{P}_n\left(\varphi^{k,k}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}}\right)$, so that (2) implies that the equation in θ_{3K+1-k}, $\mathbb{P}_n\left(\varphi^{k,k}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}}\right) = 0$, has a solution and, then, $\widehat{\theta}_{3K+1-k}$ solves that equation by definition of $\widehat{\theta}_{3K+1-k}$ in step 4 of the MR estimation algorithm if $k = 1, \ldots, K-1$ or in step 5 if $k = 0$.

Fact (iv) follows from the facts that

(a) $C \subseteq \left\{ \text{if } \mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)}}\right) = 0,\ \text{then the equation in } \theta_{3K+1-k},\ \mathbb{P}_n\left(\phi^{3K+1-k}_{\left(\widehat{\widetilde{\theta}}_{3K+1-(k+1)},\theta_{3K+1-k}\right)}\right) = 0, \right.$ $\left. \text{has at most one solution} \right\}$,

(b) $B \subseteq \left\{\mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widetilde{\widetilde{\theta}}_{3K+1-(k+1)}}\right) = 0\right\}$, and

(c) $\left\{\widehat{\widetilde{\theta}}_{3K+1-(k+1)} = \widetilde{\widetilde{\theta}}_{3K+1-(k+1)}\right\} \subseteq \left\{\mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widetilde{\widetilde{\theta}}_{3K+1-(k+1)}}\right) = \mathbb{P}_n\left(\phi^{3K+1-(k+1)}_{\widehat{\widetilde{\theta}}_{3K+1-(k+1)}}\right)\right\}$.

This concludes the proof. ■

A.3.5 Proof of Theorem 1

To prove Theorem 1, we will exploit the fact that $E_P\left(\dot{\phi}_\theta\right)$ is a *lower-triangular-block-matrix* by using a recursive formula (provided in Lemma 12) to compute each component of a vector $u = -\Delta^{-1}v$ when Δ is a nonsingular lower-triangular-block-matrix. We start by introducing the definition of *lower-triangular-block-matrix.*

Definition 1 *A lower-triangular-block-matrix is a square matrix, having main diagonal blocks square matrices, such that the upper-diagonal blocks are zero matrices. A lower-triangular-block-matrix Δ has the form*

$\Delta = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \cdots & \Delta_{1N} \\ \Delta_{12} & \Delta_{22} & \cdots & \Delta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{N1} & \Delta_{N2} & \cdots & \Delta_{NN} \end{bmatrix}$ *where* Δ_{ij} *is a* $(d_i \times d_j)$ *matrix and* $\Delta_{ij} = 0_{d_i \times d_j}$ *if* $1 \leq i < j \leq N$.

Lemma 12 *Let* Δ *be a real-valued lower-triangular-block-matrix with nonsingular diagonal blocks, i.e., let*

$\Delta = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \cdots & \Delta_{1N} \\ \Delta_{12} & \Delta_{22} & \cdots & \Delta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{N1} & \Delta_{N2} & \cdots & \Delta_{NN} \end{bmatrix}$ *where*

(i) $\Delta_{ij} \in R^{d_i \times d_j}, 1 \leq i, j \leq N$

(ii) $\Delta_{ij} = 0_{d_i \times d_j}$ *if* $1 \leq i < j \leq N$ *and*

(iii) Δ_{ii} *is nonsingular for every* $1 \leq i \leq N$.

Then

(a) Δ *is nonsingular and*

(b) given $u = \left(u_1^T, \ldots, u_N^T\right)^T$ *and* $v = \left(v_1^T, \ldots, v_N^T\right)^T$ *with* $u_i, v_i \in R^{d_i}, i = 1, \ldots, N$, *if* $u = -\Delta^{-1} v$, *then*

$$u_1 = -\Delta_{11}^{-1} v_1$$

and, for $i = 2, \ldots, N$,

$$u_i = -\Delta_{ii}^{-1} \left(v_i + \sum_{j=1}^{i-1} \Delta_{ij} u_j \right).$$

Proof of Lemma 12. Fact (a) follows from the fact that, since Δ is a lower-triangular-block-matrix, then $det(\Delta) = \prod_{i=1}^{N} \det(\Delta_{ii})$, which is non zero by assumption 3 of the Lemma. To prove fact (b), note that $v = -\Delta u$ and, hence,

$$v_i = -\sum_{j=1}^{N} \Delta_{ij} u_j \text{ for } i = 1, \ldots, N.$$

Then, since $\Delta_{ij} = 0_{d_i \times d_j}$ for $i < j$, we have that (1) $v_1 = -\Delta_{11} u_1$ and, hence, $u_1 = -\Delta_{11}^{-1} v_1$, and (2) for $i \geq 2$, $v_i = -\sum_{j=1}^{i-1} \Delta_{ij} u_j - \Delta_{ii} u_i$, which then gives $u_i = -\Delta_{ii}^{-1} \left(v_i + \sum_{j=1}^{i-1} \Delta_{ij} u_j \right)$. ■

Part (a) of this lemma implies that Condition NonSing holds if the Condition NonSing2 is verified.

Proof of Theorem 1. To see (a), note that, since $E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)$ is a lower-triangular-block-matrix with diagonal $s-$block given by $E_P\left\{\left(\frac{\partial}{\partial \theta_s} \phi_{\theta_s}^s\right)\Big|_{\overline{\theta}_s = \overline{\theta}_s^\dagger(P)}\right\}, s = 1, \ldots, 3K+1$, Condition

NonSing2 and part (a) of Lemma 12 imply that $E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)$ is nonsingular. Then, Conditions S, D(ϕ_θ), Moment2, NonSing2 and Domination and Lemma 3 imply that $\widehat{\theta}$ is an asymptotically linear estimator of $\theta^\dagger(P)$ with influence function

$$\xi(o) = -E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)^{-1}\phi_{\theta^\dagger(P)}(o).$$

Hence, for each $s = 1, \ldots, 3K+1$, $\widehat{\theta}_s$ is an asymptotically linear estimator of $\theta_s^\dagger(P)$ with influence function $\xi_s(\cdot)$ where the $\xi_s(\cdot)'s$ are such that

$$\xi(\cdot) = \left(\xi_1(\cdot)^T, \ldots, \xi_{3K+1}^T(\cdot)\right)^T$$

and each $\xi_s(\cdot)$ has the same dimension as θ_s. The fact that $E_P\left(\dot{\phi}_{\theta^\dagger(P)}\right)$ is lower-triangular-block-matrix with nonsingular diagonal blocks and with $(s,j)-$block given by $E_P\left\{\left(\frac{\partial}{\partial\theta j}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\overline{\theta}_s^\dagger(P)}\right\}, 1 \leq j, s \leq 3K+1$ imply, by part (b) of Lemma 12, that

$$\xi_1(o) = -E_P\left\{\left(\frac{\partial}{\partial\theta_1}\phi^1_{\theta_1}\right)\Big|_{\theta_1=\theta_1^\dagger(P)}\right\}^{-1}\phi^1_{\theta_1^\dagger(P)}(o)$$

and, for $s = 2, \ldots, 3K+1$,

$$\xi_s(o) = -E_P\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\overline{\theta}_s^\dagger(P)}\right\}^{-1}\left[\phi^s_{\overline{\theta}_s^\dagger(P)}(o) + \sum_{j=1}^{s-1}E_P\left\{\left(\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\overline{\theta}_s^\dagger(P)}\right\}\xi_j(o)\right].$$

Furthermore, the fact that, for $s = 2, \ldots, 2K, \phi^s_{\overline{\theta}_s}$ depends on $\overline{\theta}_s$ only through θ_s, implies that $\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}(\cdot) = 0$, $j = 1, \ldots, s-1$, and hence, for $s = 1, \ldots, 2K$,

$$\xi_s(o) = -\left[E_P\left\{\left(\frac{\partial}{\partial\theta_s}\phi^s_{\theta s}\right)\Big|_{\theta_s=\theta_s^\dagger(P)}\right\}\right]^{-1}\phi^s_{\theta_s^\dagger(P)}(o).$$

This concludes the proof of fact (a).

To prove (b), we first introduce the following notation. For $1 \leq j \leq s \leq 3K+1$, define

$$\Delta_{sj} \equiv E_P\left\{\left(\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\overline{\theta}_s^\dagger(P)}\right\}$$

and

$$\widehat{\Delta}_{sj} \equiv \mathbb{P}_n\left\{\left(\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s=\widehat{\overline{\theta}}_s}\right\}.$$

Then,

$$\xi_1(o) = -\Delta_{11}^{-1}\phi^1_{\theta_1^\dagger(P)}(o)$$

and, for $s = 2, \ldots, 3K+1$,

$$\xi_s(o) = -\Delta_{ss}^{-1}\left\{\phi^s_{\overline{\theta}^\dagger_s(P)}(o) + \sum_{j=1}^{s-1}\Delta_{sj}\xi_j(o)\right\}.$$

Also, since $\frac{\partial}{\partial\theta_j}\phi^s_{\theta_s}(\cdot) = 0$ for every $j = 1, \ldots, s-1$ and $s = 2, \ldots, 2K$, the $\widehat{\xi}'_s s$ defined in (1.38) and (1.39) verify the equations

$$\widehat{\xi}_1(o) = -\widehat{\Delta}_{11}^{-1}\phi^1_{\widehat{\theta}_1}(o)$$

if $\widehat{\Delta}_{11}$ is nonsingular and, for $s = 2, \ldots, 3K+1$,

$$\widehat{\xi}_s(o) = -\widehat{\Delta}_{ss}^{-1}\left\{\phi^s_{\widehat{\theta}_s}(o) + \sum_{j=1}^{s-1}\widehat{\Delta}_{sj}\widehat{\xi}_j(o)\right\}$$

if $\widehat{\Delta}_{ss}$ is nonsingular.

Our proof of fact (b) relies on the following facts:

(1) $\widehat{\Delta}_{sj} \xrightarrow{p} \Delta_{sj}$, $1 \le j \le s \le 3K+1$, under P,

(2) $\widehat{\Delta}_{ss}$ is nonsingular with probability tending to one and $\widehat{\Delta}_{ss}^{-1} \xrightarrow{p} \Delta_{ss}^{-1}, 1 \le s \le 3K+1$ under P, and

(3) $\mathbb{P}_n\left(\phi^s_{\widehat{\theta}_s}\phi^{jT}_{\widehat{\theta}_j}\right) \xrightarrow{p} E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\phi^{jT}_{\overline{\theta}^\dagger_j(P)}\right)$, $1 \le s, j \le 3K+1$, under P.

(4) $\mathbb{P}_n\left(\phi^s_{\widehat{\theta}_s}\widehat{\xi}_j^T\right) \xrightarrow{p} E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\xi_j^T\right)$ under P for every $1 \le j < s \le 3K+1$,

(5) $\mathbb{P}_n\left(\widehat{\xi}_1\widehat{\xi}_1^T\right) \xrightarrow{p} E_P\left(\xi_1\xi_1^T\right)$ and

(6) $\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_k^T\right) \xrightarrow{p} E_P\left(\xi_j\xi_k^T\right)$ under P for every $1 \le j, k \le 3K, j \ne k$.

By Conditions Domination and M, there exists a neighborhood of $\theta^\dagger(P)$, throughout denoted by $\mathcal{N}$ such that (a) $\phi_\theta(o)$ and its first-order partial derivatives w.r.t. θ are measurable w.r.t. o for every $\theta \in \mathcal{N}$, and (b) the second-order partial derivatives of $\phi_\theta(o)$ w.r.t. θ are dominated by a fixed integrable function in $\mathcal{N}$.

To see fact (1), note that $\widehat{\overline{\theta}}_s$ converges in probability to $\overline{\theta}^\dagger_s(P)$ under P. Then, Lemma 14 in Appendix A.4 implies that fact (1) holds if $\frac{\partial}{\partial\theta_j}\phi^s_{\theta_s}(o)$ is regular in some neighborhood of $\overline{\theta}^\dagger_s(P), 1 \le s \le j \le 3K+1$, according to Definition 2 of Appendix A.4 where o plays the roll of x and $\overline{\theta}_s$ plays the roll of β. Let $\mathcal{N}_s$ be a compact convex neighborhood of $\overline{\theta}^\dagger_s(P)$ included in $\left\{\overline{\theta}_s : \left(\overline{\theta}_s, \underline{\theta}_{s+1}\right) \in \mathcal{N} \text{ for some } \underline{\theta}_{s+1}\right\}$. Here, $\underline{\theta}_{s+1} \equiv \left(\theta^T_{s+1}, \ldots, \theta^T_{3K+1}\right)^T$. Note that $\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}(o)$ is regular in $\mathcal{N}_s, 1 \le s \le j \le 3K+1$, because of the following facts:

(I) $\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}(o)$ is measurable w.r.t. o for every $\overline{\theta}_s \in \mathcal{N}_s$ by Condition M,

(II) $\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}(o)$ is dominated by a fixed integrable function in $\mathcal{N}_s$ by the mean value theorem because (a) $\mathcal{N}_s$ is compact and convex, (b) the first-order partial derivatives of $\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}(o)$ w.r.t. θ are dominated by a fixed integrable function in $\mathcal{N}_s$ by Condition Domination, and (c) $E_P\left\{\left(\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}\right)\Big|_{\overline{\theta}_s = \overline{\theta}^\dagger_s(P)}\right\}$ exists by Condition NonSing2, and

(III) for each fixed o, $\frac{\partial}{\partial\theta_j}\phi^s_{\overline{\theta}_s}(o)$ is a continuous of $\overline{\theta}_s$ in $\mathcal{N}_s$ by Condition D(ϕ_θ).

Fact (2) follows because, for $1 \leq s \leq 3K+1$, Δ_{ss} is nonsingular by Condition NonSing2 and $\widehat{\Delta}_{ss} \xrightarrow{p} \Delta_{ss}$ under P by (1).

To prove fact (3), assume, without loss of generality, that $s \leq j$. Since $\widehat{\overline{\theta}}_j$ converges in probability to $\overline{\theta}^\dagger_j(P)$ under P, Lemma 14 in Appendix A.4 implies that fact (3) holds if $\phi^s_{\overline{\theta}_s}(o)\,\phi^j_{\overline{\theta}_j}(o)^T$ is regular in some neighborhood of $\overline{\theta}^\dagger_j(P)$ (according to Definition 2 of Appendix A.4 where o plays the roll of x and $\overline{\theta}_s$ plays the roll of β) which is verified if $\phi^s_{\overline{\theta}_s}(o)$ is regular in $\mathcal{N}_s$, for every $s = 1, \ldots, 3K+1$. Given $s \in \{1, \ldots, 3K+1\}$, the fact that $\phi^s_{\overline{\theta}_s}(o)$ is regular in $\mathcal{N}_s$ is a consequence of the following facts:

(I) $\phi^s_{\overline{\theta}_s}(o)$ is measurable w.r.t. o for every $\overline{\theta}_s \in \mathcal{N}_s$ by Condition M,

(II) $\phi^s_{\overline{\theta}_s}(o)$ is dominated by a fixed integrable function in $\mathcal{N}_s$ by the mean value theorem, because (a) $\mathcal{N}_s$ is compact and convex, (b) the first-order partial derivatives of $\phi^s_{\overline{\theta}_s}(o)$ are dominated by a fixed integrable function in $\mathcal{N}_s$, and (c) $E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\right) = 0$ by Lemma 2, and

(III) for each o, $\phi^s_{\overline{\theta}_s}(o)$ is continuous w.r.t. $\overline{\theta}_s$ in $\mathcal{N}_s$ by Condition D(ϕ_θ).

Turn to fact (4). Given $s = 2, \ldots, 3K+1$, we will show that

$$\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\widehat{\xi}^T_j\right) \xrightarrow{p} E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\xi^T_j\right) \tag{A.11}$$

under P for every $1 \leq j \leq s-1$ by induction in j. When $j = 1$, (A.11) follows from facts (2) and (3) and from the facts that

$$\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\widehat{\xi}^T_1\right) = -\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\phi^{1T}_{\widehat{\theta}_1}\right)\left(\widehat{\Delta}^{-1}_{11}\right)^T$$

if $\widehat{\Delta}_{11}$ is nonsingular, and

$$E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\xi^T_1\right) = -E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\phi^{1T}_{\theta^\dagger_1(P)}\right)\left(\Delta^{-1}_{11}\right)^T.$$

Now, given $2 \leq j \leq s-1$, assume that $\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\widehat{\xi}^T_k\right) \xrightarrow{p} E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\xi^T_k\right)$ for every $1 \leq k \leq j-1$. We want to show that $\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\widehat{\xi}^T_j\right) \xrightarrow{p} E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\xi^T_j\right)$. Note that

$$\begin{aligned}\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\widehat{\xi}^T_j\right) &= -\mathbb{P}_n\left\{\phi^s_{\widehat{\overline{\theta}}_s}\left(\phi^j_{\widehat{\overline{\theta}}_j} + \sum_{k=1}^{j-1}\widehat{\Delta}_{jk}\widehat{\xi}_k\right)^T\right\}\left(\widehat{\Delta}^{-1}_{ss}\right)^T \\ &= -\left\{\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\phi^{jT}_{\widehat{\overline{\theta}}_j}\right) + \sum_{k=1}^{j-1}\mathbb{P}_n\left(\phi^s_{\widehat{\overline{\theta}}_s}\widehat{\xi}^T_k\right)\widehat{\Delta}^T_{jk}\right\}\left(\widehat{\Delta}^{-1}_{ss}\right)^T\end{aligned}$$

if $\widehat{\Delta}_{ss}$ is nonsingular, and

$$E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\xi^T_j\right) = -\left\{E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\phi^{jT}_{\overline{\theta}^\dagger_j(P)}\right) + \sum_{k=1}^{j-1}E_P\left(\phi^s_{\overline{\theta}^\dagger_s(P)}\xi^T_k\right)\Delta^T_{jk}\right\}\left(\Delta^{-1}_{ss}\right)^T.$$

Then $\mathbb{P}_n\left(\phi^s_{\widehat{\theta}_s}\widehat{\xi}_j^T\right) \xrightarrow{p} E_P\left(\phi^s_{\theta^\dagger_s(P)}\xi_j^T\right)$ by inductive hypothesis and facts (1)-(3).
Now, fact (5) follows from facts (2) and (3) and from the facts that

$$\mathbb{P}_n\left(\widehat{\xi}_1\widehat{\xi}_1^T\right) = \widehat{\Delta}_{11}^{-1}\mathbb{P}_n\left(\phi^1_{\widehat{\theta}_1}\phi^{1T}_{\widehat{\theta}_1}\right)\left(\widehat{\Delta}_{11}^{-1}\right)^T$$

if $\widehat{\Delta}_{11}$ is nonsingular, and

$$E_P\left(\xi_1\xi_1^T\right) = \Delta_{11}^{-1}E_P\left(\phi^1_{\theta^\dagger_1(P)}\phi^{1T}_{\theta^\dagger_1(P)}\right)\left(\Delta_{11}^{-1}\right)^T.$$

Finally, to see fact (6), first note that it is equivalent to the fact that, for $k = 2, \ldots, 3K$,

$$\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_k^T\right) \xrightarrow{p} E_P\left(\xi_j\xi_k^T\right) \quad \text{for every } j = 1, \ldots, k-1. \tag{A.12}$$

We will prove (A.12) by induction in k. When $k = 2$, (A.12) reduces to the fact that $\mathbb{P}_n\left(\widehat{\xi}_1\widehat{\xi}_2^T\right) \xrightarrow{p} E_P\left(\xi_1\xi_2^T\right)$, which follows from facts (1), (2), (4) and (5) and from the facts that

$$\begin{aligned}\mathbb{P}_n\left(\widehat{\xi}_1\widehat{\xi}_2^T\right) &= -\mathbb{P}_n\left(\widehat{\xi}_1\left\{\phi^2_{\widehat{\theta}_2} + \widehat{\Delta}_{21}\widehat{\xi}_1\right\}^T\right)\left(\widehat{\Delta}_{22}^{-1}\right)^T \\ &= -\left\{\mathbb{P}_n\left(\widehat{\xi}_1\phi^{2T}_{\widehat{\theta}_2}\right) + \mathbb{P}_n\left(\widehat{\xi}_1\widehat{\xi}_1^T\right)\widehat{\Delta}_{21}^T\right\}\left(\widehat{\Delta}_{22}^{-1}\right)^T\end{aligned}$$

if $\widehat{\Delta}_{22}$ is nonsingular, and

$$E_P\left(\xi_1\xi_2^T\right) = -\left\{E_P\left(\xi_1\phi^{2T}_{\theta^\dagger_2(P)}\right) + E_P\left(\xi_1\xi_1^T\right)\Delta_{21}^T\right\}\left(\Delta_{22}^{-1}\right)^T.$$

Now, given $k = 3, \ldots, 3K$, suppose that $\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_h^T\right) \xrightarrow{p} E_P\left(\xi_j\xi_h^T\right)$ for every $1 \le j \le h-1$ and $1 \le h \le k-1$. We want to show that $\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_k^T\right) \xrightarrow{p} E_P\left(\xi_j\xi_k^T\right)$ for every $1 \le j \le k-1$. But, note that

$$\begin{aligned}\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_k^T\right) &= -\mathbb{P}_n\left\{\widehat{\xi}_j\left(\phi^k_{\widehat{\theta}_k} + \sum_{h=1}^{k-1}\widehat{\Delta}_{kh}\widehat{\xi}_h\right)^T\right\}\left(\widehat{\Delta}_{kk}^{-1}\right)^T \\ &= -\left\{\mathbb{P}_n\left(\widehat{\xi}_j\phi^{kT}_{\widehat{\theta}_k}\right) + \sum_{h=1}^{k-1}\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_h^T\right)\widehat{\Delta}_{kh}^T\right\}\left(\widehat{\Delta}_{kk}^{-1}\right)^T\end{aligned}$$

if $\widehat{\Delta}_{kk}$ is nonsingular, and

$$E_P\left(\xi_j\xi_k^T\right) = -\left\{E_P\left(\xi_j\phi^{kT}_{\theta^\dagger_k(P)}\right) + \sum_{h=1}^{k-1}E_P\left(\xi_j\xi_h^T\right)\Delta_{kh}^T\right\}\left(\Delta_{kk}^{-1}\right)^T.$$

Then, $\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_k^T\right) \xrightarrow{p} E_P\left(\xi_j\xi_k^T\right)$ by facts (1), (2), (4) and the inductive hypothesis.

We are now in conditions to show (b). We will prove it by induction. Fact (b) for $s = 1$ follows from fact (5). Now, given $s = 2, \ldots, 3K+1$, assume that $\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_j^T\right) \xrightarrow{p} E_P\left(\xi_j\xi_j^T\right)$ for every $1 \leq j \leq s-1$. We want to show that $\mathbb{P}_n\left(\widehat{\xi}_s\widehat{\xi}_s^T\right) \xrightarrow{p} E_P\left(\xi_s\xi_s^T\right)$. To see this, note that

$$\mathbb{P}_n\left(\widehat{\xi}_s\widehat{\xi}_s^T\right) = \widehat{\Delta}_{ss}^{-1}\mathbb{P}_n\left(\widehat{\Psi}_s\widehat{\Psi}_s^T\right)\left(\widehat{\Delta}_{ss}^{-1}\right)^T$$

if $\widehat{\Delta}_{ss}$ is nonsingular, and

$$E_P\left(\xi_s\xi_s^T\right) = \Delta_{ss}^{-1}E_P\left(\Psi_s\Psi_s^T\right)\left(\Delta_{ss}^{-1}\right)^T$$

with $\widehat{\Psi}_s \equiv \phi^s_{\widehat{\theta}_s} + \sum_{j=1}^{s-1}\widehat{\Delta}_{sj}\widehat{\xi}_j$ and $\Psi_s \equiv \phi^s_{\theta^\dagger_s(P)} + \sum_{j=1}^{s-1}\Delta_{sj}\xi_j$.

Also notice that

$$\mathbb{P}_n\left(\widehat{\Psi}_s\widehat{\Psi}_s^T\right) = \mathbb{P}_n\left(\phi^s_{\widehat{\theta}_s}\phi^{sT}_{\widehat{\theta}_s}\right) + \sum_{j=1}^{s-1}\mathbb{P}_n\left(\phi^s_{\widehat{\theta}_s}\widehat{\xi}_j^T\right)\widehat{\Delta}_{sj}^T + \sum_{j=1}^{s-1}\widehat{\Delta}_{sj}\mathbb{P}_n\left(\widehat{\xi}_j\phi^{sT}_{\widehat{\theta}_s}\right) + \sum_{j=1}^{s-1}\sum_{\substack{k=1\\k\neq j}}^{s-1}\widehat{\Delta}_{sj}\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_k^T\right)\widehat{\Delta}_{sk}^T$$
$$+ \sum_{j=1}^{s-1}\widehat{\Delta}_{sj}\mathbb{P}_n\left(\widehat{\xi}_j\widehat{\xi}_j^T\right)\widehat{\Delta}_{sj}^T$$

and

$$E_P\left(\Psi_s\Psi_s^T\right) = E_P\left(\phi^s_{\theta^\dagger_s(P)}\phi^{sT}_{\theta^\dagger_s(P)}\right) + \sum_{j=1}^{s-1}E_P\left(\phi^s_{\theta^\dagger_s(P)}\xi_j^T\right)\Delta_{sj}^T + \sum_{j=1}^{s-1}\Delta_{sj}E_P\left(\xi_j\phi^{sT}_{\theta^\dagger_s(P)}\right)$$
$$+ \sum_{j=1}^{s-1}\sum_{\substack{k=1\\k\neq j}}^{s-1}\Delta_{sj}E_P\left(\xi_j\xi_k^T\right)\Delta_{sk}^T + \sum_{j=1}^{s-1}\Delta_{sj}E_P\left(\xi_j\xi_j^T\right)\Delta_{sj}^T.$$

Then, facts (1)-(4) and (6) and the inductive hypothesis imply that $\mathbb{P}_n\left(\widehat{\xi}_s\widehat{\xi}_s^T\right) \xrightarrow{p} E_P\left(\xi_s\xi_s^T\right)$ as we wanted to show. ■

A.3.6 Proof of Lemma 5

First note that, for $k = 1, \ldots, K$, the identity (1.44) implies that

$$\frac{\partial}{\partial\left(\psi, \overline{\gamma}_k\right)}\eta_k\left(\overline{a}_K, \overline{l}_k; \psi, \overline{\gamma}_k, \overline{\tau}_k\right) = H_k\left(\overline{a}_K, \overline{l}_k; \overline{\tau}_k\right).$$

Then, $\phi^{2K+1}_{\overline{\alpha}_K, \overline{\tau}_K, \psi^{(K)}, \overline{\gamma}_K^{(K)}}(O)$ can be written as a term that does not depend on $\left(\psi^{(K)}, \overline{\gamma}_K^{(K)}\right)$ minus

$$\frac{H_K\left(\overline{A}_K, \overline{L}_K; \overline{\tau}_K\right)H_K\left(\overline{A}_K, \overline{L}_K; \overline{\tau}_K\right)'}{\prod_{s=1}^K \pi_s\left(\overline{A}_s, \overline{L}_s; \alpha_s\right)}\begin{pmatrix}\psi^{(K)}\\ \gamma_1^{(K)}\\ \vdots\\ \gamma_K^{(K)}\end{pmatrix}.$$

Therefore, the nonsingularity of $E_P\left\{\frac{H_K\left(\overline{A}_K,\overline{L}_K;\overline{\tau}_K^{\dagger}(P)\right)H_K\left(\overline{A}_K,\overline{L}_K;\overline{\tau}_K^{\dagger}(P)\right)'}{\prod_{s=1}^{K}\pi_s\left(\overline{A}_s,\overline{L}_s;\alpha_s^{\dagger}(P)\right)}\right\}$ implies that the equation in $\left(\psi^{(K)},\overline{\gamma}_K^{(K)}\right)$,

$$E_P\left\{\phi^{2K+1}_{\left(\overline{\alpha}_K^{\dagger}(P),\overline{\tau}_K^{\dagger}(P),\psi^{(K)},\overline{\gamma}_K^{(K)}\right)}\right\}=0,$$

has a unique solution that we denote indistinctly by $\theta_{2K+1}^{\dagger}(P)$ or $\left(\psi^{\dagger(K)}(P),\overline{\gamma}_K^{\dagger(K)}(P)\right)$. Here $\overline{\alpha}_K^{\dagger}(P)$ and $\overline{\tau}_K^{\dagger}(P)$ are the parameters defined in parts (i) and (ii) of Condition SPob respectively.

Also, for $k=K-1,\ldots,1$, $\phi^{3K+1-k}_{\left(\overline{\alpha}_K,\overline{\tau}_K,\psi^{(K)},\overline{\gamma}_K^{(K)},\ldots,\psi^{(k+1)},\overline{\gamma}_{k+1}^{(k+1)},\psi^{(k)},\overline{\gamma}_k^{(k)}\right)}(O)$ can be written as a term that does not depend on $\left(\psi^{(k)},\overline{\gamma}_k^{(k)}\right)$ minus

$$\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}}\frac{H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k\right)H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k\right)'}{\prod_{s=1}^{k}\pi_s\left(\overline{A}_s,\overline{L}_s;\alpha_s\right)}\begin{pmatrix}\psi^{(k)}\\ \gamma_1^{(k)}\\ \vdots\\ \gamma_k^{(k)}\end{pmatrix}.$$

Then, the nonsingularity of $E_P\left\{\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}}\frac{H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k^{\dagger}(P)\right)H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k^{\dagger}(P)\right)'}{\prod_{s=1}^{k}\pi_s\left(\overline{A}_s,\overline{L}_s;\alpha_s^{\dagger}(P)\right)}\right\}$ implies that the equation in $\left(\psi^{(k)},\overline{\gamma}_k^{(k)}\right)$,

$$E_P\left(\phi^{3K+1-k}_{\overline{\alpha}_K^{\dagger}(P),\overline{\tau}_K^{\dagger}(P),\psi^{\dagger(K)}(P),\overline{\gamma}_K^{\dagger(K)}(P),\ldots,\psi^{\dagger(k+1)}(P),\overline{\gamma}_{k+1}^{\dagger(k+1)}(P),\psi^{(k)},\overline{\gamma}_k^{(k)}}\right)=0,$$

has a unique solution, that we denote indistinctly by $\theta_{3K+1-k}^{\dagger}(P)$ or $\left(\psi^{\dagger(k)}(P),\overline{\gamma}_k^{\dagger(k)}(P)\right)$.

Finally, note that $\phi^{3K+1}_{\left(\overline{\alpha}_K,\overline{\tau}_K,\psi^{(K)},\overline{\gamma}_K^{(K)},\ldots,\psi^{(1)},\gamma_1^{(1)},\psi\right)}(O)$ can be written as a term that does not depend on ψ minus

$$\sum_{\underline{a}_1\in\underline{\mathcal{A}}_1}\mathbf{m}\left(\underline{a}_1,Z\right)\mathbf{m}\left(\underline{a}_1,Z\right)'\psi.$$

Hence, the nonsingularity of $E_P\left\{\sum_{\underline{a}_1\in\underline{\mathcal{A}}_1}\mathbf{m}\left(\underline{a}_1,Z\right)\mathbf{m}\left(\underline{a}_1,Z\right)'\right\}$ implies that the equation in ψ,

$$E_P\left(\phi^{3K+1}_{\overline{\alpha}_K^{\dagger}(P),\overline{\tau}_K^{\dagger}(P),\psi^{\dagger(K)}(P),\overline{\gamma}_K^{\dagger(K)}(P),\ldots,\psi^{\dagger(1)}(P),\gamma_1^{\dagger(1)}(P),\psi}\right)=0,$$

has a unique solution, which we denote indistinctly by $\theta_{3K+1}^{\dagger}(P)$ or $\psi^{\dagger}(P)$. This concludes the proof.

A.3.7 Proof of Lemma 6

Throughout this, we use the notation $\overline{\phi}^{2K}_{\overline{\theta}_{2K}}\equiv\left(\phi^1_{\theta_1},\ldots,\phi^K_{\theta_K},\phi^{K+1}_{\theta_{K+1}},\ldots,\phi^{2K}_{\theta_{2K}}\right)=\left(\phi^1_{\alpha_1},\ldots,\phi^K_{\alpha_K},\phi^{K+1}_{\tau_1},\ldots,\phi^{2K}_{\tau_K}\right)$. Note that, by Condition SLin, it suffices to show that, for each $k=0,\ldots,K$, the event $C_{n,k}$ occurs

with probability tending to one under P where $C_{n,k} \equiv$"the equation in θ_{3K+1-k}, $\mathbb{P}_n \left\{ \phi^{3K+1-k}_{\left(\widehat{\theta}_{3K+1-(k+1)}, \theta_{3K+1-k}\right)} \right\} = 0$, has a unique solution". To see that for $k = K$, note that the linearity of m and the $g'_k s$ implies that $\phi^{2K+1}_{\widehat{\theta}_{2K}, \theta_{2K+1}}(O) \left(= \phi^{2K+1}_{\widehat{\overline{\alpha}}_K, \widehat{\overline{\tau}}_K, \psi^{(K)}, \overline{\gamma}^{(K)}_K}(O) \right)$ can be written as a term that does not depend on $\left(\psi^{(K)}, \overline{\gamma}^{(K)}_K\right)$ minus

$$\frac{H_K\left(\overline{A}_K, \overline{L}_K; \widehat{\overline{\tau}}_K\right) H_K\left(\overline{A}_K, \overline{L}_K; \widehat{\overline{\tau}}_K\right)'}{\prod_{s=1}^K \pi_s\left(\overline{A}_s, \overline{L}_s; \widehat{\alpha}_s\right)} \begin{pmatrix} \psi^{(K)} \\ \gamma_1^{(K)} \\ \vdots \\ \gamma_K^{(K)} \end{pmatrix}.$$

On the other hand, note that: (1) by Condition SPobLin and Lemma 5, we have that $E_P\left\{\overline{\phi}^{2K}_{\overline{\theta}^{\dagger}_{2K}(P)}\right\} = 0$, and (2) assumptions (i) and (ii) of the lemma imply that $\overline{\phi}^{2K}_{\overline{\theta}_{2K}}(o)$ is twice continuously differentiable w.r.t. $\overline{\theta}_{2K}$. Then, assumptions (i) and (ii) of the lemma and Conditions Moment2, NonSing2, Domination and SPobLin imply that $\overline{\phi}^{2K}_{\overline{\theta}_{2K}}$ and $\overline{\theta}^{\dagger}_{2K}(P)$ verify the assumptions of q_β and β_0 in Proposition 3. Also, Condition SLin imply that the estimating equation $\mathbb{P}_n\left(\overline{\phi}^{2K}_{\overline{\theta}_{2K}}\right) = 0$ has at most one solution with probability tending to one under P. Hence, Proposition 3 and the fact that $\widehat{\overline{\theta}}_{2K} = \left(\widehat{\overline{\alpha}}_K, \widehat{\overline{\tau}}_K\right)$ solves that equation, imply that

$$\left(\widehat{\overline{\alpha}}_K, \widehat{\overline{\tau}}_K\right) \xrightarrow{p} \left(\overline{\alpha}^{\dagger}_K(P), \overline{\tau}^{\dagger}_K(P)\right)$$

under P. Then, since $\frac{H_K\left(\overline{a}_K, \overline{l}_K; \overline{\tau}_K\right) H'_K\left(\overline{a}_K, \overline{l}_K; \overline{\tau}_K\right)}{\prod_{s=1}^K \pi_s\left(\overline{a}_s, \overline{l}_s; \alpha_s\right)}$ is regular by Condition R, Lemma 14 in Appendix A.4 implies that

$$\mathbb{P}_n \left\{ \frac{H\left(\overline{A}_K, \overline{L}_K; \widehat{\overline{\tau}}_K\right) H_K\left(\overline{A}_K, \overline{L}_K; \widehat{\overline{\tau}}_K\right)'}{\prod_{s=1}^K \pi_s\left(\overline{A}_s, \overline{L}_s; \widehat{\alpha}_s\right)} \right\} \xrightarrow{p} E_P \left\{ \frac{H_K\left(\overline{A}_K, \overline{L}_K; \overline{\tau}^{\dagger}_K(P)\right) H_K\left(\overline{A}_K, \overline{L}_K; \overline{\tau}^{\dagger}_K(P)\right)'}{\prod_{s=1}^K \pi_s\left(\overline{A}_s, \overline{L}_s; \alpha^{\dagger}_s(P)\right)} \right\}$$

under P. But the expectation in the right hand side of last display is nonsingular by Condition SPobLin. Then, $\mathbb{P}_n \left\{ \frac{H\left(\overline{A}_K, \overline{L}_K; \widehat{\overline{\tau}}_K\right) H_K\left(\overline{A}_K, \overline{L}_K; \widehat{\overline{\tau}}_K\right)'}{\prod_{s=1}^K \pi_s\left(\overline{A}_s, \overline{L}_s; \widehat{\alpha}_l\right)} \right\}$ is nonsingular with probability tending to one under P and, therefore, the equation $\mathbb{P}_n\left(\phi^{2K+1}_{\widehat{\theta}_{2K}, \theta_{2K+1}}\right) = 0$ has a unique solution with probability tending to one under P, that is $C_{n,K}$ occurs with probability tending to one under P.

Analogously, for $k = 1, \ldots, K-1$, $\phi^{3K+1-k}_{\widehat{\theta}_{3K-k}, \theta_{3K+1-k}}(O) \left(= \phi^{3K+1-k}_{\widehat{\overline{\alpha}}_K, \widehat{\overline{\tau}}_K, \widehat{\psi}^{(K)}, \widehat{\overline{\gamma}}^{(K)}_K, \ldots, \widehat{\psi}^{(k+1)}, \widehat{\overline{\gamma}}^{(k+1)}_{k+1}, \psi^{(k)}, \overline{\gamma}^{(k)}_k}(O) \right)$ can be written as a term that does not depend on $\theta_{3K+1-k} = \left(\psi^{(k)}, \overline{\gamma}^{(k)}_k\right)$ minus

$$\sum_{\underline{a}_{k+1} \in \underline{\mathcal{A}}_{k+1}} \frac{H_k\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \widehat{\overline{\tau}}_k\right) H_k\left(\overline{A}_k, \underline{a}_{k+1}, \overline{L}_k; \widehat{\overline{\tau}}_k\right)'}{\prod_{s=1}^k \pi_s\left(\overline{A}_s, \overline{L}_s; \widehat{\alpha}_s\right)} \begin{pmatrix} \psi^{(k)} \\ \gamma_1^{(k)} \\ \vdots \\ \gamma_k^{(k)} \end{pmatrix},$$

Also, Lemma 14 in Appendix A.4, the convergence in probability of $\left(\widehat{\overline{\alpha}}_k, \widehat{\overline{\tau}}_k\right)$ to $\left(\overline{\alpha}_k^{\dagger}(P), \overline{\tau}_k^{\dagger}(P)\right)$ under P, the regularity of $\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}} \frac{H_k\left(\overline{a}_k,\underline{a}_{k+1},\overline{l}_k;\overline{\tau}_k\right)H_k'\left(\overline{a}_k,\underline{a}_{k+1},\overline{l}_k;\overline{\tau}_k\right)}{\prod_{s=1}^{k}\pi_s\left(\overline{a}_s,\overline{l}_s;\alpha_s\right)}$ and the nonsingularity of $E_P\left\{\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}} \frac{H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k^{\dagger}(P)\right)H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\overline{\tau}_k^{\dagger}(P)\right)'}{\prod_{s=1}^{k}\pi_s\left(\overline{A}_s,\overline{L}_s;\alpha_s^{\dagger}(P)\right)}\right\}$ imply that $\mathbb{P}_n\left\{\sum_{\underline{a}_{k+1}\in\underline{\mathcal{A}}_{k+1}} \frac{H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\widehat{\overline{\tau}}_k\right)H_k\left(\overline{A}_k,\underline{a}_{k+1},\overline{L}_k;\widehat{\overline{\tau}}_k\right)'}{\prod_{s=1}^{k}\pi_s\left(\overline{A}_s,\overline{L}_s;\widehat{\overline{\alpha}}_s\right)}\right\}$ is nonsingular with probability tending to one under P. Hence, the equation $\mathbb{P}_n\left(\phi^{3K+1-k}_{\widehat{\widetilde{\theta}}_{3K-k},\theta_{3K+1-k}}\right) = 0$ has a unique solution with probability tending to one under P, that is, for $k = 1,\ldots,K-1, C_{n,k}$ occurs with probability tending to one under P.

Finally, $\phi^{3K+1}_{\widehat{\widetilde{\theta}}_{3K},\theta_{3K+1}}(O)\left(= \phi^{3K+1}_{\overline{\alpha}_K,\overline{\tau}_K,\psi^{(K)},\overline{\gamma}_K^{(K)},\ldots,\psi^{(1)},\gamma_1^{(1)},\psi}(O)\right)$ can be written as a term that does not depend on ψ minus

$$\sum_{\underline{a}_1\in\underline{\mathcal{A}}_1} \mathbf{m}\left(\underline{a}_1, Z\right)\mathbf{m}\left(\underline{a}_1, Z\right)'\psi.$$

Furthermore, the nonsingularity of $E_P\left\{\sum_{\underline{a}_1\in\underline{\mathcal{A}}_1} \mathbf{m}\left(\underline{a}_1, Z\right)\mathbf{m}\left(\underline{a}_1, Z\right)'\right\}$ imply that $\mathbb{P}_n\left\{\sum_{\underline{a}_1\in\underline{\mathcal{A}}_1} \mathbf{m}\left(\underline{a}_1, Z\right)\mathbf{m}\left(\underline{a}_1, Z\right)'\right\}$ is nonsingular with probability tending to one under P and, hence, that the equation $\mathbb{P}_n\left(\phi^{3K+1}_{\widehat{\widetilde{\theta}}_{3K},\theta_{3K+1}}\right) = 0$ has a unique solution with probability tending to one under P, that is $C_{n,0}$ occurs with probability tending to one under P. This concludes the proof.

A.4 Technical results for Section 1.10

In this appendix, for completeness, we present some results used in Section 1.10, despite being known in the literature. Throughout this appendix, $\mathcal{B}$ denotes a subset of a Euclidean space with non-empty interior and X denotes an $m \times 1$ random vector with law G in R^m. We start by introducing the definition of *regular* function.

Definition 2 *Given a subset $\widetilde{\mathcal{B}} \subseteq \mathcal{B}$ with non-empty interior, a function $\zeta : R^m \times \mathcal{B} \to R^{p \times q}$ is said to be regular in $\widetilde{\mathcal{B}}$ if*

(i) $\zeta(x, \beta)$ is measurable w.r.t. x for each $\beta \in \widetilde{\mathcal{B}}$,

(ii) ζ is dominated in $\widetilde{\mathcal{B}}$, in the sense that there exists a function $b : R^m \to R$ such that $\|\zeta(x, \beta)\| \leq b(x)$ for every $(x, \beta) \in R^m \times \widetilde{\mathcal{B}}$ and $E_G\{b(X)\} < \infty$, and

(iii) ζ is almost sure continuos in $\widetilde{\mathcal{B}}$, in the sense that, for each fixed $\beta \in \widetilde{\mathcal{B}}$, the event $\{\lim_{\gamma \to \beta} \zeta(X, \gamma) = \zeta(X, \beta)\}$ has probability 1 (dG).

The measurability and domination assumptions (i) and (ii) ensure that the expectation

$$\Psi(\beta) \equiv E\{\zeta(X, \beta)\}$$

exists for every $\beta \in \widetilde{\mathcal{B}}$, while the almost sure continuity assumption (iii) implies, by dominated convergence, that Ψ is a continuous function of β in $\widetilde{\mathcal{B}}$.

The following lemma, proved by Tauchen [52], states that if ζ is regular in a compact subset of $\mathcal{B}$, then the sample average $\mathbb{P}_n\{\zeta(X, \beta)\}$ converges uniformly almost surely to its expectation $\Psi(\beta)$ in that subset. Although Tauchen proved this result for vector functions, the argument leading the result for matrices functions is entirely analogous.

Lemma 13 (from Lemma 1 of Tauchen, 1985) *Let $\widetilde{\mathcal{B}}$ be a compact subset of $\mathcal{B}$ with non-empty interior and let $\zeta : R^m \times \mathcal{B} \to R^{p \times q}$ be a regular function in $\widetilde{\mathcal{B}}$. Then, $E\{\zeta(X, \beta)\}$ is continuous in β and*

$$\sup_{\beta \in \widetilde{\mathcal{B}}} \|\mathbb{P}_n\{\zeta(X, \beta)\} - E\{\zeta(X, \beta)\}\| \xrightarrow{a.s.} 0.$$

The following lemma is an immediate corollary of the previous one.

Lemma 14 *Let β_0 be a point in the interior of $\mathcal{B}$. If $\zeta : R^m \times \mathcal{B} \to R^{p \times q}$ is regular in some neighborhood of β_0 then*

$$\mathbb{P}_n\left\{\zeta\left(X, \widehat{\beta}_n\right)\right\} \xrightarrow{p} E\{\zeta(X, \beta_0)\}$$

for every sequence $\widehat{\beta}_n$ that converges in probability to β_0.

Appendix B

Appendix of Chapter 2

B.1 Examples of counterfactual contrasts that correspond to a g-formula

In this appendix we provide several examples of parameters of interest in causal inference and missing data analysis that correspond to a g-formula.

B.1.1 Example 1.

Mean of an outcome in a longitudinal study with ignorable drop-out. Consider a longitudinal study with drop-outs. Define L_k to be the data vector L_k^* that is recorded on a subject randomly selected from a target population if the subject is still on study at the k^{th} study cycle and to be equal to an arbitrary vector in $\mathcal{L}_k$, say $\varkappa_k$, otherwise. Assume no subject misses the first cycle. Then $L_1 = L_1^*$. Let $A_k = 1$ if the subject is on study at the $(k+1)^{th}$ study cycle and $A_k = 0$ otherwise. Thus, $L_k = A_{k-1}L_k^* + (1 - A_{k-1})\varkappa_k$. Let $p = \prod_{j=0}^{K} g_j \prod_{j=1}^{K} h_j$ be the law of $\left(\overline{A}_K, \overline{L}_{K+1}\right)$. Under the missing at random assumption that $L_{K+1}^* \perp\!\!\!\perp A_k \mid \left(\overline{A}_{k-1} = \overline{1}, \overline{L}_k\right)$ for each $k = 1, \ldots, K$, and the positivity assumption that for all $k = 1, \ldots, K$, $\Pr\left\{h_k\left(1|\overline{A}_{k-1} = 1, \overline{L}_k\right) > 0\right\} = 1$, the mean of the, potentially missing, last cycle outcome L_{K+1}^*, i.e., of the outcome that would be recorded if the study did not suffer from drop-out, equals

$$E_{g_0}\left[E_{g_1}[\ldots E_{g_{K-1}}\left\{E_{g_K}\left(L_{K+1}|\overline{A}_K = \overline{1}, \overline{L}_K\right)\middle|\, \overline{A}_{K-1} = \overline{1}, \overline{L}_{K-1}\right\}\ldots|A_1 = 1, L_1]\right]. \tag{B.1}$$

The expression in (B.1) agrees with $\theta(p)$ if we take $h_k^*\left(a_k|\bar{l}_k, \overline{a}_{k-1}\right) = a_k$ and $\kappa\left(\bar{l}_{K+1}\right) = l_{K+1}$ ([29], [43]). Note that the positivity assumption is the same as the assumption that $gh^* << gh$. Note also that because A_k is a binary variable, $\theta(p)$ actually involves only integrals over $l_1, \ldots, l_{K+1}$ as, for each k, the integral over a_k is indeed a sum with a single non-zero term.

B.1.2 Example 2.

Outcome mean under a sequence of fixed treatments. Suppose that in a longitudinal study L_k denotes the vector of variables measured at the k^{th} study cycle on a subject randomly selected from a target population. Assume that immediately after recording L_k the subject decides which of the available treatments in a set $\mathcal{A}_k$ he will take until the next study cycle. Let $A_k \in \mathcal{A}_k$ denote the subject's treatment choice. Let $p = \prod_{j=0}^{K} g_j \prod_{j=1}^{K} h_j$ be the law of $\left(\overline{A}_K, \overline{L}_{K+1}\right)$. Also, let $L_{K+1,\overline{a}_K^*}$ be the counterfactual outcome at the end of follow-up if, possibly contrary to fact, the subject took treatment $\overline{A}_K = \overline{a}_K^*$ for some fixed $\overline{a}_K^* = (a_1^*, \ldots, a_K^*)$. Contrasts of the mean of $L_{K+1,\overline{a}_K^*}$ involving different $\overline{a}_K^*$ quantify treatment effects. For instance, the average treatment effect (ATE) comparing the *always on treatment* vs *never on treatment* regimes is defined as the mean of $L_{K+1,\overline{1}}$ minus the mean of $L_{K+1,\overline{0}}$. Under the consistency assumption that $\overline{A}_K = \overline{a}_K^* \Rightarrow L_{K+1} = L_{K+1,\overline{a}_K^*}$, the no-unmeasured confounding assumption that for $k = 1, \ldots, K$, $L_{K+1,\overline{a}_K^*} \perp\!\!\!\perp A_k \mid \left(\overline{A}_{k-1} = \overline{a}_{k-1}^*, \overline{L}_k\right)$ and the positivity assumption that for $k = 1, \ldots, K$, $\Pr\left\{h_k\left(a_k^* | \overline{a}_{k-1}^*, \overline{L}_k\right) > 0\right\} = 1$, the mean of $L_{K+1,\overline{a}_K^*}$ equals ([29])

$$E_{g_0}\left[E_{g_1}\left[\ldots E_{g_{K-1}}\left\{E_{g_K}\left(L_{K+1} | \overline{A}_K = \overline{a}_K^*, \overline{L}_K\right) \middle| \overline{A}_{K-1} = \overline{a}_{K-1}^*, \overline{L}_{K-1}\right\} \ldots | A_1 = a_1^*, L_1\right]\right]. \quad \text{(B.2)}$$

This expression agrees with $\theta(p)$ if we take $h_k^*\left(a_k | \bar{l}_k, \overline{a}_{k-1}\right) = I_{\left\{a_k^*\right\}}(a_k)$ and $\kappa\left(\bar{l}_{K+1}\right) = l_{K+1}$ where throughout, $I_D(x) = 1$ if $x \in D$ and $I_D(x) = 0$ otherwise. Note that in Example 1 we could arrive at the formula (B.1) from the formula (B.2) if, in that example we regard A_k as a sequence of time-dependent treatments indexed by k and consider estimation of the mean of L_{K+1} had, contrary to fact, all subjects followed the treatment regime specified by $a_k = 1$ for $k = 1, \ldots, K$; that is, the regime in which no subject had dropped-out. Robins ([29], p. 1491; 1987a, sec. AD.5) provided additional discussion of the usefulness of regarding missing data indicators as time-dependent treatments.

B.1.3 Example 3.

Outcome mean under a non-random dynamic treatment regime. Assume that the recorded data O are as in the longitudinal study of Example 2. However, suppose that we are now interested in estimating the mean of L_{K+1} if, contrary to fact, the entire study population followed a given *non-random dynamic* treatment regime which stipulates that right after study cycle k and until just prior to study cycle $k+1$, a patient with covariate and treatment history $\left(\overline{a}_{k-1}, \bar{l}_k\right)$ receives treatment $A_k = d_k\left(\overline{a}_{k-1}, \bar{l}_k\right)$. Similarly to Example 2, the average treatment effect for comparing the two such regimes, say d and d', is defined as the mean of $L_{K+1,d}$ minus the mean of $L_{K+1,d'}$ where for any treatment regime $d = \{d_1, \ldots, d_K\}$, $L_{K+1,d}$ denotes the counterfactual outcome at the end of the study if, possibly contrary to fact, the subject had followed treatment regime d. Under the consistency assumption that $\overline{A}_K = \overline{D}_K \Rightarrow L_{K+1} = L_{K+1,d}$, where for any $j = 1, \ldots, K$, $D_j \equiv d_j\left(\overline{A}_{j-1}, \overline{L}_j\right)$, the no-unmeasured confounding assumption that for $k = 1, \ldots, K$, $L_{K+1,d} \perp\!\!\!\perp A_k \mid \left(\overline{A}_{k-1} = \overline{D}_{k-1}, \overline{L}_k\right)$, and the positivity assumption that for $k = 1, \ldots, K$, $\Pr\left[\Pr\left(A_k = D_k | \overline{A}_{k-1} = \overline{D}_{k-1}, \overline{L}_k\right) > 0\right] = 1$, the mean of $L_{K+1,d}$ is

$$E_{g_0}\left[E_{g_1}\left[\ldots E_{g_{K-1}}\left\{E_{g_K}\left(L_{K+1} | \overline{A}_K = \overline{D}_K, \overline{L}_K\right) \middle| \overline{A}_{K-1} = \overline{D}_{K-1}, \overline{L}_{K-1}\right\} \ldots | A_1 = D_1, L_1\right]\right].$$

This expression agrees with $\theta(p)$ if we take $h_k^*\left(a_k|\bar{l}_k, \overline{a}_{k-1}\right) = I_{\left\{d_k\left(\bar{l}_k, \overline{a}_{k-1}\right)\right\}}(a_k)$ and $\kappa\left(\bar{l}_{K+1}\right) = l_{K+1}$. Note also that the positivity assumption is the same as the assumption that $gh^* << gh$.

B.1.4 Example 4.

Outcome mean under a random dynamic treatment regime. Assume that the recorded data O are as in the longitudinal study of Example 2. Suppose that we are now interested in estimating the mean of L_{K+1} if, contrary to fact, the entire population followed a *random* dynamic treatment regime which stipulates that at study cycle k a patient with covariate and treatment history $\left(\overline{a}_{k-1}, \bar{l}_k\right)$ is randomized to receive treatment $A_k = a_k$ with probability $h_k^*\left(a_k|\overline{a}_{k-1}, \bar{l}_k\right)$ where a_k is in the set $\mathcal{A}_k$ of treatments available at time k. Similarly to Example 2, the average treatment effect for comparing the two regimes, determined by h^* and h^{**}, is defined as the mean of L_{K+1,h^*} minus the mean of $L_{K+1,h^{**}}$ where for any $h^* \equiv \left\{h_k^* : k = 1, \ldots, K\right\}$, L_{K+1,h^*} denotes the counterfactual outcome if, possibly contrary to fact, the subject had followed the random treatment regime h^*. Under the consistency assumption that $\overline{A}_K = \overline{A}_{h^*,K} \Rightarrow L_{K+1} = L_{K+1,h^*}$ for all $k = 1, \ldots, K$ where $A_{h^*,k}$ is the treatment received at cycle k when the subject follows the random regime, the no-unmeasured confounding assumption that $L_{K+1,h^*} \perp\!\!\!\perp A_k \mid \left(\overline{A}_{k-1} = \overline{a}_{k-1}, \overline{L}_k = \bar{l}_k\right)$ for all $\left(\overline{a}_{k-1}, \bar{l}_k\right)$ such that $\prod_{j=1}^{k-1} h_j^*\left(a_j|\overline{a}_{j-1}, \bar{l}_j\right) > 0$ and the positivity assumption that $gh^* << gh$, the mean of L_{K+1,h^*} is precisely equal to $\theta(p)$ if we take $\kappa\left(\bar{l}_{K+1}\right) = l_{K+1}$.

B.2 Proof of Lemmas 7 and 8

B.2.1 Proof of Lemma 7

The identity (2.15) coincides with (2.16) for $j=0$ if $\overline{A}_j$ and $\overline{L}_j$ are defined as nill when $j=0$ and $\eta_j\left(\overline{L}_j\right)$ is defined as $\theta\left(\eta\right)$ if $j=0$. It thus suffices to show (2.16) for an arbitrary $j\in\{0,...,K\}$. We prove it by reverse induction. For $j=K$ the result holds by definition of $\eta_K\left(\overline{L}_K\right)$ since $Q_{K+1}\left(\overline{h}_{K+1}^{\dagger K},\overline{\eta}_{K+1}^{\dagger K}\right)=\kappa\left(\overline{L}_{K+1}\right)$. Suppose now that (2.16) holds for a given $j\in[K]$, we want to show that it also holds for $j-1$. Now,

$$
\begin{aligned}
&E_p\left\{Q_j\left(\underline{h}_j^{\dagger},\underline{\eta}_j^{\dagger}\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}-\eta_{j-1}\left(\overline{L}_{j-1}\right)=\\
&=E_p\left[\frac{I_j}{h_j^{\dagger}}\left\{Q_{j+1}\left(\underline{h}_{j+1}^{\dagger},\underline{\eta}_{j+1}^{\dagger}\right)-\eta_j^{\dagger}\right\}+\eta_j^{\dagger}\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right]-\eta_{j-1}\left(\overline{L}_{j-1}\right)\\
&=E_p\left[\frac{I_j}{h_j^{\dagger}}E_p\left\{Q_{j+1}\left(\underline{h}_{j+1}^{\dagger},\underline{\eta}_{j+1}^{\dagger}\right)\middle|\overline{A}_j=\overline{a}_j^{-*},\overline{L}_j\right\}\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right]\\
&-E_p\left\{\frac{I_j}{h_j^{\dagger}}\eta_j^{\dagger}\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}\\
&+E_p\left\{\eta_j^{\dagger}\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}-\eta_{j-1}\left(\overline{L}_{j-1}\right)\\
&=E_p\left\{\frac{I_j}{h_j^{\dagger}}\left[\sum_{k=j+1}^{K}E_p\left\{\frac{\overline{I}_{j+1}^{(k-1)}}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^{\dagger}}\right)\left(\eta_k^{\dagger}-\eta_k\right)\middle|\overline{A}_j=\overline{a}_j^{-*},\overline{L}_j\right\}+\eta_j\right]\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}\\
&-E_p\left(\frac{I_j}{h_j^{\dagger}}\eta_j^{\dagger}\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right)+E_p\left\{\eta_j^{\dagger}\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}-\eta_{j-1}\left(\overline{L}_{j-1}\right)\\
&=\sum_{k=j+1}^{K}E_p\left\{\frac{\overline{I}_j^{(k-1)}}{\pi_j^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^{\dagger}}\right)\left(\eta_k^{\dagger}-\eta_k\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}+E_p\left(\frac{I_j}{h_j^{\dagger}}\eta_j\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right)\\
&+E_p\left(\left(1-\frac{I_j}{h_j^{\dagger}}\right)\eta_j^{\dagger}\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right)-E_p\left(\eta_j|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right)\\
&=\sum_{k=j+1}^{K}E_p\left\{\frac{\overline{I}_j^{(k-1)}}{\pi_j^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^{\dagger}}\right)\left(\eta_k^{\dagger}-\eta_k\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}\\
&+E_p\left\{\left(1-\frac{I_j}{h_j^{\dagger}}\right)\left(\eta_j^{\dagger}-\eta_j\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}\\
&=\sum_{k=j+1}^{K}E_p\left\{\frac{\overline{I}_j^{(k-1)}}{\pi_j^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^{\dagger}}\right)\left(\eta_k^{\dagger}-\eta_k\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\}\\
&+E_p\left\{\left(\frac{I_j}{h_j}-\frac{I_j}{h_j^{\dagger}}\right)\left(\eta_j^{\dagger}-\eta_j\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^{*},\overline{L}_{j-1}\right\},
\end{aligned}
$$

where the third equality is by the inductive hypothesis and the last one follows from the fact that $E_p\left\{u\left(\overline{L}_j\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^*,\overline{L}_{j-1}\right\}=E_p\left\{\frac{I_j}{h_j}u\left(\overline{L}_j\right)\middle|\overline{A}_{j-1}=\overline{a}_{j-1}^*,\overline{L}_{j-1}\right\}$ for any $u\left(\cdot\right)$. This concludes the proof.

B.2.2 Proof of Lemma 8

The proof of Lemma 8 invokes the following lemma.

Lemma 15 *For $k\in[K]$ and for arbitrary $\eta_j^\dagger, j\in[K]$, it holds that*

$$E_{g_k}\left\{\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}-\eta_k\left(\overline{L}_k\right)=\sum_{j=k+1}^{K}E_{\underline{g}_k,\underline{h}_{k+1}}\left\{\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}$$

with

$$\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\equiv\eta_j^\dagger\left(\overline{L}_j\right)-E_{g_j}\left\{\eta_{j+1}^\dagger\left(\overline{L}_{j+1}\right)\middle|\overline{A}_j=\overline{a}_j^*,\overline{L}_j\right\}.$$

Proof of Lemma 15. Given $k\in[K]$ and $j\geq k+1$,

$$\begin{aligned}
&E_p\left\{\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}\\
&=E_p\left[\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\left\{\eta_j^\dagger-E_{g_j}\left(\eta_{j+1}^\dagger\middle|\overline{A}_j=\overline{a}_j^*,\overline{L}_j\right)\right\}\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right]\\
&=E_p\left[\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\left\{\eta_j^\dagger-E_{g_j}\left(\eta_{j+1}^\dagger\middle|\overline{A}_k=\overline{a}_k^*,\overline{A}_{k+1}^j,\overline{L}_j\right)\right\}\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right]\\
&=E_p\left[\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\left\{\eta_j^\dagger-\eta_{j+1}^\dagger\right\}\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right]\\
&=E_p\left(\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\eta_j^\dagger\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right)-E_p\left(\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\eta_{j+1}^\dagger\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right).
\end{aligned}$$

But, we prove bellow that, for $k\in[K]$, $j\geq k$ and any function $u\left(\bar{l}_{j+1}\right)$, it holds that

$$E\left\{\overline{I}_{k+1}^j u\left(\overline{L}_{j+1}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}=E\left\{\frac{\overline{I}_{k+1}^{j+1}}{h_{j+1}}u\left(\overline{L}_{j+1}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}. \tag{B.3}$$

Thus, invoking (B.3) with $u\left(\overline{L}_{j+1}\right)$ replaced by $\frac{\eta_{j+1}^\dagger\left(\overline{L}_{j+1}\right)}{\pi_{k+1}^j\left(\overline{L}_{j+1}\right)}$, we obtain

$$E_p\left(\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\eta_{j+1}^\dagger\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right)=E_p\left(\frac{\overline{I}_{k+1}^{j+1}}{\pi_{k+1}^{j+1}}\eta_{j+1}^\dagger\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right).$$

Hence, for $j\geq k+1$,

$$E_p\left\{\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}=E_p\left(\frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j}\eta_j^\dagger\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right)-E_p\left(\frac{\overline{I}_{k+1}^{j+1}}{\pi_{k+1}^{j+1}}\eta_{j+1}^\dagger\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right)$$

and, therefore,

$$\begin{aligned}
&\sum_{j=k+1}^{K} E_p\left\{\frac{\overline{I}_{k+1}^{j}}{\pi_{k+1}^{j}}\delta_j\left(\eta_j^{\dagger},\eta_{j+1}^{\dagger};g_j\right)\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right\}\\
&=E_p\left(\frac{I_{k+1}}{h_{k+1}}\eta_{k+1}^{\dagger}\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right)-E_p\left(\frac{\overline{I}_{k+1}^{K}}{\pi_{k+1}^{K}}\eta_{K+1}^{\dagger}\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right)\\
&=E_p\left(\frac{I_{k+1}}{h_{k+1}}\eta_{k+1}^{\dagger}\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right)-E_p\left(\frac{\overline{I}_{k+1}^{K}}{\pi_{k+1}^{K}}\kappa\left(\overline{L}_{K+1}\right)\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right).
\end{aligned}$$

Now, if in (B.3) we set j at k, we arrive at

$$E_p\left\{u\left(\overline{L}_{k+1}\right)\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right\}=E_p\left\{\frac{I_{k+1}}{h_{k+1}}u\left(\overline{L}_{k+1}\right)\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right\}. \tag{B.4}$$

Then, invoking (B.4) with $u\left(\overline{L}_{k+1}\right)$ replaced by $\eta_{k+1}^{\dagger}\left(\overline{L}_{k+1}\right)$, we arrive at

$$E_p\left(\frac{I_{k+1}}{h_{k+1}}\eta_{k+1}^{\dagger}\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right)=E_p\left(\eta_{k+1}^{\dagger}\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right).$$

Then, we would arrive at the desired result if we show that

$$\eta_k\left(\overline{L}_k\right)=E_p\left(\frac{\overline{I}_{k+1}^{K}}{\pi_{k+1}^{K}}\kappa\left(\overline{L}_{K+1}\right)\middle|\,\overline{A}_k=\overline{a}_k^{*},\overline{L}_k\right) \tag{B.5}$$

for all $k\in[K]$. We prove (B.5) by reverse induction. For $k=K$, (B.5) is verified since, by definition

$$\begin{aligned}
\eta_K\left(\overline{L}_K\right)&=E_p\left\{\kappa\left(\overline{L}_{K+1}\right)\middle|\,\overline{A}_K=\overline{a}_K^{*},\overline{L}_K\right\}\\
&=E_p\left\{\frac{\overline{I}_{K+1}^{K}}{\pi_{K+1}^{K}}\kappa\left(\overline{L}_{K+1}\right)\middle|\,\overline{A}_K=\overline{a}_K^{*},\overline{L}_K\right\}.
\end{aligned}$$

Suppose ($B.5$) holds for $k = j + 1$. We will show that it holds for $k = j$.

$$
\begin{aligned}
\eta_j\left(\overline{L}_j\right) &= E_p\left\{ \eta_{j+1}\left(\overline{L}_{j+1}\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= E_p\left\{ E_p\left(\frac{\overline{I}_{j+2}^K}{\pi_{j+2}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_{j+1} = \overline{a}_{j+1}^*, \overline{L}_{j+1} \right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= E_p\left\{ \frac{I_{j+1}}{h_{j+1}} E_p\left(\frac{\overline{I}_{j+2}^K}{\pi_{j+2}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_{j+1} = \overline{a}_{j+1}^*, \overline{L}_{j+1} \right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= E_p\left\{ \frac{I_{j+1}}{h_{j+1}} E_p\left(\frac{\overline{I}_{j+2}^K}{\pi_{j+2}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_{j+1} = \overline{a}_{j+1}^*, \overline{L}_{j+1} \right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= E_p\left\{ \frac{I_{j+1}}{h_{j+1}} E_p\left(\frac{\overline{I}_{j+2}^K}{\pi_{j+2}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{A}_{j+1}, \overline{L}_{j+1} \right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= E_p\left\{ E_p\left(\frac{I_{j+1}}{h_{j+1}} \frac{\overline{I}_{j+2}^K}{\pi_{j+2}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{A}_{j+1}, \overline{L}_{j+1} \right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= E_p\left\{ E_p\left(\frac{\overline{I}_{j+1}^K}{\pi_{j+1}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{A}_{j+1}, \overline{L}_{j+1} \right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= E_p\left\{ \frac{\overline{I}_{j+1}^K}{\pi_{j+1}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\}.
\end{aligned}
$$

Here, the second equivalence follows applying ($B.4$) with k replaced by j and $u\left(\overline{L}_{j+1}\right)$ replaced by $E_p\left(\frac{\overline{I}_{j+2}^K}{\pi_{j+2}^K} \kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_{j+1} = \overline{a}_{j+1}^*, \overline{L}_{j+1} \right)$.

Now, we show ($B.3$). Let $k \in [K]$ and $j \geq k$,

$$
\begin{aligned}
& E_p\left\{ \frac{\overline{I}_{k+1}^{j+1}}{h_{j+1}} u\left(\overline{L}_{j+1}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\} \\
&= E_p\left[E_p\left\{ \frac{\overline{I}_{k+1}^{j+1}}{h_{j+1}} u\left(\overline{L}_{j+1}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{A}_{k+1}^j, \overline{L}_{j+1} \right\} \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right] \\
&= E_p\left\{ \frac{\overline{I}_{k+1}^j}{h_{j+1}} u\left(\overline{L}_{j+1}\right) E_p\left(I_{j+1} \middle| \overline{A}_k = \overline{a}_k^*, \overline{A}_{k+1}^j, \overline{L}_{j+1} \right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\} \\
&= E_p\left\{ \frac{\overline{I}_{k+1}^j}{h_{j+1}} u\left(\overline{L}_{j+1}\right) E_p\left(I_{j+1} \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_{j+1} \right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\} \\
&= E_p\left\{ \overline{I}_{k+1}^j u\left(\overline{L}_{j+1}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\},
\end{aligned}
$$

since $h_{j+1}\left(\overline{L}_{j+1}\right) = P\left(A_{j+1} = a_{j+1}^* \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_{j+1}\right) = E_p\left(I_{j+1} \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_{j+1}\right)$. We thus conclude the proof. ■

Proof of Lemma 8. We prove by reverse induction that for $k \in \{0, 1..., K\}$,

$$\overline{I}_k \left(\eta_k^\dagger - \eta_k \right) = \Gamma_k + \overline{I}_k \sum_{s=k+1}^{K} E_p \left\{ \frac{1}{\pi_{k+1}^{s-1}} \left(\frac{1}{h_s} - \frac{1}{h_s^\dagger} \right) \Gamma_s \middle| \overline{A}_k, \overline{L}_k \right\} \tag{B.6}$$

with $I_0 \equiv 1, \eta_0 \equiv \theta(\eta)$, $\left(\overline{A}_0, \overline{L}_0\right) \equiv$ nill, $\sum_{s=K+1}^{K}(\cdot) \equiv 0$ and $\Gamma_0 = \Gamma_0\left(\overline{h}_1^{\dagger K}, \overline{\eta}_0^{\dagger K}; g, h\right) \equiv \eta_0^\dagger - E_{g,h}\left\{Q_1\left(\overline{h}_1^{\dagger K}, \overline{\eta}_1^{\dagger K}\right)\right\}$, where to simplify notation we use the shortcut $\Gamma_k \equiv \Gamma_k\left(\underline{h}_{k+1}^\dagger, \underline{\eta}_k^\dagger; \underline{g}_k, \underline{h}_{k+1}\right)$.

Applying this equality to $k = 0$ with

$$\eta_0^\dagger = E_p\left\{Q_1\left(\overline{h}_1^{\dagger K}, \overline{\eta}_1^{\dagger K}\right)\right\} \tag{B.7}$$

we obtain that

$$I_0\left[E_p\left\{Q_1\left(\overline{h}_1^{\dagger K}, \overline{\eta}_1^{\dagger K}\right)\right\} - \eta_0\right] = \Gamma_0 + I_0 \sum_{s=1}^{K} E_p\left\{\frac{1}{\pi_1^{s-1}}\left(\frac{1}{h_s} - \frac{1}{h_s^\dagger}\right)\Gamma_s \middle| \overline{A}_0 = \overline{a}_0^*, \overline{L}_0\right\}.$$

Recalling that $I_0 = 1, \eta_0 = \theta(\eta)$, $\left(\overline{A}_0, \overline{L}_0\right) =$ nill and that, with $\eta_0^\dagger$ defined as in (B.7), $\Gamma_0 \equiv \eta_0^\dagger - E_p\left\{Q_1\left(\overline{h}_1^{\dagger K}, \overline{\eta}_1^{\dagger K}\right)\right\} = 0$, we conclude that

$$E_p\left\{Q_1\left(\overline{h}_1^{\dagger K}, \overline{\eta}_1^{\dagger K}\right)\right\} - \theta(\eta) = \sum_{s=1}^{K} E_p\left\{\frac{1}{\pi^{s-1}}\left(\frac{1}{h_s} - \frac{1}{h_s^\dagger}\right)\Gamma_s\right\} \equiv b^p\left(h^\dagger, \eta^\dagger\right)$$

which, invoking Lemma 7, proves that $b^p\left(h^\dagger, \eta^\dagger\right) = a^p\left(h^\dagger, \eta^\dagger\right)$.

We now prove identity (B.6) by induction. For $k = K$, (B.6) holds because by definition

$$\begin{aligned}
\overline{I}_K\left(\eta_K^\dagger - \eta_K\right) &= \overline{I}_K\left[\eta_K^\dagger - E_{g_K}\left\{\kappa\left(\overline{L}_{K+1}\right) \middle| \overline{A}_K = \overline{a}_K^*, \overline{L}_K\right\}\right] \\
&= \overline{I}_K\left[\eta_K^\dagger - E_{g_K}\left\{Q_{K+1}\left(\overline{h}_{K+1}^{\dagger K}, \overline{\eta}_{K+1}^{\dagger K}\right) \middle| \overline{A}_K = \overline{a}_K^*, \overline{L}_K\right\}\right] \\
&= \Gamma_K.
\end{aligned}$$

Suppose (B.6) holds for $k = K, ..., j+1$. We will show that it holds for $k = j$. By Lemma 7 we have

$$\begin{aligned}
\overline{I}_j\left(\eta_j^\dagger - \eta_j\right) &= \overline{I}_j\left[\eta_j^\dagger - E_{\underline{g}_j, \underline{h}_{j+1}}\left\{Q_{j+1}\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger\right) | \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\}\right] \\
&\quad + \overline{I}_j\left[E_{\underline{g}_j, \underline{h}_{j+1}}\left\{Q_{j+1}\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger\right) | \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\} - \eta_j\right] \\
&= \Gamma_j + \overline{I}_j \sum_{k=j+1}^{K} E_{\underline{g}_j, \underline{h}_{j+1}}\left\{\frac{\overline{I}_{j+1}^{(k-1)}}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger}\right)\left(\eta_k^\dagger - \eta_k\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\} \\
&= \Gamma_j + \overline{I}_j \sum_{k=j+1}^{K} E_{\underline{g}_j, \underline{h}_{j+1}}\left\{\frac{1}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{1}{h_k} - \frac{1}{h_k^\dagger}\right)\overline{I}_k\left(\eta_k^\dagger - \eta_k\right) \middle| \overline{A}_j, \overline{L}_j\right\},
\end{aligned}$$

where the last equivalence follows from the fact that $\overline{I}_j E\left\{U(O) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\} = \overline{I}_j E\left\{\overline{I}_j U(O) \middle| \overline{A}_j, \overline{L}_j\right\}$ for any function $U(\cdot)$.

Then, invoking the inductive assumption we obtain

$$\overline{I}_j\left(\eta_j^\dagger-\eta_j\right)=\Gamma_j+\overline{I}_j\sum_{k=j+1}^{K}E_{\underline{g}_j,\underline{h}_{j+1}}\left\{\frac{1}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{1}{h_k}-\frac{1}{h_k^\dagger}\right)\Gamma_k\middle|\overline{A}_j,\overline{L}_j\right\}$$
$$+\overline{I}_j\sum_{k=j+1}^{K}\sum_{s=k+1}^{K}E_{\underline{g}_j,\underline{h}_{j+1}}\left[\frac{1}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{1}{h_k}-\frac{1}{h_k^\dagger}\right)\overline{I}_k\frac{1}{\pi_{k+1}^{s-1}}\left(\frac{1}{h_s}-\frac{1}{h_s^\dagger}\right)\Gamma_s\middle|\overline{A}_j,\overline{L}_j\right]$$

Now, rearranging the terms in the double-sum and using the fact that $\overline{I}_k\Gamma_s=\Gamma_s$ for all $s\geq k+1$, we obtain

$$\sum_{k=j+1}^{K}\sum_{s=k+1}^{K}E_{\underline{g}_j,\underline{h}_{j+1}}\left[\frac{1}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{1}{h_k}-\frac{1}{h_k^\dagger}\right)\overline{I}_k\frac{1}{\pi_{k+1}^{s-1}}\left(\frac{1}{h_s}-\frac{1}{h_s^\dagger}\right)\Gamma_s\middle|\overline{A}_j,\overline{L}_j\right]$$
$$=\sum_{s=j+2}^{K}E_{\underline{g}_j,\underline{h}_{j+1}}\left[\left(\frac{1}{h_s}-\frac{1}{h_s^\dagger}\right)\Gamma_s\sum_{k=j+1}^{s-1}\left\{\frac{1}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{1}{h_k}-\frac{1}{h_k^\dagger}\right)\frac{1}{\pi_{k+1}^{s-1}}\right\}\middle|\overline{A}_j,\overline{L}_j\right]$$

and we prove below that

$$\sum_{k=j+1}^{s-1}\left\{\frac{1}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{1}{h_k}-\frac{1}{h_k^\dagger}\right)\frac{1}{\pi_{k+1}^{s-1}}\right\}=\frac{1}{\pi_{j+1}^{s-1}}-\frac{1}{\pi_{j+1}^{\dagger(s-1)}}\tag{B.8}$$

Thus,

$$\overline{I}_j\left(\eta_j^\dagger-\eta_j\right)=\Gamma_j+\overline{I}_jE_{\underline{g}_j,\underline{h}_{j+1}}\left\{\left(\frac{1}{h_{j+1}}-\frac{1}{h_{j+1}^\dagger}\right)\Gamma_{j+1}\middle|\overline{A}_j,\overline{L}_j\right\}$$
$$+\overline{I}_j\sum_{s=j+2}^{K}E_{\underline{g}_j,\underline{h}_{j+1}}\left\{\frac{1}{\pi_{j+1}^{\dagger(s-1)}}\left(\frac{1}{h_s}-\frac{1}{h_s^\dagger}\right)\Gamma_s\middle|\overline{A}_j,\overline{L}_j\right\}$$
$$+\overline{I}_j\sum_{s=j+2}^{K}E_{\underline{g}_j,\underline{h}_{j+1}}\left\{\left(\frac{1}{\pi_{j+1}^{s-1}}-\frac{1}{\pi_{j+1}^{\dagger(s-1)}}\right)\left(\frac{1}{h_s}-\frac{1}{h_s^\dagger}\right)\Gamma_s\middle|\overline{A}_j,\overline{L}_j\right\}$$
$$=\Gamma_j+\overline{I}_j\sum_{s=j+1}^{K}E_{\underline{g}_j,\underline{h}_{j+1}}\left\{\frac{1}{\pi_{j+1}^{s-1}}\left(\frac{1}{h_s}-\frac{1}{h_s^\dagger}\right)\Gamma_s\middle|\overline{A}_j,\overline{L}_j\right\}$$

as we wish to show.

We now show (B.8) .

$$\begin{aligned}
&\sum_{k=j+1}^{s-1} \frac{1}{\pi_{j+1}^{\dagger(k-1)}} \left(\frac{1}{h_k} - \frac{1}{h_k^\dagger} \right) \frac{1}{\pi_{k+1}^{s-1}} \\
&= \sum_{k=j+1}^{s-1} \frac{1}{\pi_{j+1}^{\dagger(k-1)}} \frac{1}{\pi_k^{s-1}} - \sum_{k=j+1}^{s-1} \frac{1}{\pi_{j+1}^{\dagger k}} \frac{1}{\pi_{k+1}^{s-1}} \\
&= \frac{1}{\pi_{j+1}^{s-1}} + \sum_{k=j+2}^{s-1} \frac{1}{\pi_{j+1}^{\dagger(k-1)}} \frac{1}{\pi_k^{s-1}} - \sum_{k=j+1}^{s-2} \frac{1}{\pi_{j+1}^{\dagger k}} \frac{1}{\pi_{k+1}^{s-1}} - \frac{1}{\pi_{j+1}^{\dagger s-1}} \\
&= \frac{1}{\pi_{j+1}^{s-1}} - \frac{1}{\pi_{j+1}^{\dagger(s-1)}}
\end{aligned}$$

This concludes the proof that $b^p\left(h^\dagger, \eta^\dagger\right) = a^p\left(h^\dagger, \eta^\dagger\right)$.

We now prove that $c^p\left(h^\dagger, \eta^\dagger\right) = a^p\left(h^\dagger, \eta^\dagger\right)$. Lemma 15 implies that

$$E_{g_k}\left\{ \eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\} - \eta_k\left(\overline{L}_k\right) = \sum_{j=k+1}^{K} E_{\underline{g}_k, \underline{h}_{k+1}} \left\{ \frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j} \delta_j\left(\eta_j^\dagger, \eta_{j+1}^\dagger; g_j\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\}$$

where recall that

$$\delta_j\left(\eta_j^\dagger, \eta_{j+1}^\dagger; g_j\right) \equiv \eta_j^\dagger\left(\overline{L}_j\right) - E_{g_j}\left\{ \eta_{j+1}^\dagger\left(\overline{L}_{j+1}\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\}.$$

Hence,

$$\begin{aligned}
\eta_k^\dagger\left(\overline{L}_k\right) - \eta_k\left(\overline{L}_k\right) &= \eta_k^\dagger\left(\overline{L}_k\right) - E_{g_k}\left\{ \eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\} \\
&\quad + E_{g_k}\left\{ \eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\} - \eta_k\left(\overline{L}_k\right) \\
&= \sum_{j=k}^{K} E_{\underline{g}_k, \underline{h}_{k+1}} \left\{ \frac{\overline{I}_{k+1}^j}{\pi_{k+1}^j} \delta_j\left(\eta_j^\dagger, \eta_{j+1}^\dagger; g_j\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_k \right\}.
\end{aligned} \tag{B.9}$$

Then,

$$
\begin{aligned}
a^p\left(h^\dagger,\eta^\dagger\right) &\equiv \sum_{k=1}^{K} E_p\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^\dagger}\right)\left(\eta_k^\dagger-\eta_k\right)\right\}\\
&= \sum_{k=1}^{K} E_p\left[\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^\dagger}\right)\sum_{j=k}^{K} E_p\left\{\frac{\overline{I}_{k+1}^{j}}{\pi_{k+1}^{j}}\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\middle|\,\overline{A}_k=\overline{a}_k^*,\overline{L}_k\right\}\right]\\
&= \sum_{k=1}^{K} E_p\left[\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^\dagger}\right)\sum_{j=k}^{K} E_p\left\{\frac{\overline{I}_{k+1}^{j}}{\pi_{k+1}^{j}}\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\middle|\,\overline{A}_k,\overline{L}_k\right\}\right]\\
&= \sum_{k=1}^{K} E_p\left[\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^\dagger}\right)\sum_{j=k}^{K}\frac{\overline{I}_{k+1}^{j}}{\pi_{k+1}^{j}}\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\right]\\
&= \sum_{j=1}^{K} E_p\left[\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\sum_{k=1}^{j}\left\{\frac{\overline{I}_{k-1}}{\pi^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^\dagger}\right)\frac{\overline{I}_{k+1}^{j}}{\pi_{k+1}^{j}}\right\}\right]\\
&= \sum_{j=1}^{K} E_p\left[\delta_j\left(\eta_j^\dagger,\eta_{j+1}^\dagger;g_j\right)\overline{I}_j\sum_{k=1}^{j}\left\{\frac{1}{\pi^{\dagger(k-1)}}\left(\frac{1}{h_k}-\frac{1}{h_k^\dagger}\right)\frac{1}{\pi_{k+1}^{j}}\right\}\right].
\end{aligned}
$$

where the third equivalence follows from the fact that $\overline{I}_k u\left(\overline{A}_k,\overline{L}_k\right)=\overline{I}_k u\left(\overline{a}_k^*,\overline{L}_k\right)$ for any function $u(\cdot,\cdot)$.

The result $c^p\left(h^\dagger,\eta^\dagger\right)=a^p\left(h^\dagger,\eta^\dagger\right)$ is then proved if we show that

$$
\sum_{k=1}^{j}\left\{\frac{1}{\pi^{\dagger(k-1)}}\left(\frac{1}{h_k}-\frac{1}{h_k^\dagger}\right)\frac{1}{\pi_{k+1}^{j}}\right\}=\frac{1}{\pi^j}-\frac{1}{\pi^{\dagger j}}. \tag{B.10}
$$

But (B.10) follows from (B.8) by evaluating in (B.8) j at 0 and s at $j+1$. This concludes the proof. ■

B.3 Analysis of the empirical processes difference term

In this appendix, we show that (2.30) implies (2.31) when $\left(h^{\dagger}, \eta^{\dagger}\right)$ are replaced by the estimators $\left(\widehat{h}, \widehat{\eta}\right)$ used to compute $\widehat{\theta}^{u}$ or by the estimators $\left(\widehat{h}, \widehat{\eta}_{MR}\right)$ used to compute $\widehat{\theta}^{u}_{MR}$.

Assume that

$$E_p\left[\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}^2 |\mathcal{N}\right] = o_p\left(1\right) \text{ as } n \to \infty, \tag{B.11}$$

with $\left(h^{\dagger}, \eta^{\dagger}\right)$ equal to the estimators $\left(\widehat{h}, \widehat{\eta}\right)$ used to compute $\widehat{\theta}^{u}$ or to the estimators $\left(\widehat{h}, \widehat{\eta}_{MR}\right)$ used to compute $\widehat{\theta}^{u}_{MR}$. We must show that

$$\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\} = o_p\left(1\right) \text{ as } n \to \infty.$$

Note that

$$\begin{aligned}
&\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\} \equiv \\
&\equiv \frac{1}{\sqrt{N_u}} \sum_{i: O_i \in \mathcal{D}^u} \left\{Q\left(h^{\dagger}, \eta^{\dagger}\right)_i - Q\left(h, \eta\right)_i - E_p\left[Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)|\mathcal{N}\right]\right\} \\
&= \frac{1}{\sqrt{N_u}} \sum_{i: O_i \in \mathcal{D}^u} \left\{\delta_i^{\dagger} - E_p\left(\delta^{\dagger}|\mathcal{N}\right)\right\}
\end{aligned}$$

with

$$\begin{aligned}
\delta_i^{\dagger} &\equiv Q\left(h^{\dagger}, \eta^{\dagger}\right)_i - Q\left(h, \eta\right)_i \\
&= q\left(O_i; h^{\dagger}, \eta^{\dagger}\right) - q\left(O_i; h, \eta\right)
\end{aligned}$$

and

$$\begin{aligned}
\delta^{\dagger} &\equiv Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right) \\
&= q\left(O; h^{\dagger}, \eta^{\dagger}\right) - q\left(O; h, \eta\right).
\end{aligned}$$

Then, it becomes clear that, as noted in Section 2.5, $\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}$ is equal to $\sqrt{N_u}$ times an average of N_u random variables that, conditionally on the data in $\mathcal{N}$, are i.i.d. and have mean zero. It implies that

$$E_p\left[\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}\big|\mathcal{N}\right] = 0$$

and that

$$\begin{aligned}
Var_p\left[\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}\big|\mathcal{N}\right] &= Var_p\left[Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\big|\mathcal{N}\right] \\
&\leq E_p\left[\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}^2\Big|\mathcal{N}\right] = o_p\left(1\right)
\end{aligned}$$

as $n \to \infty$ by (B.11).

Now, given $\varepsilon > 0$ let $R_{n,\varepsilon} \equiv P\left[\left|\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}\right| > \varepsilon\big|\mathcal{N}\right]$. Then,

$$R_{n,\varepsilon} = P\left[\left|\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}\right| > \varepsilon\big|\mathcal{N}\right] \leq \frac{Var_p\left[\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger}, \eta^{\dagger}\right) - Q\left(h, \eta\right)\right\}\big|\mathcal{N}\right]}{\varepsilon^2} = o_p\left(1\right)$$

as $n \to \infty$. But, $R_{n,\varepsilon}$ is a sequence of random variables that converges to 0 in probability and is bounded (by 1). Then, $E_p\left(R_{n,\varepsilon}\right) \underset{n\to\infty}{\to} 0$, which implies that

$$P\left(\left|\mathbb{G}_{N_u}\left\{Q\left(h^{\dagger},\eta^{\dagger}\right)-Q\left(h,\eta\right)\right\}\right|>\varepsilon\right) \underset{n\to\infty}{\to} 0.$$

as we wanted to show.

B.4 Proof of Theorem 2

The proof of Theorem 2 invokes the following Lemma.

Lemma 16

For any $j \in [K]$,

$$E_p\left\{Q_j\left(\underline{h}_j^\dagger, \underline{\eta}_j^\dagger\right)\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right\} - \eta_j\left(\overline{L}_j\right) = \tag{B.12}$$
$$= \sum_{k=j}^{K} E_p\left\{\frac{\overline{I}_j^{(k-1)}}{\pi_j^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger}\right)\left(\eta_k^\dagger - \eta_k\right)\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right\}.$$

Proof of Lemma 16. Recall that, given $j \in [K]$,

$$Q_j\left(\underline{h}_j^\dagger, \underline{\eta}_j^\dagger\right) \equiv \frac{I_j}{h_j^\dagger\left(\overline{L}_j\right)}\left\{Q_{j+1}\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger\right) - \eta_j^\dagger\left(\overline{L}_j\right)\right\} + \eta_j^\dagger\left(\overline{L}_j\right),$$

then

$$E_p\left\{Q_j\left(\underline{h}_j^\dagger, \underline{\eta}_j^\dagger\right)\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right\} = \frac{1}{h_j^\dagger} E_p\left[I_j\left\{Q_{j+1}\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger\right) - \eta_j^\dagger\right\}\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right] + \eta_j^\dagger$$

and we prove below that, for any function $u(\cdot)$ with domain in the sample space of O,

$$E_p\left\{I_j u(O)\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right\} = h_j\left(\overline{L}_j\right) E_p\left\{u(O)\middle|\, \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\}. \tag{B.13}$$

Hence,

$$\begin{aligned}
E_p\left\{Q_j\left(\underline{h}_j^\dagger, \underline{\eta}_j^\dagger\right)\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right\} &= \frac{h_j}{h_j^\dagger} E_p\left\{Q_{j+1}\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger\right) - \eta_j^\dagger\middle|\, \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\} + \eta_j^\dagger \\
&= \frac{h_j}{h_j^\dagger}\left[E_p\left\{Q_{j+1}\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger\right)\middle|\, \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\} - \eta_j^\dagger\right] + \eta_j^\dagger \\
&= \frac{h_j}{h_j^\dagger}\left[a_j^p\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger; \overline{L}_j\right) + \eta_j - \eta_j^\dagger\right] + \eta_j^\dagger
\end{aligned}$$

where the last equality follows from Lemma 7. Thus,

$$E_p\left\{Q_j\left(\underline{h}_j^\dagger, \underline{\eta}_j^\dagger\right)\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right\} - \eta_j = \left(1 - \frac{h_j}{h_j^\dagger}\right)\left(\eta_j^\dagger - \eta_j\right) + \frac{h_j}{h_j^\dagger} a_j^p\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger; \overline{L}_j\right).$$

But

$$\begin{aligned}
\left(1 - \frac{h_j}{h_j^\dagger}\right)\left(\eta_j^\dagger - \eta_j\right) &= h_j\left(\frac{1}{h_j} - \frac{1}{h_j^\dagger}\right)\left(\eta_j^\dagger - \eta_j\right) \\
&= h_j E\left\{\left(\frac{1}{h_j} - \frac{1}{h_j^\dagger}\right)\left(\eta_j^\dagger - \eta_j\right)\middle|\, \overline{A}_j = \overline{a}_j^*, \overline{L}_j\right\} \\
&= E\left[\left(\frac{I_j}{h_j} - \frac{I_j}{h_j^\dagger}\right)\left(\eta_j^\dagger - \eta_j\right)\middle|\, \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j\right],
\end{aligned}$$

where the last equality follows from (B.13) . Then, we would arrive at the desired result if we show that

$$\frac{h_j}{h_j^\dagger} a_j^p\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger; \overline{L}_j\right) = \sum_{k=j+1}^{K} E_p\left\{ \frac{\overline{I}_j^{k-1}}{\pi_j^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger}\right)\left(\eta_k^\dagger - \eta_k\right) \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j \right\}.$$

Now,

$$\begin{aligned}
\frac{h_j}{h_j^\dagger} a_j^p\left(\underline{h}_{j+1}^\dagger, \underline{\eta}_{j+1}^\dagger; \overline{L}_j\right) &= \frac{h_j}{h_j^\dagger} \sum_{k=j+1}^{K} E_p\left\{ \frac{\overline{I}_{j+1}^{k-1}}{\pi_{j+1}^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger}\right)\left(\eta_k^\dagger - \eta_k\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= \sum_{k=j+1}^{K} h_j E_p\left\{ \frac{\overline{I}_{j+1}^{k-1}}{\pi_j^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger}\right)\left(\eta_k^\dagger - \eta_k\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \\
&= \sum_{k=j+1}^{K} E_p\left\{ \frac{\overline{I}_j^{k-1}}{\pi_j^{\dagger(k-1)}}\left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger}\right)\left(\eta_k^\dagger - \eta_k\right) \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j \right\},
\end{aligned}$$

again by (B.13) . We now show (B.13) . Note that

$$\begin{aligned}
E_p\left\{ I_j u\left(O\right) \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j \right\} &= E_p\left[E_p\left\{ I_j u\left(O\right) \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, A_j, \overline{L}_j \right\} \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j \right] \\
&= E_p\left[I_j E_p\left\{ u\left(O\right) \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, A_j, \overline{L}_j \right\} \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j \right] \\
&= E_p\left[I_j E_p\left\{ u\left(O\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j \right] \\
&= E_p\left\{ u\left(O\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} E_p\left(I_j \middle| \overline{A}_{j-1} = \overline{a}_{j-1}^*, \overline{L}_j \right) \\
&= E_p\left\{ u\left(O\right) \middle| \overline{A}_j = \overline{a}_j^*, \overline{L}_j \right\} h_j\left(\overline{L}_j\right).
\end{aligned}$$

This concludes the proof. ■

Proof of Theorem 2. We prove part (1) by induction. Part (2) follows immediately.

For $k = K$, (2.44) is true because $\widetilde{\eta}_{K,DR} = \eta_{K,DR}$ since $\widetilde{\eta}_{K+1,DR}\left(\overline{L}_{K+1}\right) \equiv \eta_{K+1}\left(\overline{L}_{K+1}\right) \equiv \kappa\left(\overline{L}_{K+1}\right)$.

Suppose (2.44) is true for $k = K, ..., j+1$. We will show it is true for $k = j$.

$$\begin{aligned}
\widetilde{\eta}_{j,DR} - \eta_j &= \Pi^j\left[\widetilde{\eta}_{j+1,DR}\left(\overline{L}_{j+1}\right)\right] - \eta_j \\
&= \left(\eta_{j,DR} - \eta_j\right) + \Pi^j\left[\widetilde{\eta}_{j+1,DR}\left(\overline{L}_{j+1}\right)\right] - \eta_{j,DR} \\
&= \left(\eta_{j,DR} - \eta_j\right) + \Pi^j\left[\widetilde{\eta}_{j+1,DR}\left(\overline{L}_{j+1}\right) - \eta_{j+1}\left(\overline{L}_{j+1}\right)\right]
\end{aligned}$$

and we prove bellow that, for any function $u\left(\overline{L}_{j+1}\right)$ and for $j \in [K]$,

$$\Pi^j\left[u\left(\overline{L}_{j+1}\right)\right] = \Pi_{DR}^j\left[u\left(\overline{L}_{j+1}\right)\right]. \tag{B.14}$$

Hence,

$$\begin{aligned}
\widetilde{\eta}_{j,DR} - \eta_j &= (\eta_{j,DR} - \eta_j) + \Pi^j_{DR}\left[\widetilde{\eta}_{j+1,DR} - \eta_{j+1}\right] \\
&= (\eta_{j,DR} - \eta_j) + \Pi^j_{DR}\left[\eta_{j+1,DR} - \eta_{j+1} + \sum_{k=j+2}^{K} \Pi_{DR,j+1,k}\left[\eta_{k,DR} - \eta_k\right]\right] \\
&= (\eta_{j,DR} - \eta_j) + \Pi^j_{DR}\left[\eta_{j+1,DR} - \eta_{j+1}\right] + \sum_{k=j+2}^{K} \Pi^j_{DR}\left[\Pi_{DR,j+1,k}\left\{\eta_{k,DR} - \eta_k\right\}\right] \\
&= (\eta_{j,DR} - \eta_j) + \sum_{k=j+1}^{K} \Pi_{DR,j,k}\left[\eta_{k,DR} - \eta_k\right]
\end{aligned}$$

The second equality is by the inductive hypothesis and third is by the assumed linearity of the operator Π^j which induces linearity of the operator Π^j_{DR}. This concludes the proof of part (1).

We now show (B.14) . Note that, by definition

$$\Pi^j_{DR}\left[u\left(\overline{L}_{j+1}\right)\right] = \Pi^j\left\{E_p\left(\frac{I_{j+1}}{h_{j+1}} u\left(\overline{L}_{j+1}\right)\middle| \overline{A}_j = \overline{a}^*_j, \overline{L}_{j+1}\right)\right\},$$

but

$$\begin{aligned}
E_p\left(\frac{I_{j+1}}{h_{j+1}} u\left(\overline{L}_{j+1}\right)\middle| \overline{A}_j = \overline{a}^*_j, \overline{L}_{j+1}\right) &= \frac{u\left(\overline{L}_{j+1}\right)}{h_{j+1}\left(\overline{L}_{j+1}\right)} E_p\left(I_{j+1}|\, \overline{A}_j = \overline{a}^*_j, \overline{L}_{j+1}\right) \\
&= u\left(\overline{L}_{j+1}\right),
\end{aligned}$$

from where (B.14) follows.

We now prove part (3) by induction in K. Part (4) follows immediately. First we show (2.45) is true when $K = 1$. For $K = 1$, we have

$$\begin{aligned}
&E_p\left\{\left(\frac{I_1}{h_1} - \frac{I_1}{h^\dagger_1}\right)\left(\widetilde{\eta}_{1,MR} - \eta_1\right)\middle| L_1\right\} = \\
&= E_p\left\{\left(\frac{I_1}{h_1} - \frac{I_1}{h^\dagger_1}\right)\left(\eta_{1,MR} - \eta_1\right)\middle| L_1\right\} \\
&= E_p\left\{\left(\frac{I_1}{h_1} - \frac{I_1}{h^\dagger_1}\right)\left(\eta_{1,MR} - \eta_1\right)\middle| L_1\right\} \\
&+ \sum_{1\le r_1<r_2<\ldots<r_u\le K-1}\ \sum_{k=r_u+1}^{K} E_p\left(\nabla_{0,r_1}\Pi_{MR,r_1,r_2,\ldots,r_u,k}\left[\eta_{k,MR} - \eta_k\right]|\, L_1\right)
\end{aligned}$$

where the first equality follows because when $K = 1, Q_2\left(\underline{\overline{h}}^\dagger_2, \underline{\widetilde{\eta}}_{2,MR}\right) \equiv \kappa\left(\overline{L}_2\right)$ so $\widetilde{\eta}_{1,MR} \equiv \Pi^1\left[Q_2\left(\underline{\overline{h}}^\dagger_2, \underline{\widetilde{\eta}}_{2,MR}\right)\right] \equiv \Pi^1\left[\kappa\left(\overline{L}_2\right)\right] \equiv \eta_{1,MR}$ and the second equality is true because $\sum_{1\le r_1<r_2<\ldots<r_u\le 0}(\cdot) \equiv 0$. This proves (2.45) for $K = 1$.

Next, assume (2.45) is true for $K-1$, we will show it is true for K. If (2.45) is true for $K-1$, then it holds that

$$\begin{aligned}
&\sum_{k=2}^{K} E_p\left\{\frac{\overline{I}_2^{k-1}}{\pi_2^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^\dagger}\right)\left(\widetilde{\eta}_{k,MR}-\eta_k\right)\middle| A_1,\overline{L}_2\right\}\\
&=\sum_{k=2}^{K} E_p\left\{\nabla_{1,k}\left(\widetilde{\eta}_{k,MR}-\eta_k\right)|\, A_1,\overline{L}_2\right\}\\
&=\sum_{k=2}^{K} E_p\left\{\nabla_{1,k}\left(\eta_{k,MR}-\eta_k\right)|\, A_1,\overline{L}_2\right\}\\
&+\sum_{2\le r_1<r_2<\ldots<r_u\le K-1}\ \sum_{j=r_u+1}^{K} E_p\left\{\nabla_{1,r_1}\Pi_{MR,r_1,r_2,\ldots,r_u,j}\left[\eta_{j,MR}-\eta_j\right]|\, A_1,\overline{L}_2\right\}
\end{aligned}$$

Note that in the preceding expression we used the inductive hypothesis pretending that our study started at cycle 2 instead of cycle 1, i.e. with $\left(A_1,\overline{L}_2\right)$ playing the role of L_1, with each A_j playing the role of $A_{j-1}, j=2,\ldots,K$, with each L_j playing the role of $L_{j-1}, j=3,\ldots,K+1$, and with $\nabla_{1,k}$ playing the role of $\nabla_{0,k-1}, k=2,\ldots,K$.

We also have that

$$\begin{aligned}
\widetilde{\eta}_{1,MR}-\eta_1 &= \widetilde{\eta}_{1,MR}-\eta_{1,MR}+\eta_{1,MR}-\eta_1\\
&=\left(\eta_{1,MR}-\eta_1\right)+\Pi^1\left[Q_2\left(\underline{h}_2^\dagger,\widetilde{\underline{\eta}}_{2,MR}\right)\right]-\Pi^1\left[\eta_2\left(\overline{L}_2\right)\right]\\
&-\Pi^1\left[Q_2\left(\underline{h}_2^\dagger,\widetilde{\underline{\eta}}_{2,MR}\right)-E_p\left\{Q_2\left(\underline{h}_2^\dagger,\widetilde{\underline{\eta}}_{2,MR}\right)\middle| A_1=a_1^*,\overline{L}_2\right\}\right]\\
&=\left(\eta_{1,MR}-\eta_1\right)+\Pi^1\left[E_p\left\{Q_2\left(\underline{h}_2^\dagger,\widetilde{\underline{\eta}}_{2,MR}\right)\middle| A_1=a_1^*,\overline{L}_2\right\}-\eta_2\left(\overline{L}_2\right)\right]\\
&=\left(\eta_{1,MR}-\eta_1\right)+\Pi^1\left[\sum_{k=2}^{K} E_p\left\{\frac{\overline{I}_2^{(k-1)}}{\pi_2^{\dagger(k-1)}}\left(\frac{I_k}{h_k}-\frac{I_k}{h_k^\dagger}\right)\left(\widetilde{\eta}_{k,MR}-\eta_k\right)\middle| A_1=a_1^*,\overline{L}_2\right\}\right]\\
&=\left(\eta_{1,MR}-\eta_1\right)+\Pi^1\left[\sum_{k=2}^{K} E_p\left\{\nabla_{1,k}\left(\widetilde{\eta}_{k,MR}-\eta_k\right)|\, A_1=a_1^*,\overline{L}_2\right\}\right]\\
&=\left(\eta_{1,MR}-\eta_1\right)+\Pi^1\left[\sum_{k=2}^{K} E_p\left[\nabla_{1,k}\left(\eta_{k,MR}-\eta_k\right)|\, A_1=a_1^*,\overline{L}_2\right]\right]\\
&+\Pi^1\left[\sum_{2\le r_1<r_2<\ldots<r_u\le K-1}\ \sum_{j=r_u+1}^{K} E_p\left\{\nabla_{1,r_1}\Pi_{MR,r_1,r_2,\ldots,r_u,j}\left[\eta_{j,MR}-\eta_j\right]|\, A_1=a_1^*,\overline{L}_2\right\}\right]
\end{aligned}$$

where the fourth equality follows after invoking Lemma 16 and the sixth is by the inductive hypothesis. Hence,

$$\widetilde{\eta}_{1,MR} - \eta_1 = (\eta_{1,MR} - \eta_1) + \sum_{k=2}^{K} \Pi_{MR,1,k} \left[\eta_{k,MR} - \eta_k\right] \tag{B.15}$$
$$+ \sum_{2 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{j=r_u+1}^{K} \Pi_{MR,1,r_1,r_2,\ldots,r_u,j} \left[\eta_{j,MR} - \eta_j\right].$$

So,

$$\begin{aligned}
&\sum_{k=1}^{K} E_p \left\{ \frac{\overline{I}_{k-1}}{\pi^{\dagger k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger} \right) (\widetilde{\eta}_{k,MR} - \eta_k) \middle| L_1 \right\} \\
&= \sum_{k=2}^{K} E_p \left[\frac{I_1}{h_1^\dagger} E_p \left\{ \frac{\overline{I}_2^{k-1}}{\pi_2^{\dagger k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger} \right) (\widetilde{\eta}_{k,MR} - \eta_k) \middle| A_1, \overline{L}_2 \right\} \middle| L_1 \right] \\
&+ E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) (\widetilde{\eta}_{1,MR} - \eta_1) \middle| L_1 \right\} \\
&= E_p \left\{ \frac{I_1}{h_1^\dagger} \left[\sum_{k=2}^{K} E_p \left\{ \frac{\overline{I}_2^{k-1}}{\pi_2^{\dagger k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^\dagger} \right) (\widetilde{\eta}_{k,MR} - \eta_k) \middle| A_1, \overline{L}_2 \right\} \right] \middle| L_1 \right\} \\
&+ E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) (\widetilde{\eta}_{1,MR} - \eta_1) \middle| L_1 \right\} \\
&= E_p \left\{ \frac{I_1}{h_1^\dagger} \left[\sum_{k=2}^{K} E_p \left\{ \nabla_{1,k} (\widetilde{\eta}_{k,MR} - \eta_k) | A_1, \overline{L}_2 \right\} \right] \middle| L_1 \right\} \\
&+ E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) (\widetilde{\eta}_{1,MR} - \eta_1) \middle| L_1 \right\} \\
&= E_p \left[\frac{I_1}{h_1^\dagger} \sum_{k=2}^{K} E_p \left\{ \nabla_{1,k} (\eta_{k,MR} - \eta_k) | A_1, \overline{L}_2 \right\} \middle| L_1 \right] \\
&+ E_p \left\{ \frac{I_1}{h_1^\dagger} \left[\sum_{2 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{j=r_u+1}^{K} E_p \left\{ \nabla_{1,r_1} \Pi_{MR,r_1,r_2,\ldots,r_u,j} \left[\eta_{j,MR} - \eta_j\right] | A_1, \overline{L}_2 \right\} \right] \middle| L_1 \right\} \\
&+ E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) (\eta_{1,MR} - \eta_1) \middle| L_1 \right\} \\
&+ E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) \sum_{k=2}^{K} \Pi_{MR,1,k} \left[\eta_{k,MR} - \eta_k\right] \middle| L_1 \right\} \\
&+ E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{h_1^\dagger} \right) \sum_{2 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{j=r_u+1}^{K} \Pi_{MR,1,r_1,r_2,\ldots,r_u,j} \left[\eta_{j,MR} - \eta_j\right] \middle| L_1 \right\}.
\end{aligned}$$

The last equality follows by inductive hypothesis and by (B.15). Thus,

$$\sum_{k=1}^{K} E_p \left\{ \frac{\overline{I}_{k-1}}{\pi^{\dagger k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{h_k^{\dagger}} \right) (\widetilde{\eta}_{k,MR} - \eta_k) \middle| L_1 \right\} =$$

$$= \sum_{k=1}^{K} E_p \left\{ \nabla_{0,k} \left(\eta_{k,MR} - \eta_k \right) | \, L_1 \right\}$$

$$+ \sum_{2 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{j=r_u+1}^{K} E_p \left(\nabla_{0,r_1} \Pi_{MR,r_1,r_2,\ldots,r_u,j} \left[\eta_{j,MR} - \eta_j \right] | \, L_1 \right)$$

$$+ \sum_{k=2}^{K} E_p \left\{ \nabla_{0,1} \Pi_{MR,1,k} \left[\eta_{k,MR} - \eta_k \right] | \, L_1 \right\}$$

$$+ \sum_{2 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{j=r_u+1}^{K} E_p \left(\nabla_{0,1} \Pi_{MR,1,r_1,r_2,\ldots,r_u,j} \left[\eta_{j,MR} - \eta_j \right] | \, L_1 \right)$$

$$= \sum_{k=1}^{K} E_p \left\{ \nabla_{0,k} \left(\eta_{k,MR} - \eta_k \right) | \, L_1 \right\}$$

$$+ \sum_{1 \leq r_1 < r_2 < \ldots < r_u \leq K-1} \sum_{j=r_u+1}^{K} E_p \left(\nabla_{0,r_1} \Pi_{MR,r_1,r_2,\ldots,r_u,j} \left[\eta_{j,MR} - \eta_j \right] | \, L_1 \right)$$

This concludes the proof of Theorem 2. ■

B.5 Technical results on the convergence of series estimators

The results presented in this appendix rely on results about the $L^2(F)$ and uniform rates of convergence of least squares series estimators proved in [3]. To introduce them, it will be convenient to define some notation. As in [3], we consider a sequence of models indexed by the sample size n,

$$\begin{aligned} Y_i &= g(X_i) + \varepsilon_i, \quad E(\varepsilon_i | X_i) = 0, \\ X_i &\in \mathcal{X} \subseteq \mathbb{R}^d, \ i = 1, \ldots, n \end{aligned} \tag{B.16}$$

where Y_i is a response variable, X_i is a vector of continuous covariates with distribution F, ε_i a noise and $g(x) = E(Y_i | X_i = x)$ a regression (conditional mean) function; that is, we consider a triangular array of models with $Y_i = Y_{i,n}$, $X_i = X_{i,n}, \varepsilon_i = \varepsilon_{i,n}$ and $g = g_n$. We assume that $g \in \mathcal{G}$ where $\mathcal{G}$ is some class of functions. Since we consider a sequence of models indexed by the sample size n, we allow $\mathcal{G} = \mathcal{G}_n$ to depend on n. In addition, we allow $\mathcal{X} = \mathcal{X}_n$ and $d = d_n$ to depend on n, as well but we assume that the diameter of $\mathcal{X}$ is bounded from above uniformly over n. We also assume that $\mathcal{X}$ is compact. For notational convenience we omit indexing by n where it does not lead to confusion.

Although the results in this appendix are used in Subsection 2.7.2 in the special case in which the model is fixed, i.e. not changing with n, we present them in the more general case in which the model is allowed to change with n.

Condition A.1 (Sample) For each n, random vectors $(Y_i, X_i')', i = 1, \ldots, n$ are i.i.d. and satisfy (B.16).

Suppose we approximate the function $g(x)$ by linear forms $p(x)' b$ where

$$p(x) \equiv (p_1(x), \ldots, p_m(x))'$$

is the vector of the first m elements of a dictionary of approximating functions $\{p_j(\cdot)\}_{j \geq 1}$ that can change with n; in particular, m may increase with n. The next assumption imposes regularity conditions on the regressors $p_j(X_i), j = 1, \ldots, m$.

Condition A.2 (Eigenvalues) Uniformly over n, eigenvalues of $Q \equiv E\{p(X_i) p(X_i)'\}$ are bounded above and bellow away from zero.

Condition A.2 imposes the restriction that $p_1(X_i), \ldots, p_m(X_i)$ are not too colinear.

Given $f \in \mathcal{G}$, let

$$\beta_f \equiv \arg\min_{b \in \mathbb{R}^m} E\left[\{f(X_i) - p(X_i)' b\}^2\right]$$

and, for all $x \in \mathcal{X}$, let

$$r_f(x) \equiv f(x) - p(x)' \beta_f.$$

The function $p(x)' \beta_f$ provides the best linear approximation to the function $f(x)$ in norm $L^2(F)$ and, hence, $r_f(x)$ represents the approximation error in that norm. Let

$$\beta \equiv \beta_g \equiv \arg\min_{b \in \mathbb{R}^m} E\left[\{g(X_i) - p(X_i)' b\}^2\right].$$

Model (B.16) implies that $\beta = \arg\min_{b \in \mathbb{R}^m} E\left[\left\{Y_i - p(X_i)' b\right\}^2\right]$. Hence, the least squares estimator of β is

$$\widehat{\beta} \equiv \arg\min_{b \in \mathbb{R}^m} \mathbb{P}_n\left[\left\{Y - p(X)' b\right\}^2\right]$$

This estimator induces the estimator

$$\widehat{g}(x) \equiv p(x)' \widehat{\beta}$$

for the target function $g(x)$.

For any function f in $L^2(F)$, we denote $\|f\|_{L_2(F)} \equiv \sqrt{\int_{x \in \mathcal{X}} f(x)^2 dF(x)}$. We also denote

$$\xi_m \equiv \|p\|_\infty$$

and

$$\xi_m^L \equiv \sup_{x, x' \in \mathcal{X}: x \neq x'} \frac{\|\alpha(x) - \alpha(x')\|}{\|x - x'\|}$$

with $\alpha(x) \equiv \frac{p(x)}{\|p(x)\|}$.

The following condition is related to the approximation properties of the dictionary $\{p_j(\cdot)\}_{j \geq 1}$ to the functions in the class $\mathcal{G}$.

Condition A.3 (Approximation) For each n and m there are finite constants c_m and l_m such that

i $sup_{f \in \mathcal{G}} \|r_f\|_{L_2(F)} \leq c_m$ and
ii $sup_{f \in \mathcal{G}} \|r_f\|_\infty \leq l_m c_m$.

together c_m and l_m characterize the approximating properties of the underlying functions under $L^2(F)$ and uniform distances. Note that constants $c_m = c_m(\mathcal{G})$ and $l_m = l_m(\mathcal{G})$ are allowed to depend on n but we omit indexing by n for simplicity of notation.

Let $q > 2$. The following assumption imposes restrictions on the tails of the regression errors.

Condition A.4 (Disturbances) Regression errors satisfy

$$\sup_{x \in \mathcal{X}} E\left(|\varepsilon_i|^q \middle| X_i = x\right) \lesssim 1.$$

We will also need the following assumption on the dictionary to hold with the same $q > 2$ as in that in Condition A.4

Condition A.5 (Basis) Dictionary functions are such that (i) $\xi_m^{2q/(q-2)} \frac{\log m}{n} \lesssim 1$, (ii) $\log \xi_m^L \lesssim \log m$, and (iii) $\log \xi_m \lesssim \log m$.

Finally we denote

$$\overline{R}_{1n} \equiv \sqrt{\frac{\xi_m^2 \log m}{n}} \left(n^{1/q} \sqrt{\log m} + \sqrt{m}.l_m c_m\right)$$

and

$$\overline{R}_{2n} \equiv \sqrt{\log m}.l_m c_m.$$

Next, we introduce two results derived from Theorems 4.1 and 4.3 in [3]

Theorem 7 (from Theorem 4.1 of Belloni et al. - L^2 rate of convergence) *Assume that Conditions A.1, A.2 and Condition A.3 (part i) are satisfied. In addition, assume that (i)* $\frac{\xi_m^2 \log m}{n} \underset{n\to\infty}{\to} 0$ *and (ii)* $\sup_{x\in\mathcal{X}} E\left(\varepsilon_i^2 \middle| X_i = x\right) \lesssim 1$, *then*

$$\|\widehat{g} - g\|_{L_2(F)} \lesssim_P \sqrt{m/n} + c_m.$$

Theorem 8 (from Theorem 4.3 of Belloni et al. - uniform rate of convergence) *Assume that the conditions A-1-A.5 are satisfied. Then*

$$\|\widehat{g} - g\|_\infty \lesssim_P \frac{\xi_m}{\sqrt{n}}\left(\sqrt{\log m} + \overline{R}_{1n} + \overline{R}_{2n}\right) + l_m c_m$$

In what follows, we assume that the class of functions $\mathcal{G}$ to which the regression function g belongs, is contained in a Hölder ball with finite radius and known smoothness order $s > 0$. Let

$$\gamma \equiv \frac{s}{d}$$

where recall d is the dimension of the covariates. Since $\mathcal{G}$ is allowed to change with n, s is also allowed to depend on n. Furthermore, d may change with n. However, in what follows we assume that $\gamma = \frac{s}{d}$ is constant, that is, independent of n. Recall that in Subsection 2.7.2 we apply the results of this appendix for the case in which the whole model is fixed, so that the requirement that γ is independent on n is satisfied in that setting.

The following Lemma 17 shows that, under regularity conditions, if the number m of dictionary elements is chosen to balance the trade off between approximation error and sampling error, the optimal L^2 rate of convergence of nonparametric estimators $n^{-\frac{s}{2s+d}}$ - of regression functions in a Hölder ball $\mathcal{H}(\mathcal{X}; s, \rho)$ - can be achieved by series estimators that use dictionaries satisfying certain optimal approximation properties. Then, in Lemma 18, we show that, under regularity conditions, if we choose a dictionary satisfying certain more restrictive approximation properties and if the number m of dictionary elements is chosen to yield the L_2 optimal rate of convergence, then the series estimator is L_∞ consistent if, in addition, $\gamma > \frac{1}{6}$. Furthermore, we find the L_∞ rate of convergence of the series estimator under these conditions. Both Lemmas 17 and 18 assume that (1) $\xi_m \lesssim \sqrt{m}$. In addition, Lemma 17 requires that (2) part i of Condition A.3 holds with c_m verifying that $c_m \lesssim m^{-\gamma}$. Finally, Lemma 18 also assume that (3) Condition A.3 is holds with l_m and c_m such that $l_m c_m \lesssim m^{-\gamma}$. Examples of dictionaries satisfying (1) and (2) are Cohen-Daubechies-Vial wavelets, B-spline and local polynomial partition series. If, in addition, uniformly over n, the pdf of F is bounded from above and bellow away from zero, these dictionaries also verify (3), (see [3]).

Lemma 17 *Assume that*

1. $\mathcal{G}$ *is contained in a Hölder ball with finite radius and known smoothness order* $s > 0$,
2. *Conditions A.1 and A.2 are satisfied,*
3. *Part i of Condition A.3 is verified with* $c_m \lesssim m^{-\gamma}$,
4. $\xi_m \lesssim \sqrt{m}$, *and*
5. $\underset{x\in\mathcal{X}}{sup} E\left(\varepsilon_i^2 \middle| X_i = x\right) \lesssim 1$.

Then, if we set $m \asymp n^{\frac{1}{2\gamma+1}}$, *we have that*

$$\|\widehat{g} - g\|_{L_2(F)} \lesssim_P n^{-\frac{\gamma}{2\gamma+1}}.$$

Proof. First note that, since $\xi_m \lesssim \sqrt{m}$ and $m \asymp n^{\frac{1}{2\gamma+1}}$,

$$\frac{\xi_m^2 \log m}{n} \lesssim \frac{m \log m}{n} \lesssim n^{\frac{1}{2\gamma+1}-1} \log\left(n^{\frac{1}{2\gamma+1}}\right) \underset{n\to\infty}{\to} 0$$

Thus, all the assumptions of Theorem 7 hold, so that

$$\begin{aligned}\|\widehat{g} - g\|_{L_2(F)} &\lesssim_P \sqrt{m/n} + c_m \\ &\lesssim \sqrt{m/n} + m^{-\gamma}\end{aligned}$$

by assumption 3 of the lemma. Now, the fact that $m \asymp n^{\frac{1}{2\gamma+1}}$ implies that

(I) $\sqrt{m/n} \asymp n^{-\frac{\gamma}{2\gamma+1}}$ and

(II) $m^{-\gamma} \lesssim n^{-\frac{\gamma}{2\gamma+1}}$

from where we arrive at

$$\|\widehat{g} - g\|_{L_2(F)} \lesssim_P n^{-\frac{\gamma}{2\gamma+1}}$$

as we wanted to show. ■

Lemma 18 *Assume that*

1. $\mathcal{G}$ *is contained in a Hölder ball with finite radius and known smoothness order* $s > 0$ *such that* $\gamma \equiv \frac{s}{d} > \frac{1}{6}$,
2. *Conditions A.1 and A.2 are satisfied,*
3. *Condition A.3 is verified with* $l_m c_m \lesssim m^{-\gamma}$,
4. *Condition A.4 is satisfied for some* $q > 2 + \frac{1}{\gamma}$,
5. $\xi_m \lesssim \sqrt{m}$, *and*
6. $\log \xi_m^L \lesssim \log m$.

Then, if we set $m \asymp n^{\frac{1}{2\gamma+1}}$, *we have that*

a) $\|\widehat{g} - g\|_\infty = O_p\left(n^{-\frac{6\gamma-1}{4\gamma+2}} \sqrt{\log n}\right)$, *if* $\frac{1}{6} < \gamma < \frac{1}{4}$, *and*

b) $\|\widehat{g} - g\|_\infty = O_p\left(n^{-\frac{\gamma}{2\gamma+1}} \log n\right)$ *if* $\gamma \geq \frac{1}{4}$.

In particular,

c) $\|\widehat{g} - g\|_\infty = o_p(1)$ *for all* $\gamma > \frac{1}{6}$.

Proof. Note that assumption 5 and the fact that $m \asymp n^{\frac{1}{2\gamma+1}}$ imply that

$$\xi_m^{2q/(q-2)} \frac{\log m}{n} \lesssim m^{q/(q-2)} \frac{\log m}{n} \asymp n^{\frac{1}{2\gamma+1}\frac{q}{q-2}-1} \log\left(n^{\frac{1}{2\gamma+1}}\right).$$

But, $\frac{1}{2\gamma+1}\frac{q}{q-2} - 1 < 0$ since $q > 2 + \frac{1}{\gamma}$ by assumption 4. Hence, the sequence in the right hand side of previous display tends to zero and, thus, part (i) of Condition A.5 is verified for the same q of the Condition A.4. Furthermore, part (ii) of Condition A.5 is verified by assumption 6 of the lemma. Finally, assumption 5 also implies part (iii) of Condition A.5, since $\log \xi_m \lesssim \log \sqrt{m} \lesssim \log m$. Then, all the assumptions of Theorem 8 are verified, so that

$$\|\widehat{g} - g\|_\infty \lesssim_P \frac{\xi_m}{\sqrt{n}} \left(\sqrt{\log m} + \overline{R}_{1n} + \overline{R}_{2n}\right) + l_m c_m. \tag{B.17}$$

Now, since $c_m l_m \lesssim m^{-\gamma}$ and $\xi_m \lesssim \sqrt{m}$ by assumptions 3 and 5, the bound in (B.17) becomes,

$$\|\widehat{g} - g\|_\infty \lesssim_P \sqrt{\frac{m}{n}} \left(\sqrt{\log m} + \overline{R}_{1n} + \overline{R}_{2n}\right) + m^{-\gamma}.$$

Here, recall that $\overline{R}_{2n} = \sqrt{\log m}.l_m c_m$. But $c_m l_m \lesssim m^{-\gamma} \lesssim 1$ and, hence, $\overline{R}_{2n} \lesssim \sqrt{\log m}$. Thus,

$$\sqrt{\frac{m}{n}} \left(\sqrt{\log m} + \overline{R}_{2n}\right) \lesssim \sqrt{\frac{m \log m}{n}}$$

which yields

$$\|\widehat{g} - g\|_\infty \lesssim_P \sqrt{\frac{m \log m}{n}} + \sqrt{\frac{m}{n}} \overline{R}_{1n} + m^{-\gamma}.$$

Now, since $m \asymp n^{\frac{1}{2\gamma+1}}$, we have that

(I) $\sqrt{\frac{m \log m}{n}} \lesssim \sqrt{n^{\frac{-2\gamma}{2\gamma+1}} \log\left(n^{\frac{1}{2\gamma+1}}\right)} = n^{\frac{-\gamma}{2\gamma+1}} \sqrt{\frac{1}{2\gamma+1} \log n} \leq n^{\frac{-\gamma}{2\gamma+1}} \sqrt{\log n}$, and

(II) $m^{-\gamma} \lesssim n^{\frac{-\gamma}{2\gamma+1}}$.

Then,

$$\sqrt{\frac{m \log m}{n}} + m^{-\gamma} \lesssim n^{\frac{-\gamma}{2\gamma+1}} \sqrt{\log n} \tag{B.18}$$

Next, we will show that

$$\sqrt{\frac{m}{n}} \overline{R}_{1n} \lesssim \max\left\{n^{-\frac{\gamma}{2\gamma+1}} \log n, n^{-\frac{6\gamma-1}{4\gamma+2}} \sqrt{\log n}\right\}. \tag{B.19}$$

To see this, note that, since $\xi_m \lesssim \sqrt{m}$ and $c_m l_m \lesssim m^{-\gamma}$,

$$\begin{aligned}
\overline{R}_{1n} &= \sqrt{\frac{\xi_m^2 \log m}{n}} \left(n^{1/q} \sqrt{\log m} + \sqrt{m}.l_m c_m\right) \\
&\lesssim \sqrt{\log m} \sqrt{\frac{m}{n}} \left(n^{1/q} \sqrt{\log m} + \sqrt{m}.m^{-\gamma}\right) \\
&= \underbrace{\sqrt{m} n^{\left(\frac{1}{q} - \frac{1}{2}\right)} \log m}_{a_n} + \underbrace{\sqrt{\log m} \frac{m^{(1-\gamma)}}{\sqrt{n}}}_{b_n}.
\end{aligned}$$

Now, using again that $m \asymp n^{\frac{1}{2\gamma+1}}$, we arrive at

$$\sqrt{\frac{m}{n}} a_n = m.n^{\left(\frac{1}{q}-1\right)} \log m \lesssim n^{\frac{-2\gamma}{2\gamma+1}+\frac{1}{q}} \log\left(n^{\frac{1}{2\gamma+1}}\right) \leq n^{\frac{-2\gamma}{2\gamma+1}+\frac{1}{q}} \log n.$$

But $\frac{-2\gamma}{2\gamma+1} + \frac{1}{q} < -\frac{\gamma}{2\gamma+1}$ since $q > 2 + \frac{1}{\gamma}$ by assumption 4 of the lemma. Thus, we have that

$$\sqrt{\frac{m}{n}} a_n \lesssim n^{-\frac{\gamma}{2\gamma+1}} \log n. \tag{B.20}$$

Also

$$\sqrt{\frac{m}{n}} b_n = m^{\left(\frac{3}{2}-\gamma\right)} n^{-1} \sqrt{\log m} \lesssim n^{-\frac{6\gamma-1}{4\gamma+2}} \sqrt{\log\left(n^{\frac{1}{2\gamma+1}}\right)} \leq n^{-\frac{6\gamma-1}{4\gamma+2}} \sqrt{\log n}. \tag{B.21}$$

Finally, (B.20) and (B.21) imply (B.19).

Then, (B.18) and (B.19) imply that

$$\|\widehat{g} - g\|_\infty \lesssim_P \max\left\{n^{-\frac{\gamma}{2\gamma+1}} \log n, n^{-\frac{6\gamma-1}{4\gamma+2}} \sqrt{\log n}\right\}.$$

But, $\frac{\gamma}{2\gamma+1} \leq \frac{6\gamma-1}{4\gamma+2}$ iff $\gamma \geq \frac{1}{4}$, then $\|\widehat{g} - g\|_\infty \lesssim_P n^{-\frac{6\gamma-1}{4\gamma+2}} \sqrt{\log n}$ if $\frac{1}{6} < \gamma < \frac{1}{4}$ and $\|\widehat{g} - g\|_\infty \lesssim_P n^{-\frac{6\gamma-1}{4\gamma+2}} \sqrt{\log n}$ if $\gamma \geq \frac{1}{4}$, which concludes the proof. ■

B.6 Proofs of Theorems 3 to 6

In this appendix, we prove Theorems 3 to 6. To do that we need to introduce some results, which we do now.

B.6.1 Previous technical results

The following result is used in the proofs of Theorems 3 - 6 to bound the L_∞ norm of $\left(\widehat{h}_k\right)^{-2}, k \in [K]$.

Lemma 19 *Let G be a fixed real-valued function and let $\widehat{G}_n$ be a sequence of random real-valued function, both with domain in the same subset $\mathcal{X} \subseteq \mathbb{R}^d$. If $\left\|\widehat{G}_n - G\right\|_\infty = o_p(1)$ and $|G(x)| > \sigma > 0$ for all $x \in \mathcal{X}$, then*

$$\left\|\frac{1}{\widehat{G}_n^2}\right\|_\infty = O_p(1).$$

This lemma is an immediate corollary of the following result.

Lemma 20 *Let G be a fixed real-valued function and let $\widehat{G}_n$ be a sequence of random real-valued function, both with domain in the same subset $\mathcal{X} \subseteq \mathbb{R}^d$. If $\left\|\widehat{G}_n - G\right\|_\infty = o_p(1)$ and $|G(x)| > \sigma > 0$ for all $x \in \mathcal{X}$, then*

$$\left\|\frac{1}{\widehat{G}_n^2} - \frac{1}{G^2}\right\|_\infty = o_p(1).$$

Proof. For each fixed x there exists $\widehat{T}_n(x)$ satisfying $\left|\widehat{T}_n(x) - G(x)\right| \leq \left|\widehat{G}_n(x) - G(x)\right|$ such that

$$\frac{1}{\widehat{G}_n(x)^2} - \frac{1}{G(x)^2} = \frac{-2}{\widehat{T}_n(x)^3}\left\{\widehat{G}_n(x) - G(x)\right\}.$$

Then, $\sup_x \left|\frac{1}{\widehat{G}_n(x)^2} - \frac{1}{G(x)^2}\right| \leq \frac{2}{\left(\inf_x\left|\widehat{T}_n(x)\right|\right)^3} \sup_x \left|\widehat{G}_n(x) - G(x)\right|$. Since $\sup_x \left|\widehat{G}_n(x) - G(x)\right| = o_p(1)$, if we prove that $\inf_x \left|\widehat{T}_n(x)\right| \geq \sigma/2$ with probability tending to one, then we have the desired result. But,

$$\begin{aligned}
\inf_x \left|\widehat{T}_n(x)\right| &= \inf_x \left|G(x) - \left[G(x) - \widehat{T}_n(x)\right]\right| \\
&\geq \inf_x \left\{|G(x)| - \left|\widehat{T}_n(x) - G(x)\right|\right\} \\
&\geq \inf_x |G(x)| - \sup_x \left|\widehat{T}_n(x) - G(x)\right| \\
&\geq \sigma - \sup_x \left|\widehat{G}_n(x) - G(x)\right|
\end{aligned}$$

Now, given $\varepsilon > 0$ choose n_0 such that $P\left(\sup_{x\in\mathcal{X}} \left|\widehat{G}_n(x) - G(x)\right| > \frac{\sigma}{2}\right) < \varepsilon$ for $n \geq n_0$. For such

n,

$$P\left(\inf_x \left|\widehat{T}_n(x)\right| \geq \sigma/2\right) \geq P\left(\sup_{x\in\mathcal{X}} \left|\widehat{G}_n(x) - G(x)\right| < \frac{\sigma}{2}\right) \geq 1-\varepsilon,$$

then $\inf_x \left|\widehat{T}_n(x)\right| \geq \sigma/2$ with probability tending to one, as we wanted to prove. ■

The results in the following Lemmas 21 and 22 are crucial to prove Theorems 3 to 6. Specifically, part (d) of Lemma 21 is used to bound the terms involving successive application of the linear operators $\Pi^j[\cdot]$ to different functions of the estimation errors $\eta_{k,DR-}\eta_k$ or $\eta_{k,MR-}\eta_k$, for $j<k$, by terms involving only the functions of the estimation errors $\eta_{k,DR-}\eta_k$ or $\eta_{k,MR-}\eta_k$. On the other hand, Lemma 22 is used to show that in the estimation error

$$\begin{aligned}\eta_{k,MR} - \eta_k &= \eta_{k,DR} - \eta_k \\ &+ \Pi^k\left[q_{k+1}\left(\cdot,\cdot;\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right) - E_p\left\{Q_{k+1}\left(\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1}=\cdot,\underline{\mathcal{N}}^{k+1}\right\}\right],\end{aligned}$$

the term that dominates the L_2 rate of convergence is $\eta_{k,DR-}\eta_k$. As will become clear next, to apply this lemmas to our problem, we strongly use the fact that each η_k is estimated using a different subsample $\mathcal{N}^k$ of the nuisance sample $\mathcal{N}$. The proofs of these results rely on arguments used in the proof of Theorem 4.1 of [3].

To present these results, we need first to introduce some notation. Consider a sequence of datasets

$$\mathcal{S}_n \equiv \{O_i : i=1,\ldots,n\} \tag{B.22}$$

of i.i.d. copies of $O\equiv(X,A,W)$ with X a random vector with sample space $\mathcal{X}\subseteq\mathbb{R}^d$, A a discrete random variable with sample space $\mathcal{A}$ and W another random vector. We also consider a sequence of vector-valued functions $p^{(n)}(\cdot):\mathcal{X}\to\mathbb{R}^{m(n)}$,

$$p^{(n)}(x) \equiv \left(p_1(x),\ldots,p_{m(n)}(x)\right)', \tag{B.23}$$

containing the first $m(n)$ elements of a dictionary $\{p_j(\cdot)\}_{j\geq 1}$, where $m(n)$ may change with n. Throughout, to alleviate the notation, we omit indexing by n where it does not lead to confusion. Then, we write m instead of $m(n)$, $p(x)$ instead of $p^{(n)}(x)$.

Given some fixed $a\in\mathcal{A}$, we define

$$I_a \equiv I_{\{a\}}(A), \tag{B.24}$$

$$Q_a \equiv E\left\{I_a p(X)\, p(X)'\right\} \tag{B.25}$$

and

$$\widehat{Q}_a \equiv \mathbb{P}_n\left\{I_a p(X)\, p(X)'\right\}. \tag{B.26}$$

Note that, although not explicit in the notation, $\widehat{Q}_a$ depends on n, moreover Q_a depends on n since $p(\cdot)$ does. Now, consider another sequence of datasets $\mathcal{S}_n^*$ independent of $\mathcal{S}_n$ and let $\widehat{y}(\cdot)$ be a real-valued function with domain in the sample space of O that may depend on the data in $\mathcal{S}_n^*$. Finally denote

$$\Pi[\widehat{y}](\cdot) \equiv \widehat{\beta}' p(\cdot) \tag{B.27}$$

with

$$\begin{aligned}\widehat{\beta} &\equiv \widehat{Q}_a^{-1}\mathbb{P}_n\left\{I_a p\left(X\right)\widehat{y}\left(O\right)\right\} \\ &= \left\{\sum_{i=1}^{n} I_{\{a\}}\left(A_i\right) p\left(X_i\right) p\left(X_i\right)'\right\}^{-1}\left\{\sum_{i=1}^{n} I_{\{a\}}\left(A_i\right) p\left(X_i\right)'\widehat{y}\left(O_i\right)\right\}\end{aligned}$$

the least squares coefficient in the regression of $\widehat{Y} \equiv \widehat{y}\left(O\right)$ on $p\left(X\right)$ in the subsample of $\mathcal{S}_n$ of observations with $A_i = a$. Note that $\widehat{\beta}$ is a function of $\mathcal{S}_n$ and $\mathcal{S}_n^*$.

Lemma 21 *Let $\mathcal{S}_n$, $p\left(\cdot\right), I_a, Q_a$, $\widehat{Q}_a, \mathcal{S}_n^*, \widehat{y}\left(\cdot\right)$ and $\Pi\left[\widehat{y}\right]\left(\cdot\right)$ as defined in (B.22) – (B.27). Also, let $\mathbb{P}_n$ be the empirical distribution of units in $\mathcal{S}_n$. Suppose that*

1. *uniformly over n, eigenvalues of Q_a are bounded above and bellow away from zero,*
2. *$\left\|p\right\|_\infty \lesssim \sqrt{m}$ and*
3. *$m \asymp n^\alpha$ with $0 < \alpha < 1$,*

Then, for $O \equiv (X, A, W)$ independent of $\mathcal{S}_n$ and $\mathcal{S}_n^*$,

a) $E\left[I_a\left\{\Pi\left[\widehat{y}\right]\left(X\right)\right\}^2\middle|\mathcal{S}_n, \mathcal{S}_n^*\right] \lesssim \left\|\widehat{\beta}\right\|^2$,

b) $\left\|\widehat{\beta}\right\|^2 \lesssim_P E\left\{I_a\widehat{y}\left(O\right)^2\middle|\mathcal{S}_n^*\right\}$,

c) $\left\|\widehat{\beta}\right\|^2 \lesssim_P \left\|\mathbb{P}_n\left\{I_a\widehat{y}\left(O\right)p\left(X\right)\right\}\right\|^2$,

and therefore,

d) $E\left[I_a\left\{\Pi\left[\widehat{y}\right]\left(X\right)\right\}^2\middle|\mathcal{S}_n, \mathcal{S}_n^*\right] \lesssim_P E\left\{I_a\widehat{y}\left(O\right)^2\middle|\mathcal{S}_n^*\right\}$ *and*

e) $E\left[I_a\left\{\Pi\left[\widehat{y}\right]\left(X\right)\right\}^2\middle|\mathcal{S}_n, \mathcal{S}_n^*\right] \lesssim_P \left\|\mathbb{P}_n\left\{I_a\widehat{y}\left(O\right)p\left(X\right)\right\}\right\|^2$.

Lemma 22 *Let $\mathcal{S}_n$, $p\left(\cdot\right), I_a, Q_a$, $\widehat{Q}_a, \mathcal{S}_n^*, \widehat{y}\left(\cdot\right)$ and $\Pi\left[\widehat{y}\right]\left(\cdot\right)$ as defined in (B.22) – (B.27). Suppose that*

1. *uniformly over n, eigenvalues of Q_a are bounded above and bellow away from zero,*
2. *$\left\|p\right\|_\infty \lesssim \sqrt{m}$,*
3. *$m \asymp n^\alpha$ with $0 < \alpha < 1$,*

Also, suppose that, for $O \equiv (X, A, W)$ independent of $\mathcal{S}_n$ and $\mathcal{S}_n^*$,

4. *$E\left\{I_a\widehat{y}\left(O\right)\middle|X, \mathcal{S}_n^*\right\} = 0$ and*
5. *$E\left\{I_a\widehat{y}\left(O\right)^2\middle|\mathcal{S}_n^*\right\} \lesssim_P 1$.*

Then, for O independent of $\mathcal{S}_n$ and $\mathcal{S}_n^*$,

$$E\left[I_a\left\{\Pi\left[\widehat{y}\right](X)\right\}^2\middle|\mathcal{S}_n,\mathcal{S}_n^*\right]\lesssim_P \frac{m}{n}.$$

To prove Lemmas 21 and 22 we first need to show the following results.

Lemma 23 *For any sequence of non negative random variables X_n with finite expectation it holds that $X_n = O_p\left\{E\left(X_n\right)\right\}$.*

Proof. The result follows immediately after noticing that for any $M > 0$ and any n,

$$P\left(\left|\frac{X_n}{E\left(X_n\right)}\right| > M\right) \leq \frac{1}{M}E\left(\left|\frac{X_n}{E\left(X_n\right)}\right|\right) = \frac{1}{M},$$

where the inequality follows from Markov's inequality and the equality follows because $X_n \geq 0$. ■

Lemma 24 *Let X_n and W_n be sequences of random variables such that $W_n = O_p\left(1\right)$, and such that $W_n = W_n\left(D_n\right)$ is a function of data D_n. Suppose that for any $M > 0$ it holds that*

$$P\left(|X_n| > M|D_n\right) < \frac{W_n\left(D_n\right)}{M}.$$

Then, $X_n = O_p\left(1\right)$.

Proof. We want to show that given any $\varepsilon > 0$ there exists M_ε such that for all n

$$P\left(|X_n| > M_\varepsilon\right) < \varepsilon.$$

Let $\delta > 0$ and let K_δ be such that for all n,

$$P\left(W_n > K_\delta\right) < \delta.$$

Then, for any C

$$\begin{aligned}
P\left(|X_n| > C\right) &= E\left[P\left(|X_n| > C|D_n\right)\right] \\
&= E\left[P\left(|X_n| > C|D_n, W_n > K_\delta\right)P\left(W_n > K_\delta|D_n\right)\right] \\
&+ E\left[P\left(|X_n| > C|D_n, W_n < K_\delta\right)P\left(W_n < K_\delta|D_n\right)\right] \\
&\leq E\left[P\left(W_n > K_\delta|D_n\right)\right] + E\left[P\left(|X_n| > C|D_n, W_n < K_\delta\right)\right] \\
&\leq P\left(W_n > K_\delta\right) + \frac{K_\delta}{C} \\
&\leq \delta + \frac{K_\delta}{C}.
\end{aligned}$$

Now, take $\delta = \varepsilon/2$ and take $C = \frac{K_\delta}{(\varepsilon/2)}$. Then, $\delta + \frac{K_\delta}{C} = \varepsilon$. So,

$$P\left(|X_n| > \frac{K_\delta}{(\varepsilon/2)}\right) < \varepsilon$$

which shows that X_n is bounded in probability. ■

Corollary 1 *For any sequence of non negative random variables X_n and any a sequence of random vectors D_n, it holds that $X_n \lesssim_P E(X_n|D_n)$.*

Proof. Since $X_n \geq 0$, Markov's inequality implies that

$$P\left(\left|\frac{X_n}{E(X_n|D_n)}\right| > M \middle| D_n\right) \leq \frac{E\left\{\frac{X_n}{E(X_n|D_n)} \middle| D_n\right\}}{M} = \frac{1}{M}.$$

Then, the assumptions of Lemma 24 hold with $\frac{X_n}{E(X_n|D_n)}$ playing the roll of X_n and with 1 playing the roll of W_n, from where we conclude that $\frac{X_n}{E(X_n|D_n)} = O_p(1)$ as we wanted to show. ■

In what follows, for any matrix M, we use $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ to denote the minimum and maximum eigenvalue of M respectively. Also, throughout, we will use repeatedly the fact that if $M \in \mathbb{R}^{m\times m}$ is symmetric, then $\lambda_{\min}(M) = \min_{\|u\|=1}\{u'Mu\}$, $\lambda_{\max}(M) = \max_{\|u\|=1}\{u'Mu\}$ and, hence $v'Mv \leq \lambda_{\max}(M)\|v\|^2$ for any $v \in \mathbb{R}^m$. Likewise, we will use the fact that if $M \in \mathbb{R}^{m\times m}$ is symmetric, then $\|M\| \leq \lambda_{\max}(M)$.

Lemma 25 *Let $\{A_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^{m\times m}$ be a sequence of symmetric fixed real-valued matrices and let $\left\{\widehat{A}_n\right\}_{n\in\mathbb{N}} \subset \mathbb{R}^{m\times m}$ be a sequence of symmetric random real-valued matrices. Assume that $\left\|\widehat{A}_n - A_n\right\| \xrightarrow{P} 0$. Then,*

a) if uniformly over n, the eigenvalues of A_n are bounded above by a constant $C < \infty$, then with probability going to one, the eigenvalues of $\widehat{A}_n$ are bounded above by $\frac{3}{2}C$, and

b) if uniformly over n, the eigenvalues of A_n are bounded bellow by a constant $D > 0$, then with probability going to one, the eigenvalues of $\widehat{A}_n$ are bounded bellow by $\frac{1}{2}D$.

Proof. First note that, since $\widehat{A}_n$ and A_n are symmetric real-valued matrices, $\lambda_{\min}(A_n) = \min_{\|u\|=1}\{u'A_nu\}$, $\lambda_{\max}(A_n) = \max_{\|u\|=1}\{u'A_nu\}$, $\lambda_{\min}\left(\widehat{A}_n\right) = \min_{\|u\|=1}\left\{u'\widehat{A}_nu\right\}$ and $\lambda_{\max}\left(\widehat{A}_n\right) = \max_{\|u\|=1}\left\{u'\widehat{A}_nu\right\}$.

To see (a), note that for any n,

$$\begin{aligned}
P\left(\lambda_{\max}\left(\widehat{A}_n\right) \leq \frac{3}{2}C\right) &= P\left(u'\widehat{A}_nu \leq \frac{3}{2}C \text{ for all } u \in \mathbb{R}^m \text{ with } \|u\| = 1\right) \\
&\geq P\left(\left|u'\widehat{A}_nu - u'A_nu\right| \leq \frac{1}{2}C \text{ for all } u \in \mathbb{R}^m \text{ with } \|u\| = 1\right) \\
&\geq P\left(\left\|\widehat{A}_n - A_n\right\| \leq \frac{1}{2}C\right),
\end{aligned}$$

where the first inequality follows from the fact that $\lambda_{\max}(A_n) \leq C$ for all n. But $P\left(\left\|\widehat{A}_n - A_n\right\| \leq \frac{1}{2}C\right) \to 1$ since $\left\|\widehat{A}_n - A_n\right\| \xrightarrow{P} 0$ and, hence, $P\left(\lambda_{\max}\left(\widehat{A}_n\right) \leq \frac{3}{2}C\right) \to 1$ as we wanted to show.

Fact (b) follows from an analogous argument by noticing that

$$\begin{aligned} P\left(\lambda_{\min}\left(\widehat{A}_n\right) \geq \frac{1}{2}D\right) &= P\left(u'\widehat{A}_n u \geq \frac{1}{2}D \text{ for all } u \text{ with } \|u\| = 1\right) \\ &\geq P\left(\left|u'\widehat{A}_n u - u'A_n u\right| \leq \frac{1}{2}D \text{ for all } u \text{ with } \|u\| = 1\right) \\ &\geq P\left(\left\|\widehat{A}_n - A_n\right\| \leq \frac{1}{2}D\right) \to 1, \end{aligned}$$

where the first inequality now follows from the fact that $\lambda_{\min}(A_n) \geq D$ for all n. This concludes the proof of fact (b). ■

The following result follows from Lemma 6.2 of [3] and is a variant of a fundamental result obtained by Rudelson ([48]), which is a sort of law of large numbers for matrices.

Lemma 26 (from Lemma 6.2 of Belloni et al.) *Let $P_1^{(n)}, \ldots, P_n^{(n)}$ be i.i.d. copies of a $M(n) \times 1$ vector $P^{(n)}$. Assume that there exists a sequence $\{\xi_n\}_{n\geq 1}$ such that $\left\|P^{(n)}\right\| \leq \xi_{M(n)}$ a.s. Let $R_n \equiv E\left\{P^{(n)}P^{(n)\prime}\right\}$ and $\widehat{R}_n \equiv \mathbb{P}_n\left\{P^{(n)}P^{(n)\prime}\right\}$. Then,*

$$E\left\{\left\|\widehat{R}_n - R_n\right\|\right\} \lesssim \xi^2_{M(n)}\frac{\log M(n)}{n} + \sqrt{\xi^2_{M(n)}\frac{\log M(n)}{n}\|R_n\|}.$$

The following lemma follows from Lemma 26 applying arguments used in the proof of Theorem 4.1 of [3].

Lemma 27 *Let $p(\cdot), m, Q_a$ and $\widehat{Q}_a$ as defined in (B.23), (B.25) and (B.26). If $\|p\|^2_\infty \frac{\log m}{n} \to 0$ and, uniformly over n, eigenvalues of Q_a are bounded above, then $\left\|\widehat{Q}_a - Q_a\right\| \xrightarrow{P} 0$.*

Proof. By Lemma 23, it is enough to show that $E\left\{\left\|\widehat{Q}_a - Q_a\right\|\right\} \to 0$. Now, note that

$$Q_a = E\left\{\widetilde{p}(X,A)\widetilde{p}(X,A)'\right\}$$

and

$$\widehat{Q}_a = \mathbb{P}_n\left\{\widetilde{p}(X,A)\widetilde{p}(X,A)'\right\}$$

with $\widetilde{p}(X,A) \equiv I_a p(X)$. But, $\|\widetilde{p}(X,A)\| \leq \|p\|_\infty$, then we can apply Lemma 26 replacing $P^{(n)}$ by $\widetilde{p}(X,A)$ and $\xi_{m(n)}$ by $\|p\|_\infty$, from where we conclude that

$$E\left\{\left\|\widehat{Q}_a - Q_a\right\|\right\} \lesssim \|p\|^2_\infty \frac{\log m}{n} + \sqrt{\|p\|^2_\infty \frac{\log m}{n}\|Q_a\|}.$$

Also, $\|Q_a\| \leq \lambda_{\max}(Q_a) \lesssim 1$ where the inequality follows because Q_a is symmetric and the bound follows because the eigenvalues of Q_a are bounded above, by the assumption of the lemma. Thus, we arrive at

$$E\left\{\left\|\widehat{Q}_a - Q_a\right\|\right\} \lesssim \|p\|^2_\infty \frac{\log m}{n} \to 0,$$

by assumption of the lemma. This concludes the proof. ■

Proof of Lemma 21. To see (a), note that

$$\begin{aligned} E\left[I_a\left\{\Pi\left[\widehat{y}\right](X)\right\}^2\middle|\mathcal{S}_n,\mathcal{S}_n^*\right] &= E\left\{I_a\widehat{\beta}'p(X)\,p(X)'\,\widehat{\beta}\middle|\mathcal{S}_n,\mathcal{S}_n^*\right\} \\ &= \widehat{\beta}'E\left\{I_a p(X)\,p(X)'\right\}\widehat{\beta} \\ &= \widehat{\beta}'Q_a\widehat{\beta} \leq \lambda_{\max}(Q_a)\left\|\widehat{\beta}\right\|^2. \end{aligned}$$

Now, assumption 1 of the lemma, implies that there exist constants C_1 and C_2 such that

$$0 < C_1 < \lambda_{\min}(Q_a) \leq \lambda_{\max}(Q_a) < C_2 < \infty \tag{B.28}$$

for all n. Here recall that, although not explicit in the notation, Q_a may depend on n. Then,

$$E\left[I_a\left\{\Pi\left[\widehat{y}\right](X)\right\}^2\middle|\mathcal{S}_n,\mathcal{S}_n^*\right] \leq C_2\left\|\widehat{\beta}\right\|^2,$$

which concludes the proof of fact (a). To see (b), note that

$$\begin{aligned} \left\|\widehat{\beta}\right\|^2 &= \left\|\widehat{Q}_a^{-1}\mathbb{P}_n\left\{I_a\widehat{y}(O)\,p(X)\right\}\right\|^2 = \left\|\widehat{Q}_a^{-1/2}\underbrace{\widehat{Q}_a^{-1/2}\mathbb{P}_n\left\{I_a\widehat{y}(O)\,p(X)\right\}}_{w}\right\|^2 \\ &= w'\widehat{Q}_a^{-1}w \leq \lambda_{\max}\left(\widehat{Q}_a^{-1}\right)\|w\|^2. \end{aligned}$$

But, (I) $\lambda_{\min}(Q_a) > C_1 > 0$ for all n by B.28 and (II) $\left\|\widehat{Q}_a - Q_a\right\| \xrightarrow{P} 0$. To see fact (II), note that assumptions 2 and 3 of the lemma imply that

$$\|p\|_\infty^2\frac{\log m}{n} \lesssim m\frac{\log m}{n} \lesssim n^{(\alpha-1)}\log\left(n^\alpha\right) \to 0, \tag{B.29}$$

since $\alpha < 1$. Thus, B.29, assumption 1 of the lemma and Lemma 27 imply that

$$\left\|\widehat{Q}_a - Q_a\right\| \xrightarrow{P} 0.$$

Now, facts (I) and (II) and Lemma 25 imply that $\lambda_{\min}\left(\widehat{Q}_a\right) > \frac{1}{2}C_1$ with probability going to one, so that $\lambda_{\max}\left(\widehat{Q}_a^{-1}\right) = \left\{\lambda_{\min}\left(\widehat{Q}_a\right)\right\}^{-1} < \frac{2}{C_1}$ with probability going to one. Hence, $\lambda_{\max}\left(\widehat{Q}_a^{-1}\right)$ is bounded in probability, thus yielding

$$\begin{aligned} \left\|\widehat{\beta}\right\|^2 &\lesssim_P \left\|\widehat{Q}_a^{-1/2}\mathbb{P}_n\left\{I_a\widehat{y}(O)\,p(X)\right\}\right\|^2 \\ &= \mathbb{P}_n\left\{I_a\widehat{y}(O)\,p(X)\right\}'\widehat{Q}_a^{-1}\mathbb{P}_n\left\{I_a\widehat{y}(O)\,p(X)\right\} \\ &= \mathbb{P}_n\left\{I_a\widehat{y}(O)\,p(X)\right\}'\widehat{\beta} \\ &= \mathbb{P}_n\left\{I_a\widehat{y}(O)\,p(X)'\,\widehat{\beta}\right\} \\ &= \mathbb{P}_n\left\{I_a\widehat{y}(O)\,\Pi\left[\widehat{y}\right](X)\right\} \\ &= \mathbb{P}_n\left\{I_a\Pi\left[\widehat{y}\right](X)^2\right\} \\ &\leq \mathbb{P}_n\left\{I_a\widehat{y}(O)^2\right\}, \end{aligned} \tag{B.30}$$

where the last equality and the last inequality follow from the fact that

$$\mathbb{P}_n \left[I_a \left\{ \widehat{y}(O) - \Pi\left[\widehat{y}\right](X) \right\} \Pi\left[\widehat{y}\right](X) \right] = 0$$

by projection's properties. But,

$$\mathbb{P}_n \left\{ I_a \widehat{y}(O)^2 \right\} \lesssim_P E\left[\mathbb{P}_n \left\{ I_a \widehat{y}(O)^2 \right\} \middle| \mathcal{S}_n^* \right] \tag{B.31}$$

by Corollary 1 and

$$E\left[\mathbb{P}_n \left\{ I_a \widehat{y}(O)^2 \right\} \middle| \mathcal{S}_n^* \right] = E\left\{ I_a \widehat{y}(O)^2 \middle| \mathcal{S}_n^* \right\}. \tag{B.32}$$

Hence, from (B.30), (B.31) and (B.32), we arrive at

$$\left\| \widehat{\beta} \right\|^2 \lesssim_P E\left\{ I_a \widehat{y}(O)^2 \middle| \mathcal{S}_n^* \right\}$$

as we wanted to show.

To see (c), note that

$$\begin{aligned} \left\| \widehat{\beta} \right\|^2 &= \left\| \widehat{Q}_a^{-1} \mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p(X) \right\} \right\|^2 \\ &= \mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p(X) \right\}' \widehat{Q}_a^{-2} \mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p(X) \right\} \\ &\leq \lambda_{\max}\left(\widehat{Q}_a^{-2} \right) \left\| \mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p(X) \right\} \right\|^2. \end{aligned}$$

But $\lambda_{\max}\left(\widehat{Q}_a^{-2} \right) = \lambda_{\max}\left(\widehat{Q}_a^{-1} \right)^2$, which is bounded in probability as shown above. Hence,

$$\left\| \widehat{\beta} \right\|^2 \lesssim_P \left\| \mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p(X) \right\} \right\|^2,$$

thus concluding the proof. ■

Proof of Lemma 22. First note that

$$E\left[I_a \left\{ \Pi\left[\widehat{y}\right](X) \right\}^2 \middle| \mathcal{S}_n, \mathcal{S}_n^* \right] \lesssim_P \left\| \mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p(X) \right\} \right\|^2$$

by Lemma 21 and assumptions 1 to 3 of the lemma. Now

$$\begin{aligned} \left\| \mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p(X) \right\} \right\|^2 &= \sum_{j=1}^m \left[\mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p_j(X) \right\} \right]^2 \\ &\lesssim_P E\left\{ \sum_{j=1}^m \left[\mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p_j(X) \right\} \right]^2 \middle| \mathcal{S}_n^* \right\} \\ &= \sum_{j=1}^m E\left\{ \left[\mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p_j(X) \right\} \right]^2 \middle| \mathcal{S}_n^* \right\} \end{aligned}$$

where the "$\lesssim_P$" in the second row follows from Corollary 1. Then

$$E\left[I_a \left\{ \Pi\left[\widehat{y}\right](X) \right\}^2 \middle| \mathcal{S}_n, \mathcal{S}_n^* \right] \lesssim_P \sum_{j=1}^m E\left\{ \left[\mathbb{P}_n \left\{ I_a \widehat{y}(O)\, p_j(X) \right\} \right]^2 \middle| \mathcal{S}_n^* \right\}. \tag{B.33}$$

But

$$E\left\{\mathbb{P}_n\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\right\}\middle|\mathcal{S}_n^*\right\}=E\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\middle|\mathcal{S}_n^*\right\}$$

and

$$E\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\middle|\mathcal{S}_n^*\right\}=0 \tag{B.34}$$

by assumption 4 of the lemma. Hence,

$$\begin{aligned}E\left\{\left[\mathbb{P}_n\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\right\}\right]^2\middle|\mathcal{S}_n^*\right\}&=Var\left\{\mathbb{P}_n\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\right\}\middle|\mathcal{S}_n^*\right\}\\&=\frac{1}{n}Var\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\middle|\mathcal{S}_n^*\right\}\\&=\frac{1}{n}E\left[\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\right\}^2\middle|\mathcal{S}_n^*\right]\end{aligned}$$

where the last equivalence follows from (B.34). Then,

$$\begin{aligned}\sum_{j=1}^{m}E\left\{\left[\mathbb{P}_n\left\{I_a\widehat{y}\left(O\right)p_j\left(X\right)\right\}\right]^2\middle|\mathcal{S}_n^*\right\}&=\sum_{j=1}^{m}\frac{1}{n}E\left[\left\{I_{\{a\}}\left(A_i\right)\widehat{y}\left(O_i\right)p_j\left(X_i\right)\right\}^2\middle|\mathcal{S}_n^*\right]\\&=\frac{1}{n}E\left[\sum_{j=1}^{m}\left\{I_{\{a\}}\left(A_i\right)\widehat{y}\left(O_i\right)p_j\left(X_i\right)\right\}^2\middle|\mathcal{S}_n^*\right]\\&=\frac{1}{n}E\left[I_{\{a\}}\left(A_i\right)\widehat{y}\left(O_i\right)^2\left\|p\left(X_i\right)\right\|^2\middle|\mathcal{S}_n^*\right]\\&\lesssim\frac{m}{n}E\left[I_{\{a\}}\left(A_i\right)\widehat{y}\left(O_i\right)^2\middle|\mathcal{S}_n^*\right]\\&\lesssim_P\frac{m}{n}.\end{aligned} \tag{B.35}$$

where the "$\lesssim$" in the 4^{th} row follows from assumption 2 of the lemma and the "$\lesssim_P$" in the last row follows by assumption 5 of the lemma. Finally, equations (B.33) and (B.35) imply that

$$E\left[I_a\left\{\Pi\left[\widehat{y}\right]\left(X\right)\right\}^2\middle|\mathcal{S}_n,\mathcal{S}_n^*\right]\lesssim_P\frac{m}{n}$$

as we wanted to show. ■

B.6.2 Proofs of Theorems 3 to 6

Throughout this section, we will use repeatedly the fact that for any fixed $p\in\mathbb{N}$ and any sequence $\{a_n\}_{n\in\mathbb{N}}$ of $p\times 1$ vectors with $a_n\equiv(a_{n,1},\ldots,a_{n,p})$, it holds that

$$\left(\sum_{i=1}^{p}a_{n,i}\right)^2\lesssim\sum_{i=1}^{p}a_{n,i}^2.$$

This is because, as can be easily shown by induction, $\left(\sum_{i=1}^{p}v_i\right)^2\leq 2^p\sum_{i=1}^{p}v_i^2$ for every $p\in\mathbb{N}$ and $(v_1,\ldots,v_p)\in\mathbb{R}^p$. Likewise, we will use the fact that $\lim_{n\to\infty}n^{-a}\left(\log n^b\right)^c=0$ for any $a>0,b>0$ and $c>0$, which follows straightforwardly by L'Hôpital's rule.

To prove Theorems 3 to 6 we need first to show some results. Like the Theorems 3 to 6, each of the following results assumes a specific subset of the conditions defined in Subsection 2.7.2. We recall them here, for easy of reference.

Condition Hölder(k) $\eta_k(\cdot)$ lies in a Hölder ball $\mathcal{H}\left(\overline{\mathcal{L}}_k; s_k, \rho_k\right)$ with $\rho_k < \infty$ and known smoothness order $s_k > 0$.

Condition R(k) Assumptions 2 - 5 of Lemma 17 in Appendix B.5 are verified with $\overline{L}_k$ in the place of X, $\eta_{k+1}\left(\overline{L}_{k+1}\right)$ in the place of Y, $\eta_k(\cdot)$ in the place of $g(\cdot)$, the distribution of $\overline{L}_{k+1}\big|\,\overline{A}_k = \overline{a}_k^*$ in the place of the distribution of (Y, X'), $\phi_k\left(\overline{l}_k\right)$ in the place of $p(x)$, $\mathcal{H}\left(\overline{\mathcal{L}}_k; s_k, \rho_k\right)$ in the place of $\mathcal{G}$ and $\gamma_k \equiv s_k / \dim\left(\overline{L}_k\right)$ in the place of γ.

Condition B(k) there exists $\xi > 0$ such that $h_k\left(\overline{l}_k\right) > \xi$ for all $\overline{l}_k$.

Condition Hconvergence(k) $\left\|\widehat{h}_k - h_k\right\|_\infty = o_p(1)$.

Condition HrateInf(k) $\left\|\widehat{h}_k - h_k\right\|_\infty = O_p(\alpha_{k,n})$ for some sequence $\alpha_{k,n}$ converging to 0 as n goes to ∞.

Condition HrateL2(k) $\sqrt{E_p\left\{\overline{I}_k\left(\widehat{h}_k - h_k\right)^2 \middle| \mathcal{N}\right\}} = O_p(\beta_{k,n})$ for some sequence $\beta_{k,n}$ converging to 0 as n goes to ∞.

Also recall that, whenever Condition Hölder(k) holds and d_k denotes the dimension of the vector $\overline{L}_k$, we let

$$\gamma_k \equiv \frac{s_k}{d_k} \text{ and } r_k \equiv \frac{\gamma_k}{2\gamma_k + 1}.$$

In addition, recall that m_k is the dimension of the vector $\phi_k\left(\overline{L}_k\right)$ used to construct the series estimators $\widehat{\eta}_k$ and $\widehat{\eta}_{k,MR}$ (see definition (2.42)). Finally, $\eta_{k,DR}$ and $\eta_{k,MR}$ are as defined in (b) and (d) of Subsection 2.7.1 for the special linear operator $\Pi^k[\cdot]$ defined in (2.42) and with $h^\dagger$ replaced by the estimator $\widehat{h}$ used to compute $\widehat{\theta}^u$ and $\widehat{\theta}^u_{MR}$.

Next, we introduce and show the results used in the proofs of Theorems 3 to 6.

Lemma 28 *Suppose that, for a given $k \in [K]$,*

1. *Condition B(k) holds, and*
2. *Condition Hconvergence(k) holds.*

Then,

a) $\left\|\left(\widehat{h}_k\right)^{-2}\right\|_\infty = O_p(1)$ *and*

b) if, in addition, Condition HrateInf(k) holds, then

$$E_p\left\{\overline{I}_k\left(\frac{1}{h_k} - \frac{1}{\widehat{h}_k}\right)^2 u(O,\mathcal{N}) \middle| \mathcal{N}\right\} = O_p\left(\alpha_{k,n}^2\right) E_p\left\{u(O,\mathcal{N})|\mathcal{N}\right\}$$

for any nonnegative real-valued function $u(\cdot,\cdot)$.

Proof. Fact a) follows straightforwardly by Lemma 19 in Appendix B.6.1 and the assumptions of the lemma.

To see b), note that, for $k \in [K]$,

$$\begin{aligned}\overline{I}_k \left(\frac{1}{h_k} - \frac{1}{\widehat{h}_k} \right)^2 &= \overline{I}_k \left(\frac{1}{h_k \widehat{h}_k} \right)^2 \left(\widehat{h}_k - h_k \right)^2 \\ &\leq \frac{1}{\xi^2} \left\| \widehat{h}_k^{-2} \right\|_\infty \left\| \widehat{h}_k - h_k \right\|_\infty^2\end{aligned}$$

by Condition B(k). Thus, since $u(\cdot,\cdot)$ is a nonnegative real-valued function,

$$0 \leq E_p \left\{ I_k \left(\frac{1}{h_k} - \frac{1}{\widehat{h}_k} \right)^2 u(O,\mathcal{N}) \middle| \mathcal{N} \right\} \lesssim \left\| \widehat{h}_k^{-2} \right\|_\infty \left\| \widehat{h}_k - h_k \right\|_\infty^2 E_p \left\{ u(O,\mathcal{N}) | \mathcal{N} \right\}.$$

But $\left\| \widehat{h}_k^{-2} \right\|_\infty = O_p(1)$ by part a) of the lemma. Then, if $\left\| \widehat{h}_k - h_k \right\|_\infty = O_p(\alpha_{k,n})$, we arrive at

$$E_p \left\{ I_k \left(\frac{1}{h_k} - \frac{1}{\widehat{h}_k} \right)^2 u(O,\mathcal{N}) \middle| \mathcal{N} \right\} = O_p\left(\alpha_{k,n}^2\right) E_p \left\{ u(O,\mathcal{N}) | \mathcal{N} \right\}.$$

■

Lemma 29 *Suppose that, for each $k \in [K]$,*

1. *Condition B(k) holds,*
2. *Condition Hconvergence(k) holds and*
3. *Condition HrateL2(k) holds.*

Then, for $k \in [K]$,

$$\sqrt{E_p \left[\left\{ \frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{\widehat{h}_k} \right) \right\}^2 \middle| \mathcal{N} \right]} = O_p(\beta_{k,n}).$$

Proof. Note that, for $k \in [K]$,

$$\begin{aligned}\left\{ \frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{\widehat{h}_k} \right) \right\}^2 &= \left(\frac{\overline{I}_k}{\widehat{\pi}^k h_k} \right)^2 \left(\widehat{h}_k - h_k \right)^2 \\ &\leq \frac{1}{\xi^2} \left\| \left(\widehat{\pi}^k \right)^{-2} \right\|_\infty \overline{I}_k \left(\widehat{h}_k - h_k \right)^2\end{aligned}$$

by Condition B(k). Hence, since $\left\| \left(\widehat{\pi}^k \right)^{-2} \right\|_\infty$ depends only on $\mathcal{N}$,

$$E_p \left[\left\{ \frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{\widehat{h}_k} \right) \right\}^2 \middle| \mathcal{N} \right] \lesssim \left\| \left(\widehat{\pi}^k \right)^{-2} \right\|_\infty E_p \left\{ \overline{I}_k \left(\widehat{h}_k - h_k \right)^2 \middle| \mathcal{N} \right\}.$$

But $\left\|\left(\widehat{\pi}^k\right)^{-2}\right\|_\infty \leq \Pi_{j=1}^k \left\|\left(\widehat{h}_j\right)^{-2}\right\|_\infty = O_p(1)$ by part a) of Lemma 28 and assumptions 1 and 2 of the lemma. Then, assumption 3 implies that

$$\sqrt{E_p\left[\left\{\frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}}\left(\frac{I_k}{h_k}-\frac{I_k}{\widehat{h}_k}\right)\right\}^2\middle|\mathcal{N}\right]} = O_p\left(\beta_{k,n}\right)$$

as we wanted to show. ■

The following lemma follows immediately from Lemma 17 in Appendix B.5.

Lemma 30 *Suppose that, for each* $k \in [K]$,

1. *Condition Hölder*(k) *holds and*
2. *Condition* $R(k)$ *holds.*

If $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ *for* $k \in [K]$, *then*

$$\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right]} = O_p\left(n^{-r_k}\right).$$

Lemma 31 *Assume that* $K=2$ *and suppose that*

1. *Condition Hölder*(k) *holds for* $k=1,2$,
2. *Condition* $R(k)$ *holds for* $k=1,2$,
3. *Condition* $B(k)$ *holds for* $k=2$, *and*
4. *Condition Hconvergence*(k) *holds for* $k=2$.

If $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ *for* $k=1,2$, *then*

$$\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,MR}-\eta_k\right)^2\middle|\mathcal{N}\right]} = O_p\left(n^{-r_k}\right).$$

Proof. Recall that

$$\eta_{2,MR} = \eta_{2,DR}$$

and

$$\eta_{1,MR}(\cdot) = \eta_{1,DR}(\cdot) + \Pi^1[\varepsilon_2](\cdot)$$

with

$$\varepsilon_2(\cdot) \equiv q_2\left(\cdot,\cdot;\widehat{h}_2,\widehat{\eta}_{2,MR}\right) - E_p\left\{Q_2\left(\widehat{h}_2,\widehat{\eta}_{2,MR}\right)\middle|\, A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N}^2\right\}.$$

Hence,

$$E_p\left[\overline{I}_2\left(\eta_{2,MR}-\eta_2\right)^2\middle|\mathcal{N}\right] = E_p\left[\overline{I}_2\left(\eta_{2,DR}-\eta_2\right)^2\middle|\mathcal{N}\right]$$

and

$$E_p\left[I_1\left(\eta_{1,MR}-\eta_1\right)^2\middle|\mathcal{N}\right] \lesssim E_p\left[I_1\left(\eta_{1,DR}-\eta_1\right)^2\middle|\mathcal{N}\right] + E_p\left[I_1\Pi^1[\varepsilon_2]^2\middle|\mathcal{N}\right].$$

Also, from assumptions 1 and 2 of the lemma and Lemma 30, we have hat

$$E_p\left[\overline{I}_k\left(\eta_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right]\lesssim_P\left(n^{-r_k}\right)^2, k=1,2.$$

Hence, we would arrive at the desired result if we show that

$$E_p\left[I_1\Pi^1\left[\varepsilon_2\right]^2\middle|\mathcal{N}\right]\lesssim_P\left(n^{-r_1}\right)^2. \tag{B.36}$$

Note that, since we are taking $m_1\asymp n^{\frac{1}{2\gamma_1+1}}$, we have that $\frac{m_1}{n}\asymp n^{-\frac{2\gamma_1}{2\gamma_1+1}}\asymp n^{-\frac{2\gamma_1}{2\gamma_1+1}}=\left(n^{-r_1}\right)^2$. Then, to show (B.36), it suffices to show that

$$E_p\left[I_1\Pi^1\left[\varepsilon_2\right]^2\middle|\mathcal{N}\right]\lesssim_P\frac{m_1}{n}. \tag{B.37}$$

Note that $E_p\left[I_1\Pi^1\left[\varepsilon_2\right]^2\middle|\mathcal{N}\right]=E_p\left[I_1\Pi^1\left[\varepsilon_2\right]^2\middle|\mathcal{N}^1,\mathcal{N}^2\right]$ and that ε_2 is function of the observed data O and $\mathcal{N}^2$. Also, note that assumption 2 of the lemma imply that

(i) the eigenvalues of $E_p\left\{I_1\phi_1\left(L_1\right)\phi_1\left(L_1\right)'\right\}$ are bounded above and bellow away from zero, and

(ii) $\left\|\phi_1\right\|_\infty\leq\sqrt{m_1}$.

Then, assumptions 1 and 2 of Lemma 22 in Appendix B.6.1 hold with $\mathcal{S}_n,\mathcal{S}_n^*,X,A,a,p\left(\cdot\right)$ and $\widehat{y}\left(O\right)$ replaced by $\mathcal{N}^1,\mathcal{N}^2,L_1,A_1,a_1^*,\phi_1\left(\cdot\right)$ and ε_2 respectively. Furthermore, the fact that $m_1\asymp n^{\frac{1}{2\gamma_1+1}}$ implies that the assumption 3 of that lemma also holds. Note that here we are using the fact that each $\widehat{\eta}_{k,MR}, k=1,2$, is computed from a different subsample of $\mathcal{N}$, because we are strongly using the fact that $\mathcal{N}^1$ and $\mathcal{N}^2$ are independent datasets, which is one of the assumptions of Lemma 22. Then, we would arrive at (B.37) if we show that

(I) $E_p\left\{I_1\varepsilon_2\middle|L_1,\mathcal{N}^2\right\}=0$ and

(II) $E_p\left\{I_1\varepsilon_2^2\middle|\mathcal{N}^2\right\}\lesssim_P 1$.

This is because, (I) and (II) imply that assumptions 4 and 5 of Lemma 22 also hold and, hence, we would conclude (B.37) from that lemma.

To see fact (I), note that, by definition, $E_p\left\{\varepsilon_2\middle|A_1=a_1^*,\overline{L}_2,\mathcal{N}^2\right\}=0$. Then,

$$E_p\left\{I_1\varepsilon_2\middle|\overline{L}_2,\mathcal{N}^2\right\}=E_p\left\{\varepsilon_2\middle|A_1=a_1^*,\overline{L}_2,\mathcal{N}^2\right\}P\left(A_1=a_1^*\middle|\overline{L}_2\right)=0$$

and, hence,

$$E_p\left\{I_1\varepsilon_2\middle|L_1,\mathcal{N}^2\right\}=0.$$

To see fact (II) note that, since

$$Q_2\left(\widehat{h}_2,\widehat{\eta}_{2,MR}\right)=\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)+\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}$$

and $\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)$ depends only on $\overline{L}_2$ and $\mathcal{N}^2$, then

$$\varepsilon_2\equiv\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}-E_p\left[\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\middle|A_1=a_1^*,\overline{L}_2,\mathcal{N}^2\right]$$

Thus, to prove (II), it suffices to show that

(II.a) $E_p\left\{ I_1\left[\frac{I_2}{\widehat{h}_2(\overline{L}_2)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\right]^2\middle|\mathcal{N}^2\right\}\lesssim_P 1$ and

(II.b) $E_p\left\{ I_1 E_p\left[\frac{I_2}{\widehat{h}_2(\overline{L}_2)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\middle| A_1=a_1^*,\overline{L}_2,\mathcal{N}^2\right]^2\middle|\mathcal{N}^2\right\}\lesssim_P 1.$

We start by fact (II.a). Note that

$$I_1\left[\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\right]^2\leq\left\|\widehat{h}_2^{-2}\right\|_\infty I_2\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}^2$$

Hence,

$$\begin{aligned}
&E_p\left\{ I_1\left[\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\right]^2\middle|\mathcal{N}^2\right\}\\
&\leq\left\|\widehat{h}_2^{-2}\right\|_\infty E_p\left[\overline{I}_2\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2\right]\\
&\lesssim_P E_p\left[\overline{I}_2\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2\right]\\
&\lesssim E_p\left[\overline{I}_2\left\{\kappa\left(\overline{L}_3\right)-\eta_2\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2\right]+E_p\left[\overline{I}_2\left\{\eta_2\left(\overline{L}_2\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2\right]
\end{aligned}$$

where the bound in the third row follows from part a) of Lemma 28 and assumptions 3 and 4. But,

$$E_p\left[\overline{I}_2\left\{\kappa\left(\overline{L}_3\right)-\eta_2\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2\right]\lesssim 1$$

by assumption 2 of the lemma and

$$E_p\left[\overline{I}_2\left\{\eta_2\left(\overline{L}_2\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2\right]=o_p\left(1\right)$$

by Lemma 30 and by the fact that $\widehat{\eta}_{2,MR}=\eta_{2,DR}$. Then

$$E_p\left\{ I_1\left[\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\right]^2\middle|\mathcal{N}^2\right\}\lesssim_P 1$$

as we wanted to show.

Fact (II.b) follows immediately from fact (II.a), after noticing that

$$
\begin{aligned}
&E_p\left\{ I_1 E_p\left[\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\middle| A_1=a_1^*,\overline{L}_2,\mathcal{N}^2\right]^2\middle|\mathcal{N}^2\right\}\\
&\leq E_p\left[I_1 E_p\left\{ \left[\frac{I_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\right]^2\middle| A_1=a_1^*,\overline{L}_2,\mathcal{N}^2\right\}\middle|\mathcal{N}^2\right]\\
&= E_p\left[E_p\left\{ \left[\frac{\overline{I}_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\right]^2\middle| A_1,\overline{L}_2,\mathcal{N}^2\right\}\middle|\mathcal{N}^2\right]\\
&= E_p\left\{ \left[\frac{\overline{I}_2}{\widehat{h}_2\left(\overline{L}_2\right)}\left\{\kappa\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}\right]^2\middle|\mathcal{N}^2\right\}
\end{aligned}
$$

■

Lemma 32 *Assume that $K=3$ and suppose that*

1. *Condition Hölder(k) holds for $k=1,2,3$,*
2. *Condition R(k) holds for $k=1,2,3$,*
3. *Condition B(k) holds for $k=2,3$, and*
4. *Condition Hconvergence(k) holds $k=2,3$.*

If $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k=1,2,3$, then

$$\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,MR}-\eta_k\right)^2\middle|\mathcal{N}\right]}=O_p\left(n^{-r_k}\right).$$

Proof. First, note that assumption 2 of the lemma imply that, for each $k=1,2,3$,

$$\text{the eigenvalues of } E_p\left\{\overline{I}_k\phi_k\left(\overline{L}_k\right)\phi_k\left(\overline{L}_k\right)'\right\} \text{ are bounded above and bellow away from zero,} \tag{B.38}$$

$$\left\|\phi_k\right\|_\infty \lesssim \sqrt{m_k}, \tag{B.39}$$

$$E_p\left[\overline{I}_2\left\{\eta_3\left(\overline{L}_3\right)-\eta_2\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2\right]\lesssim 1 \tag{B.40}$$

and

$$E_p\left[\overline{I}_3\left\{\kappa\left(\overline{L}_4\right)-\eta_3\left(\overline{L}_3\right)\right\}^2\right]\lesssim 1. \tag{B.41}$$

Recall that

$$\eta_{3,MR}=\eta_{3,DR}$$

and, for $k=1,2$,

$$\eta_{k,MR}=\eta_{k,DR}+\Pi^k\left[\varepsilon_{k+1}\right]$$

with

$$\varepsilon_{k+1}(\cdot) \equiv q_{k+1}\left(\cdot,\cdot;\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right) - E_p\left\{Q_{k+1}\left(\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right)\middle|\,\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1}=\cdot,\underline{\mathcal{N}}^{k+1}\right\}$$

where recall that $\underline{\mathcal{N}}^{k+1} \equiv \cup_{j=k+1}^{3}\mathcal{N}^j$.

Hence,

$$E_p\left[\overline{I}_3\left(\eta_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}\right] = E_p\left[\overline{I}_3\left(\eta_{3,DR}-\eta_3\right)^2\middle|\mathcal{N}\right]$$

and, for $k=1,2$,

$$E_p\left[\overline{I}_k\left(\eta_{k,MR}-\eta_k\right)^2\middle|\mathcal{N}\right] \lesssim E_p\left[\overline{I}_k\left(\eta_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right] + E_p\left[\overline{I}_k\Pi^k\left[\varepsilon_{k+1}\right]^2\middle|\mathcal{N}\right].$$

Also, from Lemma 30 and assumptions 1 and 2 of the lemma, we have hat

$$E_p\left[\overline{I}_k\left(\eta_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right] \lesssim_P \left(n^{-r_k}\right)^2, k=1,2,3. \tag{B.42}$$

Hence, we would arrive at the desired result if we show that

$$E_p\left[\overline{I}_k\Pi^k\left[\varepsilon_{k+1}\right]^2\middle|\mathcal{N}\right] \lesssim_P \left(n^{-r_k}\right)^2 \text{ for } k=1,2. \tag{B.43}$$

As in the case $K=2$, since we are taking $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$, we have that $\frac{m_k}{n} \asymp n^{-\frac{2\gamma_k}{2\gamma_k+1}} \asymp n^{-\frac{2\gamma_k}{2\gamma_k+1}} = \left(n^{-r_k}\right)^2$ and, hence, to see (B.43) it suffices to show that

$$E_p\left[\overline{I}_k\Pi^k\left[\varepsilon_{k+1}\right]^2\middle|\mathcal{N}\right] \lesssim_P \frac{m_k}{n} \text{ for } k=1,2. \tag{B.44}$$

To see (B.44), we proceed as in the proof of Lemma 31. Specifically, we first note that, for $k=1,2$, (i) $E_p\left[\overline{I}_k\Pi^k\left[\varepsilon_{k+1}\right]^2\middle|\mathcal{N}\right] = E_p\left[\overline{I}_k\Pi^k\left[\varepsilon_{k+1}\right]^2\middle|\mathcal{N}^k,\underline{\mathcal{N}}^{k+1}\right]$ and (ii) $\varepsilon_{k+1}(O)$ depends on the observed data O and $\underline{\mathcal{N}}^{k+1}$. Hence, to prove (B.44), we can apply, for each $k=1,2$, Lemma 22 of Appendix B.6.1, with $\mathcal{S}_n, \mathcal{S}_n^*, X, A, a, p(\cdot)$ and $\widehat{y}(O)$ replaced by $\mathcal{N}^k, \underline{\mathcal{N}}^{k+1}, \overline{L}_k, \overline{A}_k, \overline{a}_k^*, \phi_k(\cdot)$ and ε_{k+1} respectively. Facts (B.38) and (B.39) and the fact that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ imply that all assumptions 1 - 3 of that lemma are verified with the replacements mentioned above, for $k=1,2$. Also, for $k=1,2$, assumptions 4 and 5 of that lemma hold if

(I) $E_p\left\{\overline{I}_k\varepsilon_{k+1}\middle|\,\overline{A}_k,\overline{L}_k,\underline{\mathcal{N}}^{k+1}\right\}=0$ and

(II) $E_p\left\{\overline{I}_k\varepsilon_{k+1}^2\middle|\,\underline{\mathcal{N}}^{k+1}\right\} \lesssim_P 1$.

Fact (I) follows again from the fact that, for $k=1,2$, $E_p\left\{\varepsilon_{k+1}\middle|\,\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}=0$ by definition.

To see (II) note that, since

$$Q_{k+1}\left(\widehat{\underline{h}}_{k+1},\widehat{\underline{\eta}}_{k+1,MR}\right) = \widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right) + \sum_{j=k+1}^{3}\frac{\overline{I}_k^j}{\widehat{\pi}_{k+1}^j}\left\{\widehat{\eta}_{j+1,MR}\left(\overline{L}_{j+1}\right)-\widehat{\eta}_{j,MR}\left(\overline{L}_j\right)\right\}$$

and $\widehat{\eta}_{k+1,MR}\left(\overline{L}_{k+1}\right)$ depends only on $\overline{L}_{k+1}$ and $\underline{\mathcal{N}}^{k+1}$, then

$$\varepsilon_{k+1}(O) \equiv u_{k+1} - v_{k+1}$$

with

$$u_{k+1} \equiv \sum_{j=k+1}^{3} \frac{\overline{I}_k^j}{\widehat{\pi}_{k+1}^j} \left\{ \widehat{\eta}_{j+1,MR}\left(\overline{L}_{j+1}\right) - \widehat{\eta}_{j,MR}\left(\overline{L}_j\right) \right\}$$

and

$$v_{k+1} \equiv E_p\left\{ u_{k+1} \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_{k+1}, \underline{\mathcal{N}}^{k+1} \right\}.$$

Then, to prove (II), it suffices to show that, for $k = 1, 2$,

(II.a) $E_p\left\{ \overline{I}_k u_{k+1}^2 \middle| \underline{\mathcal{N}}^{k+1} \right\} \lesssim_P 1$ and

(II.b) $E_p\left\{ \overline{I}_k v_{k+1}^2 \middle| \underline{\mathcal{N}}^{k+1} \right\} \lesssim_P 1$.

We start by fact (II.a). Note that

$$\begin{aligned} \overline{I}_k \left(u_{k+1}\right)^2 &\lesssim \overline{I}_k \sum_{j=k+1}^{3} \left(\frac{\overline{I}_k^j}{\widehat{\pi}_{k+1}^j} \right)^2 \left\{ \widehat{\eta}_{j+1,MR}\left(\overline{L}_{j+1}\right) - \widehat{\eta}_{j,MR}\left(\overline{L}_j\right) \right\}^2 \\ &\leq \sum_{j=k+1}^{3} \left\| \left(\widehat{\pi}_{k+1}^j\right)^{-2} \right\|_\infty \overline{I}_j \left\{ \widehat{\eta}_{j+1,MR}\left(\overline{L}_{j+1}\right) - \widehat{\eta}_{j,MR}\left(\overline{L}_j\right) \right\}^2. \end{aligned}$$

Hence,

$$\begin{aligned} E_p\left\{ \overline{I}_k u_{k+1}^2 \middle| \underline{\mathcal{N}}^{k+1} \right\} &\lesssim \sum_{j=k+1}^{3} \left\| \left(\widehat{\pi}_{k+1}^j\right)^{-2} \right\|_\infty E_p\left[\overline{I}_j \left\{ \widehat{\eta}_{j+1,MR}\left(\overline{L}_{j+1}\right) - \widehat{\eta}_{j,MR}\left(\overline{L}_j\right) \right\}^2 \middle| \underline{\mathcal{N}}^{k+1} \right] \\ &\lesssim_P \sum_{j=k+1}^{3} E_p\left[\overline{I}_j \left\{ \widehat{\eta}_{j+1,MR}\left(\overline{L}_{j+1}\right) - \widehat{\eta}_{j,MR}\left(\overline{L}_j\right) \right\}^2 \middle| \underline{\mathcal{N}}^{k+1} \right]. \end{aligned}$$

since, for $k = 1, 2$ and $j \geq k+1$, $\left\| \left(\widehat{\pi}_{k+1}^j\right)^{-2} \right\|_\infty \lesssim_P 1$ by part a) of Lemma 28 and assumptions 3 and 4 of the lemma. Then, to prove (II.a), it suffices to show that

$$E_p\left[\overline{I}_j \left\{ \widehat{\eta}_{j+1,MR}\left(\overline{L}_{j+1}\right) - \widehat{\eta}_{j,MR}\left(\overline{L}_j\right) \right\}^2 \middle| \underline{\mathcal{N}}^{k+1} \right] \lesssim_P 1$$

for $j \geq k+1$ and $k = 1, 2$. That is, it suffices to show that

$$E_p\left[\overline{I}_3 \left\{ \kappa\left(\overline{L}_4\right) - \widehat{\eta}_{3,MR}\left(\overline{L}_3\right) \right\}^2 \middle| \mathcal{N}^3 \right] \lesssim_P 1 \tag{B.45}$$

and

$$E_p\left[\overline{I}_2 \left\{ \widehat{\eta}_{3,MR}\left(\overline{L}_3\right) - \widehat{\eta}_{2,MR}\left(\overline{L}_2\right) \right\}^2 \middle| \mathcal{N}^2 \cup \mathcal{N}^3 \right] \lesssim_P 1. \tag{B.46}$$

To see (B.45) , note that

$$\begin{aligned} E_p\left[\overline{I}_3 \left\{ \kappa\left(\overline{L}_4\right) - \widehat{\eta}_{3,MR}\left(\overline{L}_3\right) \right\}^2 \middle| \mathcal{N}^3 \right] &\lesssim E_p\left[\overline{I}_3 \left\{ \kappa\left(\overline{L}_4\right) - \eta_3\left(\overline{L}_3\right) \right\}^2 \right] \\ &\quad + E_p\left[\overline{I}_3 \left\{ \eta_3\left(\overline{L}_3\right) - \widehat{\eta}_{3,MR}\left(\overline{L}_3\right) \right\}^2 \middle| \mathcal{N}^3 \right]. \end{aligned}$$

But

$$E_p\left[\overline{I}_3\left\{\kappa\left(\overline{L}_4\right)-\eta_3\left(\overline{L}_3\right)\right\}^2\right]\lesssim 1$$

by fact B.41 and

$$E_p\left[\overline{I}_3\left\{\eta_3\left(\overline{L}_3\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}^2\middle|\mathcal{N}^3\right]=o_p\left(1\right)$$

again by Lemma 30 and from the fact that $\widehat{\eta}_{3,MR}=\eta_{3,DR}$.

To see (B.46) note that

$$\begin{aligned}\widehat{\eta}_{2,MR}&=\Pi^2\left[Q_3\left(\widehat{h}_3,\widehat{\eta}_{3,MR}\right)\right]\\&=\Pi^2\left[\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)+\frac{I_3}{\widehat{h}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\right]\\&=\Pi^2\left[\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right]+\Pi^2\left[\frac{I_3}{\widehat{h}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\right].\end{aligned}$$

Then,

$$\begin{aligned}&E_p\left[\overline{I}_2\left\{\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)-\widehat{\eta}_{2,MR}\left(\overline{L}_2\right)\right\}^2\middle|\mathcal{N}^2,\mathcal{N}^3\right]\\&\lesssim E_p\left[\overline{I}_2\left\{\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)-\Pi^2\left[\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right]\right\}^2\middle|\mathcal{N}^2,\mathcal{N}^3\right]\\&+E_p\left[\overline{I}_2\Pi^2\left[\frac{I_3}{\widehat{h}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\right]^2\middle|\mathcal{N}^3\right]\end{aligned}$$

Hence, to prove (B.46) , it suffices to show that

$$E_p\left[\overline{I}_2\left\{\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)-\Pi^2\left[\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right]\right\}^2\middle|\mathcal{N}^2,\mathcal{N}^3\right]\lesssim_P 1 \tag{B.47}$$

and

$$E_p\left\{\overline{I}_2\Pi^2\left[\frac{I_3}{\widehat{h}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\right]^2\middle|\mathcal{N}^2,\mathcal{N}^3\right\}\lesssim_P 1. \tag{B.48}$$

To see (B.48) , we will apply Lemma 21 of Appendix B.6.1, replacing $\mathcal{S}_n,\mathcal{S}_n^*,X,A,a$ and $\widehat{y}\left(O\right)$ by $\mathcal{N}^2,\mathcal{N}^3,\overline{L}_2,\overline{A}_2,\overline{a}_2^*$ and $\frac{I_3}{\widehat{h}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}$. The assumptions of that lemma are verified with the corresponding replacements from the facts (B.38) and (B.39) for $k=2$ and the fact that

$m_2 \asymp n^{\frac{1}{2\gamma_2+1}}$. Thus,

$$
\begin{aligned}
& E_p\left\{\overline{I}_2\Pi^2\left[\frac{I_3}{\widehat{\overline{h}}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\right]^2\middle|\mathcal{N}^2\cup\mathcal{N}^3\right\} \\
& \lesssim_P E_p\left\{\overline{I}_2\left[\frac{I_3}{\widehat{\overline{h}}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\right]^2\middle|\mathcal{N}^3\right\} \\
& = E_p\left\{\overline{I}_3\left[\frac{1}{\widehat{\overline{h}}_3}\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}\right]^2\middle|\mathcal{N}^3\right\} \\
& \leq \left\|\widehat{\overline{h}}_3^{-2}\right\|_\infty E_p\left[\overline{I}_3\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}^2\middle|\mathcal{N}^3\right] \\
& \lesssim_P E_p\left[\overline{I}_3\left\{\kappa\left(\overline{L}_4\right)-\widehat{\eta}_{3,MR}\left(\overline{L}_3\right)\right\}^2\middle|\mathcal{N}^3\right] \\
& \lesssim_P 1,
\end{aligned}
$$

where the bounds in the last two rows follow from Lemma 28 and by fact (B.45) respectively.

To see (B.47) , note that

$$
\begin{aligned}
& E_p\left[\overline{I}_2\left\{\widehat{\eta}_{3,MR}-\Pi^2\left[\widehat{\eta}_{3,MR}\right]\right\}^2\middle|\mathcal{N}^2\cup\mathcal{N}^3\right] \\
& = E_p\left[\overline{I}_2\left\{\widehat{\eta}_{3,MR}-\eta_3+\eta_3-\Pi^2\left[\widehat{\eta}_{3,MR}-\eta_3\right]-\Pi^2\left[\eta_3\right]\right\}^2\middle|\mathcal{N}^2\cup\mathcal{N}^3\right] \\
& \lesssim E_p\left[\overline{I}_2\left(\widehat{\eta}_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}^3\right]+E_p\left[\overline{I}_2\left\{\Pi^2\left[\eta_3\right]-\eta_3\right\}^2\middle|\mathcal{N}^2\right] \\
& + E_p\left[\overline{I}_2\Pi^2\left[\widehat{\eta}_{3,MR}-\eta_3\right]^2\middle|\mathcal{N}^2\cup\mathcal{N}^3\right]
\end{aligned}
$$

Also, note that

$$
E_p\left[\overline{I}_2\Pi^2\left[\widehat{\eta}_{3,MR}-\eta_3\right]^2\middle|\mathcal{N}^2\cup\mathcal{N}^3\right]\lesssim_P E_p\left[\overline{I}_2\left(\widehat{\eta}_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}^3\right].
$$

It follows from applying again Lemma 21 of Appendix B.6.1 with $\mathcal{S}_n, \mathcal{S}_n^*, X, A, a$ replaced by $\mathcal{N}^2, \mathcal{N}^3, \overline{L}_2, \overline{A}_2, \overline{a}_2^*$ as before, but with $\widehat{y}(O)$ now replaced by $(\widehat{\eta}_{3,MR}-\eta_3)$. Then, to see (B.47) , it suffices to show that

$$
E_p\left[\overline{I}_2\left(\widehat{\eta}_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}^3\right]\lesssim_P 1
$$

and

$$
E_p\left[\overline{I}_2\left\{\Pi^2\left[\eta_3\right]-\eta_3\right\}^2\middle|\mathcal{N}^2\right]\lesssim_P 1.
$$

Now,

$$
E_p\left[\overline{I}_2\left(\widehat{\eta}_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}^3\right]\lesssim E_p\left[\overline{I}_3\left(\widehat{\eta}_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}^3\right]
$$

since $h_3\left(\overline{l}_3\right)>\xi$ for all $\overline{l}_3$ by the positivity assumption 3. But, $\widehat{\eta}_{3,MR}=\eta_{3,MR}=\eta_{3,DR}$, then

$$
E_p\left[\overline{I}_3\left(\widehat{\eta}_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}^3\right]=E_p\left[\overline{I}_3\left(\eta_{3,DR}-\eta_3\right)^2\middle|\mathcal{N}^3\right]\lesssim_P\left(n^{-r_3}\right)^2
$$

by Lemma 30. Hence,

$$E_p\left[\overline{I}_2\left(\widehat{\eta}_{3,MR}-\eta_3\right)^2\middle|\mathcal{N}^3\right]\lesssim_P 1.$$

Likewise, $\Pi^2\left[\eta_3\right]=\eta_{2,DR}$, so that

$$\begin{aligned}E_p\left[\overline{I}_2\left\{\Pi^2\left[\eta_3\right]-\eta_3\right\}^2\middle|\mathcal{N}^2\right] &= E_p\left[\overline{I}_2\left\{\eta_{2,DR}-\eta_3\right\}^2\middle|\mathcal{N}^2\right]\\ &= E_p\left[\overline{I}_2\left\{\eta_{2,DR}-\eta_2+\eta_2-\eta_3\right\}^2\middle|\mathcal{N}^2\right]\\ &\lesssim \underbrace{E_p\left[\overline{I}_2\left\{\eta_{2,DR}-\eta_2\right\}^2\middle|\mathcal{N}^2\right]}_{a}+\underbrace{E_p\left[\overline{I}_2\left\{\eta_3-\eta_2\right\}^2\middle|\mathcal{N}^2\right]}_{b}.\end{aligned}$$

But, $a\lesssim_P\left(n^{-r_2}\right)^2$ as noted above and $b\lesssim 1$ by fact B.40. Then,

$$E_p\left[\overline{I}_2\left\{\Pi^2\left[\eta_3\right]-\eta_3\right\}^2\middle|\mathcal{N}^2\right]\lesssim_P 1,$$

which concludes the proof of fact (B.47) and, hence, of fact (II.a).

Turn to fact (II.b). We must show that $E_p\left\{\overline{I}_k v_{k+1}^2\middle|\underline{\mathcal{N}}^{k+1}\right\}\lesssim_P 1$ for $k=1,2$, where recall that $v_{k+1}=E_p\left\{u_{k+1}\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}$. Then,

$$\begin{aligned}E_p\left\{\overline{I}_k v_{k+1}^2\middle|\underline{\mathcal{N}}^{k+1}\right\} &= E_p\left\{\overline{I}_k\left[E_p\left\{u_{k+1}\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}\right]^2\middle|\underline{\mathcal{N}}^{k+1}\right\}\\ &\le E_p\left[\overline{I}_k E_p\left\{u_{k+1}^2\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}\middle|\underline{\mathcal{N}}^{k+1}\right]\\ &= E_p\left[\overline{I}_k E_p\left\{u_{k+1}^2\middle|\overline{A}_k,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}\middle|\underline{\mathcal{N}}^{k+1}\right]\\ &= E_p\left[E_p\left\{\overline{I}_k u_{k+1}^2\middle|\overline{A}_k,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}\middle|\underline{\mathcal{N}}^{k+1}\right]\\ &= E_p\left\{\overline{I}_k u_{k+1}^2\middle|\underline{\mathcal{N}}^{k+1}\right\}\\ &\lesssim_P 1\end{aligned}$$

by fact (II.a), which concludes the proof. ■

Lemma 33 *Given $k\in[K]$, if Condition $B(k)$ holds then*

$$E_p\left\{\overline{I}_{k-1}u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)\middle|\underline{\mathcal{N}}^k\right\}\lesssim E_p\left\{\overline{I}_k u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)\middle|\underline{\mathcal{N}}^k\right\}$$

for any non negative real-valued function $u\left(\cdot,\cdot\right)$, where $\overline{I}_0\equiv 1$.

Proof. Note that

$$\begin{aligned}E_p\left\{\overline{I}_k u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)\middle|\underline{\mathcal{N}}^k\right\} &= E_p\left[E_p\left\{I_k\overline{I}_{k-1}u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)\middle|\overline{A}_{k-1},\overline{L}_k\underline{\mathcal{N}}^k\right\}\middle|\underline{\mathcal{N}}^k\right]\\ &= E_p\left\{\overline{I}_{k-1}u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)E_p\left(I_k\middle|\overline{A}_{k-1}=\overline{a}_{k-1}^*,\overline{L}_k\right)\middle|\underline{\mathcal{N}}^k\right\}\\ &= E_p\left\{\overline{I}_{k-1}u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)h_k\left(\overline{L}_k\right)\middle|\underline{\mathcal{N}}^k\right\}\\ &\ge \xi E_{gh}\left\{\overline{I}_{k-1}u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)\middle|\underline{\mathcal{N}}^k\right\}.\end{aligned}$$

with $\xi > 0$ by Condition B(k) . Then

$$E_p\left\{\overline{I}_{k-1}u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)\middle|\underline{\mathcal{N}}^k\right\} \leq \xi^{-1}E_p\left\{\overline{I}_k u\left(\overline{L}_k,\underline{\mathcal{N}}^k\right)\middle|\underline{\mathcal{N}}^k\right\},$$

which concludes the proof. ■

Proof of Theorem 3. We want to show that

$$\delta_k^{DR} \lesssim_P \beta_{k,n}n^{-r_k} \text{ for } k = 1,2 \tag{B.49}$$

and

$$\xi_{1,2}^{DR} \lesssim_P \beta_{1,n}n^{-r_2}. \tag{B.50}$$

To see (B.49) note that, for $k = 1, 2$,

$$\begin{aligned}\left|\delta_k^{DR}\right| &= \left|E_p\left\{\frac{\overline{I}_k}{\widehat{\pi}^{k-1}}\left(\frac{1}{h_k}-\frac{1}{\widehat{h}_k}\right)(\eta_{k,DR}-\eta_k)\middle|\mathcal{N}\right\}\right| \\ &\leq \sqrt{E_p\left[\left\{\frac{\overline{I}_k}{\widehat{\pi}^{k-1}}\left(\frac{1}{h_k}-\frac{1}{\widehat{h}_k}\right)\right\}^2\middle|\mathcal{N}\right]}\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right]}\end{aligned}$$

by Cauchy-Schwarz . But, for each $k = 1, 2$,

$$\sqrt{E_p\left[\left\{\frac{\overline{I}_k}{\widehat{\pi}^{k-1}}\left(\frac{1}{h_k}-\frac{1}{\widehat{h}_k}\right)\right\}^2\middle|\mathcal{N}\right]} \lesssim_P \beta_{k,n}$$

by Lemma 29 and Conditions B(k), Hconvergence(k) and HrateL2(k) for $k = 1, 2$. Also, for each $k = 1, 2$,

$$\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right]} \lesssim_P n^{-r_k}$$

by Lemma 30, Conditions Hölder(k) and R(k) and the fact that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k = 1, 2$. Thus, (B.49) holds.

To see (B.50) , we apply again Cauchy-Schwarz and write

$$\begin{aligned}\left|\xi_{1,2}^{DR}\right| &\equiv \left|E_p\left\{I_1\left(\frac{1}{h_1}-\frac{1}{\widehat{h}_1}\right)\Pi^1\left[\eta_{2,DR}-\eta_2\right]\middle|\mathcal{N}\right\}\right| \\ &\leq \sqrt{E_p\left[\left\{I_1\left(\frac{1}{h_1}-\frac{1}{\widehat{h}_1}\right)\right\}^2\middle|\mathcal{N}\right]}\sqrt{E_p\left\{I_1\Pi^1\left[\eta_{2,DR}-\eta_2\right]^2\middle|\mathcal{N}\right\}}\end{aligned}$$

Now,

$$\sqrt{E_p\left[\left\{I_1\left(\frac{1}{h_1}-\frac{1}{\widehat{h}_1}\right)\right\}^2\middle|\mathcal{N}\right]} \lesssim_P \beta_{1,n}$$

again by Lemma 29. Then, to arrive at (B.50) it suffices to show that

$$E_p\left\{I_1\Pi^1\left[\eta_{2,DR}-\eta_2\right]^2\middle|\mathcal{N}^1,\mathcal{N}^2\right\} \lesssim_P \left(n^{-r_2}\right)^2.$$

Now, Condition R(k) for $k=1$ implies that

(i) the eigenvalues of $E_p\left\{I_1\phi_1\left(L_1\right)\phi_1\left(L_1\right)'\right\}$ are bounded above and bellow away from zero, and

(ii) $\left\|\phi_1\right\|_\infty \leq \sqrt{m_1}$.

Then, assumptions 1 and 2 of Lemma 21 in Appendix B.6.1 hold with $\mathcal{S}_n, \mathcal{S}_n^*, X, A, a$ and $\widehat{y}\left(O\right)$ replaced by $\mathcal{N}^1, \mathcal{N}^2 L_1, A_1, a_1^*$ and $\eta_{2,DR}\left(\overline{L}_2\right) - \eta_2\left(\overline{L}_2\right)$. Furthermore, the fact that $m_1 \asymp n^{\frac{1}{2\gamma_1+1}}$ implies that the assumption 3 of that lemma also holds. Therefore,

$$\begin{aligned} E_p\left\{ I_1\Pi^1\left[\eta_{2,DR}-\eta_2\right]^2 \middle| \mathcal{N}^1, \mathcal{N}^2 \right\} &\lesssim_P E_p\left\{ I_1\left(\eta_{2,DR}-\eta_2\right)^2 \middle| \mathcal{N}^2 \right\} \\ &\lesssim E_p\left\{ \overline{I}_2\left(\eta_{2,DR}-\eta_2\right)^2 \middle| \mathcal{N}^2 \right\} \\ &\lesssim_P \left(n^{-r_2}\right)^2, \end{aligned}$$

where the second bound follows by Lemma 33 and Condition B(k) for $k=2$ and the last bound follows from again Lemma 30. This concludes the proof. ■

Remark 2 *Notice that for the validity of the sequence of bounds in the last display of the preceding proof it was crucial that $\eta_{2,DR}$ was a data dependent function that depended only on the data in the sample $\mathcal{N}^2$ independent of $\mathcal{N}^1$, as otherwise we could have not invoked Lemma 21. Observe that in Lemma 21, the independence of the sample $\mathcal{S}_n^*$ from the sample $\mathcal{S}_n$ is crucial for the validity of the equality (B.32). This highlights the reason why we have chosen to sample split the nuisance estimation sample so as to estimate the $\eta_k's$ from independent samples. Without sample splitting we wouldn't know how to bound $E_p\left\{ I_1\Pi^1\left[\eta_{2,DR}-\eta_2\right]^2 \middle| \mathcal{N} \right\}$.*

Proof of Theorem 4. First, note that Condition R(k) for $k=1$ implies that

(i) uniformly over n, eigenvalues of $E_p\left\{I_1\phi_1\left(L_1\right)\phi_1\left(L_1\right)'\right\}$ are bounded above and bellow away from zero and

(ii) $\left\|\phi_1\right\|_\infty \lesssim \sqrt{m_1}$.

Then, we have that

A) observations (i) and (ii) and the fact that $m_1 \asymp n^{\frac{1}{2\gamma_1+1}}$ imply that the assumptions of Lemma 21 of Appendix B.6.1 hold with $\mathcal{S}_n, \mathcal{S}_n^*, X, A$ and a replaced by $\mathcal{N}^1, \mathcal{N}^2, L_1, A_1, a_1^*$ and with $\widehat{y}\left(O\right)$ replaced by any function of the observed data O and the dataset $\mathcal{N}^2$.

Also, note that

B) the Conditions B(k) for $k=1,2$, Hconvergence(k) for $k=1$ and HrateInf(k) for $k=2$ imply that the assumptions of Lemma 28 hold for $K=2$ and $k=1,2$,

C) the Conditions B(k) for $k=1,2$, Hconvergence(k) for $k=1$, HrateInf(k) for $k=2$ and HrateL2(k) for $k=1,2$ imply that the assumptions of Lemma 29 hold for $K=2$, and

D) the Conditions Hölder(k) and R(k) for $k=1,2$, Conditions B(k) and HrateInf(k) for $k=2$ and the assumption that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k=1,2$ of the theorem imply that the assumptions of Lemma 31 hold.

We want to show that

$$\delta_k^{MR} \lesssim_P \beta_{k,n} n^{-r_k} \text{ for } k = 1, 2 \tag{B.51}$$

and

$$\xi_{1,2}^{MR} \lesssim_P \beta_{1,n}\alpha_{2,n} n^{-r_2}, \tag{B.52}$$

where, recall that

$$\delta_1^{MR} \equiv E_p\left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1}\right)(\eta_{1,MR} - \eta_1)\middle| \mathcal{N}\right\},$$
$$\delta_2^{MR} \equiv E_p\left\{ \frac{I_1}{\widehat{h}_1}\left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2}\right)(\eta_{2,MR} - \eta_2)\middle| \mathcal{N}\right\}$$

and

$$\xi_{1,2}^{MR} \equiv E_p\left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1}\right)\Pi^1\left[E_p\left\{ \left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2}\right)(\eta_{2,MR} - \eta_2)\middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N}\right\}\right]\middle| \mathcal{N}\right\}.$$

To see (B.51) , note that, for $k = 1, 2$,

$$\begin{aligned}\left|\delta_k^{MR}\right| &= E_p\left|\left\{ \frac{\overline{I}_k}{\widehat{\pi}^{k-1}}\left(\frac{1}{h_k} - \frac{1}{\widehat{h}_k}\right)(\eta_{k,MR} - \eta_k)\middle| \mathcal{N}\right\}\right| \\ &\leq \sqrt{E_p\left[\left\{ \frac{\overline{I}_k}{\widehat{\pi}^{k-1}}\left(\frac{1}{h_k} - \frac{1}{\widehat{h}_k}\right)\right\}^2\middle| \mathcal{N}\right]}\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,MR} - \eta_k\right)^2\middle| \mathcal{N}\right]}\end{aligned}$$

by Cauchy-Schwarz . But, for each $k = 1, 2$,

$$\sqrt{E_p\left[\left\{ \frac{\overline{I}_k}{\widehat{\pi}^{k-1}}\left(\frac{1}{h_k} - \frac{1}{\widehat{h}_k}\right)\right\}^2\middle| \mathcal{N}\right]} \lesssim_P \beta_{k,n} \tag{B.53}$$

by Lemma 29 and, for $k = 1, 2$,

$$\sqrt{E_p\left[\overline{I}_k\left(\eta_{k,MR} - \eta_k\right)^2\middle| \mathcal{N}\right]} \lesssim_P n^{-r_k} \tag{B.54}$$

by Lemma 31. Then, from (B.53) and (B.54) , we conclude (B.51).

To see (B.52) , we apply again Cauchy-Schwarz and arrive at

$$\left|\xi_{1,2}^{MR}\right| = \left|E_p\left\{ I_1\left(\frac{1}{h_1} - \frac{1}{\widehat{h}_1}\right) I_1\Pi^1\left[E_p\left\{ I_2\left(\frac{1}{h_2} - \frac{1}{\widehat{h}_2}\right)(\eta_{2,MR} - \eta_2)\middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N}\right\}\right]\middle| \mathcal{N}\right\}\right|$$
$$\leq \sqrt{E_p\left[\left\{ I_1\left(\frac{1}{h_1} - \frac{1}{\widehat{h}_1}\right)\right\}^2\middle| \mathcal{N}\right]}\sqrt{E_p\left[I_1\Pi^1\left[E_p\left\{ I_2\left(\frac{1}{h_2} - \frac{1}{\widehat{h}_2}\right)(\eta_{2,MR} - \eta_2)\middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N}^2\right\}\right]^2\middle| \mathcal{N}\right]}.$$

Now,

$$\sqrt{E_p\left[\left\{ I_1\left(\frac{1}{h_1} - \frac{1}{\widehat{h}_1}\right)\right\}^2\middle| \mathcal{N}\right]} \lesssim_P \beta_{1,n}$$

again by Lemma 29. Then, to arrive at (B.52), it suffices to show that

$$\sqrt{E_p\left\{I_1\Pi^1\left[E_p\left\{I_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)(\eta_{2,MR}-\eta_2)\middle|A_1=a_1^*,\overline{L}_2=\cdot,\mathcal{N}^2\right\}\right]^2\middle|\mathcal{N}^1,\mathcal{N}^2\right\}}\lesssim_P \alpha_{2,n}n^{-r_2}. \tag{B.55}$$

Let

$$f\left(\cdot,\mathcal{N}^2\right)\equiv E_p\left\{I_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)(\eta_{2,MR}-\eta_2)\middle|A_1=a_1^*,\overline{L}_2=\cdot,\mathcal{N}^2\right\}.$$

Then

$$\begin{aligned}&E_p\left\{I_1\Pi^1\left[E_p\left\{I_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)(\eta_{2,MR}-\eta_2)\middle|A_1=a_1^*,\overline{L}_2=\cdot,\mathcal{N}^2\right\}\right]^2\middle|\mathcal{N}^1,\mathcal{N}^2\right\}\\&=E_p\left\{I_1\Pi^1\left[f\left(\cdot,\mathcal{N}^2\right)\right]^2\middle|\mathcal{N}^1,\mathcal{N}^2\right\}.\end{aligned} \tag{B.56}$$

Now, applying Lemma 21 of Appendix B.6.1 with $\mathcal{N}^1,\mathcal{N}^2L_1,A_1,a_1^*$ and $f\left(\overline{L}_2,\mathcal{N}^2\right)$ playing the roll of $\mathcal{S}_n,\mathcal{S}_n^*,X,A,a$ and $\widehat{y}\left(O\right)$, we arrive at

$$E_p\left\{I_1\Pi^1\left[f\left(\cdot,\mathcal{N}^2\right)\right]^2\middle|\mathcal{N}^1,\mathcal{N}^2\right\}\lesssim_P E_p\left\{I_1f\left(\overline{L}_2,\mathcal{N}^2\right)^2\middle|\mathcal{N}^2\right\}. \tag{B.57}$$

But

$$\begin{aligned}0&\leq E_p\left\{I_1f\left(\overline{L}_2,\mathcal{N}^2\right)^2\middle|\mathcal{N}^2\right\}\\&=E_p\left\{I_1E_p\left\{I_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)(\eta_{2,MR}-\eta_2)\middle|A_1,\overline{L}_2,\mathcal{N}^2\right\}^2\middle|\mathcal{N}^2\right\} &\text{(B.58)}\\&\leq E_p\left\{I_1E_p\left\{I_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)^2(\eta_{2,MR}-\eta_2)^2\middle|A_1,\overline{L}_2,\mathcal{N}^2\right\}\middle|\mathcal{N}^2\right\}\\&=E_p\left\{E_p\left\{\overline{I}_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)^2(\eta_{2,MR}-\eta_2)^2\middle|A_1,\overline{L}_2,\mathcal{N}^2\right\}\middle|\mathcal{N}^2\right\}\\&=E_p\left\{\overline{I}_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)^2(\eta_{2,MR}-\eta_2)^2\middle|\mathcal{N}^2\right\}\\&=E_p\left\{\overline{I}_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)^2\overline{I}_2(\eta_{2,MR}-\eta_2)^2\middle|\mathcal{N}^2\right\}\\&\lesssim_P \alpha_{2,n}^2E_p\left\{\overline{I}_2(\eta_{2,MR}-\eta_2)^2\middle|\mathcal{N}^2\right\} &\text{(B.59)}\end{aligned}$$

by part b) of Lemma 28 and Condition HrateInf(k) for $k=2$. Then, equations (B.56) – (B.59) imply that

$$\sqrt{E_p\left\{I_1\Pi^1\left[E_p\left\{I_2\left(\frac{1}{h_2}-\frac{1}{\widehat{h}_2}\right)(\eta_{2,MR}-\eta_2)\middle|A_1=a_1^*,\overline{L}_2=\cdot,\mathcal{N}^2\right\}\right]^2\middle|\mathcal{N}^1,\mathcal{N}^2\right\}}$$
$$\lesssim_P \alpha_{2,n}\sqrt{E_p\left\{\overline{I}_2\left(\eta_{2,MR}-\eta_2\right)^2\middle|\mathcal{N}^2\right\}}$$

Furthermore,

$$\sqrt{E_p\left\{\overline{I}_2\left(\eta_{2,MR}-\eta_2\right)^2\middle|\mathcal{N}^2\right\}}\lesssim_P n^{-r_2} \tag{B.60}$$

by Lemma 31, from where we conclude that equation (B.55) holds, as we wanted to show. ■

Proof of Theorem 5. First, note that Condition R(k), for $k=1,2$ implies that, for $k=1,2$,

(i) uniformly over n, eigenvalues of $E_p\left\{\overline{I}_k\phi_k\left(\overline{L}_k\right)\phi_k\left(\overline{L}_k\right)'\right\}$ are bounded above and bellow away from zero, and

(ii) $\|\phi_k\|_\infty \lesssim \sqrt{m_k}$.

Then, we have that

A) observations (i) and (ii) for $k=1$ and the fact that $m_1 \asymp n^{\frac{1}{2\gamma_1+1}}$ imply that the assumptions of Lemma 21 of Appendix B.6.1 hold with $\mathcal{S}_n, \mathcal{S}_n^*, X, A$ and a replaced by $\mathcal{N}^1, \underline{\mathcal{N}^2} \equiv \mathcal{N}^2 \cup \mathcal{N}^3, L_1, A_1$ and a_1^* respectively and with $\widehat{y}(O)$ replaced by any function of the observed data O and the dataset $\underline{\mathcal{N}^2}$.

B) observations (i) and (ii) for $k=2$ and the fact that $m_2 \asymp n^{\frac{1}{2\gamma_2+1}}$ imply that the assumptions of Lemma 21 of Appendix B.6.1 hold with $\mathcal{S}_n, \mathcal{S}_n^*, X, A$ and a replaced by $\mathcal{N}^2, \mathcal{N}^3\overline{L}_2, \overline{A}_2$ and $\overline{a}_2^*$ respectively and with $\widehat{y}(O)$ replaced by any function of the observed data O and the dataset $\mathcal{N}^3$.

Finally, note that:

C) Conditions B(k), Hconvergence(k) and HrateL2(k) for $k=1,2,3$ imply that the assumptions of Lemma 29 for $K=3$ hold,

D) Conditions Hölder(k) and R(k) and the fact that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k=1,2,3$ imply that the assumptions of Lemma 30 for $K=3$ hold and

E) Condition B(k) for $k=2,3$ implies that the assumptions of Lemma 33 hold for $k=2,3$.

We must show that

$$\left|\delta_k^{DR}\right| \lesssim_P \beta_{k,n}n^{-r_k} \text{ for } k=1,2,3, \tag{B.61}$$
$$\left|\xi_{1,2}^{DR}\right| \lesssim_P \beta_{1,n}n^{-r_2}, \tag{B.62}$$
$$\left|\xi_{2,3}^{DR}\right| \lesssim_P \beta_{2,n}n^{-r_3}, \tag{B.63}$$

and

$$\left|\xi_{1,3}^{DR}\right| \lesssim_P \beta_{1,n}n^{-r_3}, \tag{B.64}$$

where recall that, for $k = 1, 2, 3$

$$\delta_k^{DR} \equiv E_p \left\{ \frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{\widehat{h}_k} \right) (\eta_{k,DR} - \eta_k) \middle| \mathcal{N} \right\}$$

and

$$\begin{aligned}
\xi_{1,2}^{DR} &\equiv E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[\eta_{2,DR} - \eta_2 \right] \middle| \mathcal{N} \right\}, \\
\xi_{2,3}^{DR} &\equiv E_p \left\{ \frac{\overline{I}_2}{\widehat{h}_1} \left(\frac{1}{h_2} - \frac{1}{\widehat{h}_2} \right) \Pi^2 \left[\eta_{3,DR} - \eta_3 \right] \middle| \mathcal{N} \right\}, \text{ and} \\
\xi_{1,3}^{DR} &\equiv E_p \left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[\Pi^2 \left[\eta_{3,DR} - \eta_3 \right] \right] \middle| \mathcal{N} \right\}.
\end{aligned}$$

As in the proof of Theorem 3, (B.61) follows applying Cauchy–Schwarz and invoking Lemmas 29 and 30.

Also as in the proof of Theorem 3, (B.62) follows applying Cauchy–Schwarz and invoking Lemmas 29, Lemma 21 of Appendix B.6.1 - now with $\mathcal{S}_n, \mathcal{S}_n^*, X, A, a$ and $\widehat{y}(O)$ replaced by $\mathcal{N}^1, \underline{\mathcal{N}^2} = \mathcal{N}^2 \cup \mathcal{N}^3, L_1, A_1, a_1^*$ and $\left\{ \eta_{2,DR}\left(\overline{L}_2\right) - \eta_2\left(\overline{L}_2\right) \right\}$ respectively - Lemma 33 and Lemma 30.

To see (B.63), we apply again Cauchy–Schwarz and arrive at

$$\begin{aligned}
\left| \xi_{2,3}^{DR} \right| &\equiv \left| E_p \left\{ \frac{\overline{I}_2}{\widehat{h}_1} \left(\frac{1}{h_2} - \frac{1}{\widehat{h}_2} \right) \overline{I}_2 \Pi^2 \left[\eta_{3,DR} - \eta_3 \right] \middle| \mathcal{N} \right\} \right| \\
&\leq \sqrt{E_p \left[\left\{ \frac{\overline{I}_2}{\widehat{h}_1} \left(\frac{1}{h_2} - \frac{1}{\widehat{h}_2} \right) \right\}^2 \middle| \mathcal{N} \right]} \sqrt{E_p \left[\overline{I}_2 \left\{ \Pi^2 \left[\eta_{3,DR} - \eta_3 \right] \right\}^2 \middle| \mathcal{N}^2, \mathcal{N}^3 \right]}
\end{aligned}$$

Now,

$$\sqrt{E_p \left[\left\{ \frac{\overline{I}_2}{\widehat{h}_1} \left(\frac{1}{h_2} - \frac{1}{\widehat{h}_2} \right) \right\}^2 \middle| \mathcal{N} \right]} \lesssim_P \beta_{2,n}$$

by Lemma 29. Then, to arrive at (B.63), it suffices to show that

$$\sqrt{E_p \left[\overline{I}_2 \Pi^2 \left[\eta_{3,DR} - \eta_3 \right]^2 \middle| \mathcal{N}^2, \mathcal{N}^3 \right]} \lesssim_P n^{-r_3}. \tag{B.65}$$

Now, invoking Lemma 21 in Appendix B.6.1 with $\mathcal{N}^2, \mathcal{N}^3 \overline{L}_2, \overline{A}_2, \overline{a}_2^*$ and $\left\{ \eta_{3,DR}\left(\overline{L}_3\right) - \eta_3\left(\overline{L}_3\right) \right\}$ playing the roll of $\mathcal{S}_n, \mathcal{S}_n^*, X, A, a$ and $\widehat{y}(O)$, we conclude that

$$\begin{aligned}
& E_p \left[\overline{I}_2 \Pi^2 \left[\eta_{3,DR} - \eta_3 \right]^2 \middle| \mathcal{N}^2, \mathcal{N}^3 \right] \\
& \lesssim_P E_p \left\{ \overline{I}_2 \left(\eta_{3,DR} - \eta_3 \right)^2 \middle| \mathcal{N}^3 \right\} \\
& \lesssim E_p \left\{ \overline{I}_3 \left(\eta_{3,DR} - \eta_3 \right)^2 \middle| \mathcal{N}^3 \right\} \\
& \lesssim_P \left(n^{-r_3} \right)^2,
\end{aligned}$$

where the bound in the third row follows from Lemma 33 for $k = 3$ and the last bound follows again from Lemma 30. This concludes the proof of (B.63).

Turn to fact (B.64) . Applying Cauchy–Schwarz once more, we have that

$$\begin{aligned}\left|\xi_{1,3}^{DR}\right| &\equiv \left| E_p\left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1}\right) \Pi^1\left[\Pi^2\left[\eta_{3,DR} - \eta_3\right]\right] \middle| \mathcal{N} \right\} \right| \\ &\leq \sqrt{E_p\left[\left\{ I_1\left(\frac{1}{h_1} - \frac{1}{\widehat{h}_1}\right)\right\}^2 \middle| \mathcal{N}\right]} \sqrt{E_p\left[I_1 \left\{\Pi^1\left[\Pi^2\left[\eta_{3,DR} - \eta_3\right]\right]\right\}^2 \middle| \mathcal{N}\right]}\end{aligned}$$

But

$$\sqrt{E_p\left[\left\{ I_1\left(\frac{1}{h_1} - \frac{1}{\widehat{h}_1}\right)\right\}^2 \middle| \mathcal{N}\right]} \lesssim_P \beta_{1,n}$$

by Lemma 29, so that, to see (B.64) it suffices to show that

$$\sqrt{E_p\left[I_1 \left\{\Pi^1\left[\Pi^2\left[\eta_{3,DR} - \eta_3\right]\right]\right\}^2 \middle| \mathcal{N}\right]} \lesssim_P n^{-r_3}.$$

Invoking Lemma 21 in Appendix B.6.1 - now with $\mathcal{N}^1, \underline{\mathcal{N}^2} \equiv \mathcal{N}^2 \cup \mathcal{N}^3, L_1, A_1, a_1^*$ and $\Pi^2\left[\eta_{3,DR} - \eta_3\right]\left(\overline{L}_2\right)$ playing the roll of $\mathcal{S}_n, \mathcal{S}_n^*, X, A, a$ and $\widehat{y}\left(O\right)$ respectively - we have that

$$\begin{aligned}&E_p\left[I_1 \left\{\Pi^1\left[\Pi^2\left[\eta_{3,DR} - \eta_3\right]\right]\right\}^2 \middle| \mathcal{N}\right] \\ &= E_p\left[I_1 \left\{\Pi^1\left[\Pi^2\left[\eta_{3,DR} - \eta_3\right]\right]\right\}^2 \middle| \mathcal{N}^1, \underline{\mathcal{N}^2}\right] \\ &\lesssim_P E_p\left\{ I_1 \Pi^2\left[\eta_{3,DR} - \eta_3\right]^2 \middle| \underline{\mathcal{N}^2}\right\} \\ &\lesssim E_p\left\{ \overline{I}_2 \Pi^2\left[\eta_{3,DR} - \eta_3\right]^2 \middle| \underline{\mathcal{N}^2}\right\},\end{aligned}$$

where the last bound follows from Lemma 33. But

$$E_p\left[\overline{I}_2 \Pi^2\left[\eta_{3,DR} - \eta_3\right]^2 \middle| \underline{\mathcal{N}^2}\right] \lesssim_P \left(n^{-r_3}\right)^2$$

by (B.65) and, hence, $\sqrt{E_p\left[I_1 \left\{\Pi^1\left[\Pi^2\left[\eta_{3,DR} - \eta_3\right]\right]\right\}^2 \middle| \mathcal{N}\right]} \lesssim_P n^{-r_3}$, as we wanted to show. ■

Proof of Theorem 6. First, note that Condition R(k) for $k = 1, 2$ implies that, for $k = 1, 2$,

(i) uniformly over n, eigenvalues of $E_p\left\{ \overline{I}_k \phi_k\left(\overline{L}_k\right) \phi_k\left(\overline{L}_k\right)' \right\}$ are bounded above and bellow away from zero, and

(ii) $\left\|\phi_k\right\|_\infty \lesssim \sqrt{m_k}$.

Then, we have that

A) observations (i) and (ii) for $k = 1$ and the fact that $m_1 \asymp n^{\frac{1}{2\gamma_1 + 1}}$ imply that the assumptions of Lemma 21 of Appendix B.6.1 hold with $\mathcal{S}_n, \mathcal{S}_n^*, X, A$ and a replaced by $\mathcal{N}^1, \underline{\mathcal{N}^2} \equiv \mathcal{N}^2 \cup \mathcal{N}^3, L_1, A_1$ and a_1^* respectively and with $\widehat{y}\left(O\right)$ replaced by any function of the observed data O and the dataset $\underline{\mathcal{N}^2}$, and

B) observations (i) and (ii) for $k=2$ and the fact that $m_2 \asymp n^{\frac{1}{2\gamma_2+1}}$ imply that the assumptions of Lemma 21 of Appendix B.6.1 hold with $\mathcal{S}_n, \mathcal{S}_n^*, X, A$ and a replaced by $\mathcal{N}^2, \mathcal{N}^3\overline{L}_2, \overline{A}_2$ and $\overline{a}_2^*$ respectively and with $\widehat{y}(O)$ replaced by any function of the observed data O and the dataset $\mathcal{N}^3$.

Also, note that

C) Conditions B(k) and Hconvergence(k) for $k=1,2,3$ imply that the assumptions of Lemma 28 for $K=3$ hold for $k=1,2,3$,

D) Conditions B(k) , Hconvergence(k) and HrateL2(k) for $k=1,2,3$ imply that the assumptions of Lemma 29 for $K=3$ hold, and

E) Conditions Hölder(k) and R(k) for $k=1,2,3$, Conditions B(k) and HrateInf(k) for $k=2,3$ and the assumption that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k=1,2,3$ of the theorem imply that the assumptions of Lemma 32 hold.

We want to show that

$$\delta_k^{MR} \lesssim_P \beta_{k,n} n^{-r_k} \text{ for } k=1,2,3, \tag{B.66}$$

$$\xi_{1,2}^{MR} \lesssim_P \beta_{1,n}\alpha_{2,n} n^{-r_2}, \tag{B.67}$$

$$\xi_{2,3}^{MR} \lesssim_P \beta_{2,n}\alpha_{3,n} n^{-r_3}, \tag{B.68}$$

$$\xi_{1,3}^{MR} \lesssim_P \beta_{1,n}\alpha_{3,n} n^{-r_3} \tag{B.69}$$

and

$$\varkappa_{1,2,3}^{MR} \lesssim_P \beta_{1,n}\alpha_{2,n}\alpha_{3,n} n^{-r_3}, \tag{B.70}$$

where recall that, for $k=1,2,3$,

$$\delta_k^{MR} \equiv E_p\left\{ \frac{\overline{I}_{k-1}}{\widehat{\pi}^{k-1}} \left(\frac{I_k}{h_k} - \frac{I_k}{\widehat{h}_k} \right) (\eta_{k,MR} - \eta_k) \middle| \mathcal{N} \right\},$$

and

$$\xi_{1,2}^{MR} \equiv E_p\left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[E_p\left\{ \left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2} \right) (\eta_{2,MR} - \eta_2) \middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\},$$

$$\xi_{2,3}^{MR} \equiv E_p\left\{ \frac{I_1}{\widehat{h}_1} \left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2} \right) \Pi^2 \left[E_p\left\{ \left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3} \right) (\eta_{3,MR} - \eta_3) \middle| \overline{A}_2 = \overline{a}_2^*, \overline{L}_3 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\},$$

$$\xi_{1,3}^{MR} \equiv E_p\left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[E_p\left\{ \frac{I_2}{\widehat{h}_2} \left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3} \right) (\eta_{3,MR} - \eta_3) \middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\}$$

and

$$\varkappa_{1,2,3}^{MR} \equiv E_p\left\{ \left(\frac{I_1}{h_1} - \frac{I_1}{\widehat{h}_1} \right) \Pi^1 \left[E_p\left\{ \left(\frac{I_2}{h_2} - \frac{I_2}{\widehat{h}_2} \right) \Pi^2 \left[E_p\left\{ \left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3} \right) (\eta_{3,MR} - \eta_3) \middle| \overline{A}_2 = \overline{a}_2^*, \overline{L}_3 = \cdot, \mathcal{N} \right\} \right] \right.\right.$$
$$\left.\left. \middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N} \right\} \right] \middle| \mathcal{N} \right\}.$$

Fact (B.66) follows from the same arguments invoked in the proof of Theorem 4 to show (B.51), but applying Lemma 32 instead of 31.

Fact (B.67) follows also from the same arguments used in the proof of Theorem 4 to show (B.52), but applying Lemma 32 instead of Lemma 31 and replacing $\mathcal{N}^2$ with $\underline{\mathcal{N}}^2$.

Fact (B.68) can also be shown with the same arguments used to prove (B.52), but now applying Lemma 21 with $\mathcal{N}^2, \mathcal{N}^3\overline{L}_2, \overline{A}_2, L_3, \overline{a}_2^*$ and $g\left(\overline{L}_3, \mathcal{N}^3\right)$ playing the roll of $\mathcal{S}_n, \mathcal{S}_n^*, X, A, W, a$ and $\widehat{y}\left(O\right)$ respectively, where

$$g\left(\cdot, \mathcal{N}^3\right) \equiv E_p\left\{\left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3}\right)\left(\eta_{3,MR} - \eta_3\right)\middle| \overline{A}_2 = \overline{a}_2^*, \overline{L}_3 = \cdot, \mathcal{N}^3\right\}.$$

Likewise, to see fact (B.69), we can apply the same arguments used in the proof of (B.52), except for two points. First, in this case, we must apply Lemma 21 with $\mathcal{N}^1, \mathcal{N}^3 L_1, A_1, L_2, a_1^*$ and $h\left(\overline{L}_2, \mathcal{N}^3\right)$ playing the roll of $\mathcal{S}_n, \mathcal{S}_n^*, X, A, W, a$ and $\widehat{y}\left(O\right)$, where

$$h\left(\cdot, \mathcal{N}^3\right) \equiv E_p\left\{\frac{I_2}{\widehat{h}_2}\left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3}\right)\left(\eta_{3,MR} - \eta_3\right)\middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \mathcal{N}^3\right\}.$$

Second, we also need to invoke Lemma 28 to bound $\left\|\widehat{h}_2^{-2}\right\|_\infty$.

Finally, fact (B.70) follows applying the same arguments used to show (B.52) but now using twice the Lemma 21 and 28, instead of only once. The Lemma 21 is applied first with $\mathcal{N}^1, \underline{\mathcal{N}}^2 L_1, A_1, L_2, a_1^*$ and $u_1\left(\overline{L}_2, \underline{\mathcal{N}}^2\right)$ playing the roll of $\mathcal{S}_n, \mathcal{S}_n^*, X, A, W, a$ and $\widehat{y}\left(O\right)$ respectively, where

$$u_1\left(\cdot, \underline{\mathcal{N}}^2\right) \equiv E_p\left\{\left(\tfrac{I_2}{h_2} - \tfrac{I_2}{\widehat{h}_2}\right)\Pi^2\left[E_p\left\{\left(\tfrac{I_3}{h_3} - \tfrac{I_3}{\widehat{h}_3}\right)\left(\eta_{3,MR} - \eta_3\right)\middle| \overline{A}_2 = \overline{a}_2^*, \overline{L}_3 = \cdot, \mathcal{N}^3\right\}\right]\middle| A_1 = a_1^*, \overline{L}_2 = \cdot, \underline{\mathcal{N}}^2\right\}.$$

Then, Lemma 21 is applied a second time with $\mathcal{N}^2, \mathcal{N}^3\overline{L}_2, \overline{A}_2, L_3, \overline{a}_2^*$ and $u_2\left(\overline{L}_3, \mathcal{N}^3\right)$ playing the roll of $\mathcal{S}_n, \mathcal{S}_n^*, X, A, W, a$ and $\widehat{y}\left(O\right)$ respectively, where

$$u_2\left(\cdot, \mathcal{N}^3\right) \equiv E_p\left\{\left(\frac{I_3}{h_3} - \frac{I_3}{\widehat{h}_3}\right)\left(\eta_{3,MR} - \eta_3\right)\middle| \overline{A}_2 = \overline{a}_2^*, \overline{L}_3 = \cdot, \mathcal{N}^3\right\}.$$

■

B.7 Proof of Lemma 11

In this appendix, we prove Lemma 11 for the special cases $K = 2$ and $K = 3$. As we will see, the arguments in this proof are valid for any K except when we invoke Lemmas 31 and 32 to ensure the L^2 convergence of the $\eta'_{k,MR}s$, since these results assume $K = 2$ and $K = 3$ respectively.

Throughout this proof, as in Appendix B.6, we will use repeatedly the fact that for any fixed $p \in \mathbb{N}$ and any sequence $\{a_n\}_{n\in\mathbb{N}}$ of $p \times 1$ vectors with $a_n \equiv (a_{n,1}, \ldots, a_{n,p})$, it holds that

$$\left(\sum_{i=1}^{p} a_{n,i}\right)^2 \lesssim \sum_{i=1}^{p} a_{n,i}^2.$$

Recall that, for arbitrary $\left(h^\dagger, \eta^\dagger\right)$,

$$Q\left(h^\dagger, \eta^\dagger\right) \equiv \eta_1^\dagger\left(L_1\right) + \sum_{k=1}^{K} \frac{\overline{I}_k}{\pi^{\dagger k}} \left\{\eta_{k+1}^\dagger\left(\overline{L}_{k+1}\right) - \eta_k^\dagger\left(\overline{L}_k\right)\right\}$$

Now, using the formula

$$\widehat{a}\widehat{b} - ab = \widehat{a}\left(\widehat{b} - b\right) + b\left(\widehat{a} - a\right) \tag{B.71}$$

we obtain that,

$$\begin{aligned} Q\left(h^\dagger, \eta^\dagger\right) - Q\left(h, \eta\right) &= \eta_1^\dagger - \eta_1 \\ &+ \sum_{k=1}^{K} \left[\frac{\overline{I}_k}{\pi^{\dagger k}} \left\{\left(\eta_{k+1}^\dagger - \eta_{k+1}\right) - \left(\eta_k^\dagger - \eta_k\right)\right\} + \left(\frac{\overline{I}_k}{\pi^{\dagger k}} - \frac{\overline{I}_k}{\pi^k}\right)\left(\eta_{k+1} - \eta_k\right)\right]. \end{aligned}$$

Hence,

$$\begin{aligned} \left\{Q\left(h^\dagger, \eta^\dagger\right) - Q\left(h, \eta\right)\right\}^2 &\lesssim \left(\eta_1^\dagger - \eta_1\right)^2 \\ + \sum_{k=1}^{K} \left\{\left[\frac{\overline{I}_k}{\pi^{\dagger k}} \left\{\left(\eta_{k+1}^\dagger - \eta_{k+1}\right) - \left(\eta_k^\dagger - \eta_k\right)\right\}\right]^2 + \left[\left(\frac{\overline{I}_k}{\pi^{\dagger k}} - \frac{\overline{I}_k}{\pi^k}\right)\left(\eta_{k+1} - \eta_k\right)\right]^2\right\} \end{aligned} \tag{B.72}$$

Throughout, $\eta^\dagger$ stands for a placeholder for $\widehat{\eta}$ or $\widehat{\eta}_{MR}$. Likewise, for $k \in [K], \eta_k^\dagger$ stands for a placeholder for $\widehat{\eta}_k$ or $\widehat{\eta}_{k,MR}$ and $\eta_{K+1}^\dagger\left(L_{K+1}\right) \equiv \kappa\left(L_{K+1}\right)$. Equation (B.72) implies that

$$\begin{aligned} E_p\left[\left\{Q\left(\widehat{h}, \eta^\dagger\right) - Q\left(h, \eta\right)\right\}^2 \middle| \mathcal{N}\right] &\lesssim E_p\left[\left(\eta_1^\dagger - \eta_1\right)^2 \middle| \mathcal{N}\right] \\ + \sum_{k=1}^{K} \left[E_p\left\{\left[\frac{\overline{I}_k}{\widehat{\pi}^k} \left\{\left(\eta_{k+1}^\dagger - \eta_{k+1}\right) - \left(\eta_k^\dagger - \eta_k\right)\right\}\right]^2 \middle| \mathcal{N}\right\} + E_p\left\{\left[\left(\frac{\overline{I}_k}{\widehat{\pi}^k} - \frac{\overline{I}_k}{\pi^k}\right)\left(\eta_{k+1} - \eta_k\right)\right]^2 \middle| \mathcal{N}\right\}\right]. \end{aligned}$$

Now,

$$E_p\left[\left\{\eta_1^\dagger - \eta_1\right\}^2 \middle| \mathcal{N}\right] \lesssim E_p\left[I_1\left\{\eta_1^\dagger - \eta_1\right\}^2 \middle| \mathcal{N}\right] \tag{B.73}$$

by Lemma 33 and Condition B(k) for $k=1$. Also, for $k \in [K]$,

$$\begin{aligned} & E_p\left\{\left[\frac{\overline{I}_k}{\widehat{\pi}^k}\left\{\left(\eta_{k+1}^{\dagger}-\eta_{k+1}\right)-\left(\eta_k^{\dagger}-\eta_k\right)\right\}\right]^2\middle|\mathcal{N}\right\} \\ & \lesssim\left\|\left(\widehat{\pi}^k\right)^{-2}\right\|_{\infty}\left[E_p\left\{\overline{I}_k\left(\eta_{k+1}^{\dagger}-\eta_{k+1}\right)^2\middle|\mathcal{N}\right\}+E_p\left\{\overline{I}_k\left(\eta_k^{\dagger}-\eta_k\right)^2\middle|\mathcal{N}\right\}\right] \\ & \lesssim_P E_p\left\{\overline{I}_{k+1}\left(\eta_{k+1}^{\dagger}-\eta_{k+1}\right)^2\middle|\mathcal{N}\right\}+E_p\left\{\overline{I}_k\left(\eta_k^{\dagger}-\eta_k\right)^2\middle|\mathcal{N}\right\} \end{aligned} \tag{B.74}$$

by Lemmas 33 and 28 and by Conditions B(k) and Hconvergence(k) for $k \in [K]$. Finally, for $k \in [K]$,

$$\begin{aligned} E_p\left\{\left[\left(\frac{\overline{I}_k}{\widehat{\pi}^k}-\frac{\overline{I}_k}{\pi^k}\right)\left(\eta_{k+1}-\eta_k\right)\right]^2\middle|\mathcal{N}\right\} & \leq\left\|\left(\widehat{\pi}^k\right)^{-2}\right\|_{\infty}\left\|\left(\pi^k\right)^{-2}\right\|_{\infty}\left\|\widehat{\pi}^k-\pi^k\right\|_{\infty}^2 E_p\left\{\overline{I}_k\left(\eta_{k+1}-\eta_k\right)^2\right\} \\ & \lesssim_P\left\|\widehat{\pi}^k-\pi^k\right\|_{\infty}^2 E_p\left\{\overline{I}_k\left(\eta_{k+1}-\eta_k\right)^2\right\} \end{aligned} \tag{B.75}$$

by Conditions B(k) for $k \in [K]$ and Lemma 28. But, for $k \in [K]$,

$$\left\|\widehat{\pi}^k-\pi^k\right\|_{\infty}=o_p(1)$$

by Condition Hconvergence(k) for $k \in [K]$. Thus, equations (B.73) – (B.75) imply that, to arrive at the desired result it suffices to show that

$$E_p\left\{\overline{I}_k\left(\eta_{k+1}-\eta_k\right)^2\right\} \lesssim 1 \text{ for } k \in [K] \tag{B.76}$$

and

$$E_p\left\{\overline{I}_k\left(\eta_k^{\dagger}-\eta_k\right)^2\middle|\mathcal{N}\right\}=o_p(1) \text{ for } k \in [K]. \tag{B.77}$$

Fact (B.76) follows from Condition R(k), since it implies that

$$sup_{\overline{l}_k \in \overline{\mathcal{L}}_k} E\left\{\left\{\eta_{k+1}\left(\overline{L}_{k+1}\right)-\eta_k\left(\overline{L}_k\right)\right\}^2\middle|\overline{A}_k=\overline{a}_k^*, \overline{L}_k=\overline{l}_k\right\} \lesssim 1,$$

for $k \in [K]$.

Then, to conclude the proof, it suffices to show (B.77) for $\eta^{\dagger}=\widehat{\eta}$ and $\eta^{\dagger}=\widehat{\eta}_{MR}$.

First note that Condition R(k) and the assumption that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k \in [K]$ imply that the assumptions of Lemma 21 of Appendix B.6.1 are verified with $\mathcal{S}_n, \mathcal{S}_n^*, X, A$ and a replaced by $\mathcal{N}^k, \underline{\mathcal{N}}^{k+1}, \overline{L}_k, \overline{A}_k$ and $\overline{a}_k^*$ respectively and with $\widehat{y}(O)$ replaced by any function of the observed data O and the dataset $\underline{\mathcal{N}}^{k+1}$ for any $k \in [K]$. Also note that Condition B(k) imply that the assumptions of Lemma 33 hold for $k \in [K]$.

Now, we prove (B.77) for $\eta^{\dagger}=\widehat{\eta}$. First note that Conditions Hölder(k) and R(k) and the assumption that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k \in [K]$, imply that the assumptions of Lemma 30 hold, so that

$$E_p\left\{\overline{I}_k\left(\eta_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right\}=o_p(1) \text{ for } k \in [K]. \tag{B.78}$$

We will show that

$$E_p\left\{\overline{I}_k\left(\widehat{\eta}_{k,DR}-\eta_k\right)^2\middle|\mathcal{N}\right\}=o_p(1) \tag{B.79}$$

for $k \in [K]$ by reverse induction. First note that, when $k = K, \widehat{\eta}_{K,DR} = \eta_{K,DR}$, then (B.78) for $k = K$ implies (B.79) for $k = K$. Now, assume that (B.79) holds for $k = j+1$. We must show that it holds for $k = j$. First, note that

$$E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,DR} - \eta_j\right)^2 \middle| \mathcal{N}\right\} \lesssim E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,DR} - \eta_{j,DR}\right)^2 \middle| \mathcal{N}\right\} + E_p\left\{\overline{I}_j\left(\eta_{j,DR} - \eta_j\right)^2 \middle| \mathcal{N}\right\}.$$

But, $E_p\left\{\overline{I}_j\left(\eta_{j,DR} - \eta_j\right)^2 \middle| \mathcal{N}\right\} = o_p(1)$ by (B.78), then it suffices to show that $E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,DR} - \eta_{j,DR}\right)^2 \middle| \mathcal{N}\right\} = o_p(1)$. Now, note that $\widehat{\eta}_{j,DR} - \eta_{j,DR} = \Pi^j\left[\widehat{\eta}_{j+1,DR} - \eta_{j+1}\right]$, so

$$\begin{aligned} E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,DR} - \eta_{j,DR}\right)^2 \middle| \mathcal{N}\right\} &= E_p\left\{\overline{I}_j \Pi^j\left[\widehat{\eta}_{j+1,DR} - \eta_{j+1}\right]^2 \middle| \mathcal{N}^j, \underline{\mathcal{N}}^{j+1}\right\} \\ &\lesssim_P E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j+1,DR} - \eta_{j+1}\right)^2 \middle| \underline{\mathcal{N}}^{j+1}\right\} \\ &\lesssim E_p\left\{\overline{I}_{j+1}\left(\widehat{\eta}_{j+1,DR} - \eta_{j+1}\right)^2 \middle| \underline{\mathcal{N}}^{j+1}\right\} \end{aligned}$$

where the bound in the second row follow from Lemma 21 - with $\mathcal{S}_n, \mathcal{S}_n^*, X, A, a$ and $\widehat{y}(O)$ replaced by $\mathcal{N}^k, \underline{\mathcal{N}}^{k+1}, \overline{L}_k, \overline{A}_k, \overline{a}_k^*$ and $\{\widehat{\eta}_{j+1,DR}(L_{j+1}) - \eta_{j+1}(L_{j+1})\}$ respectively - and the bound in the third row follows by Lemma 33. But, $E_p\left\{\overline{I}_{j+1}\left(\widehat{\eta}_{j+1,DR} - \eta_{j+1}\right)^2 \middle| \underline{\mathcal{N}}^{j+1}\right\} = o_p(1)$ by inductive hypothesis and, hence, $E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,DR} - \eta_{j,DR}\right)^2 \middle| \mathcal{N}\right\} = o_p(1)$ as we wanted to show.

Now, we prove (B.77) for $\eta^\dagger = \widehat{\eta}_{MR}$. Note that Conditions Hölder(k), R(k), B(k) and Hconvergence(k) and the assumption that $m_k \asymp n^{\frac{1}{2\gamma_k+1}}$ for $k \in [K]$ imply that the assumptions of Lemma 31 hold if $K = 2$ and also imply that the assumptions of Lemma 32 hold if $K = 3$. Then, if $K = 2$ or $K = 3$, we have that

$$E_p\left\{\overline{I}_j\left(\eta_{j,MR} - \eta_j\right)^2 \middle| \mathcal{N}\right\} = o_p(1) \text{ for } j \in [K]. \tag{B.80}$$

We will show that

$$E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,MR} - \eta_j\right)^2 \middle| \mathcal{N}\right\} = o_p(1) \tag{B.81}$$

for $j \in [K]$ again by reverse induction assuming that $K = 2$ or $K = 3$. When $j = K$, fact (B.81) follows from the fact that $\widehat{\eta}_{K,MR} \equiv \eta_{K,MR} \equiv \eta_{K,DR}$ and from equation (B.80) for $j = K$. Now, assume that (B.81) holds for every $j = k+1, \ldots, K$. We want to show that it holds for $j = k$. Since $E_p\left\{\overline{I}_k\left(\eta_{k,MR} - \eta_k\right)^2 \middle| \mathcal{N}\right\} = o_p(1)$ by (B.80), it suffices to show that

$$E_p\left\{\overline{I}_k\left(\widehat{\eta}_{k,MR} - \eta_{k,MR}\right)^2 \middle| \mathcal{N}\right\} = o_p(1).$$

Note that,

$$\begin{aligned} &\widehat{\eta}_{k,MR} - \eta_{k,MR} = \\ &= \Pi^k\left[q_{k+1}\left(\cdot, \cdot; \underline{\widehat{h}}_{k+1}, \underline{\widehat{\eta}}_{k+1,MR}\right)\right] - \Pi^k\left[\eta_{k+1}\right] \\ &- \Pi^k\left[q_{k+1}\left(\cdot, \cdot; \underline{\widehat{h}}_{k+1}, \underline{\widehat{\eta}}_{k+1,MR}\right) - E_p\left\{Q_{k+1}\left(\underline{\widehat{h}}_{k+1}, \underline{\widehat{\eta}}_{k+1,MR}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_{k+1} = \cdot, \underline{\mathcal{N}}^{k+1}\right\}\right] \\ &= \Pi^k\left[E_p\left\{Q_{k+1}\left(\underline{\widehat{h}}_{k+1}, \underline{\widehat{\eta}}_{k+1,MR}\right) \middle| \overline{A}_k = \overline{a}_k^*, \overline{L}_{k+1} = \cdot, \underline{\mathcal{N}}^{k+1}\right\} - \eta_{k+1}\right]. \end{aligned}$$

Now,

$$E_p\left\{Q_{k+1}\left(\underline{\widehat{h}}_{k+1},\underline{\widehat{\eta}}_{k+1,MR}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}=$$
$$=\widehat{\eta}_{k+1,MR}+\sum_{j=k+1}^{K}E_p\left\{\frac{\overline{I}_{k+1}^j}{\widehat{\pi}_{k+1}^j}\left(\widehat{\eta}_{j+1,MR}-\widehat{\eta}_{j,MR}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}.$$

Hence,

$$\widehat{\eta}_{k,MR}-\eta_{k,MR}=\Pi^k\left[\widehat{\eta}_{k+1,MR}-\eta_{k+1}\right]+\sum_{j=k+1}^{K}\Pi^k\left[e_{j,k}\right]$$

with

$$e_{j,k}\left(\cdot\right)\equiv E_p\left\{\frac{\overline{I}_{k+1}^j}{\widehat{\pi}_{k+1}^j}\left(\widehat{\eta}_{j+1,MR}-\widehat{\eta}_{j,MR}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1}=\cdot,\underline{\mathcal{N}}^{k+1}\right\}.$$

Then,

$$E_p\left\{\overline{I}_k\left(\widehat{\eta}_{k,MR}-\eta_{k,MR}\right)^2\middle|\mathcal{N}\right\}\lesssim E_p\left\{\overline{I}_k\Pi^k\left[\widehat{\eta}_{k+1,MR}-\eta_{k+1}\right]^2\middle|\mathcal{N}\right\}+\sum_{j=k+1}^{K}E_p\left\{\overline{I}_k\Pi^k\left[e_{j,k}\right]^2\middle|\mathcal{N}\right\}.$$

Now,

$$E_p\left\{\overline{I}_k\Pi^k\left[\widehat{\eta}_{k+1,MR}-\eta_{k+1}\right]^2\middle|\mathcal{N}\right\}=E_p\left\{\overline{I}_k\Pi^k\left[\widehat{\eta}_{k+1,MR}-\eta_{k+1}\right]^2\middle|\mathcal{N}^k,\underline{\mathcal{N}}^{k+1}\right\}$$
$$\lesssim_P E_p\left\{\overline{I}_{k+1}\left(\widehat{\eta}_{k+1,MR}-\eta_{k+1}\right)^2\middle|\underline{\mathcal{N}}^{k+1}\right\}$$

by Lemmas 21 and 33. But, $E_p\left\{\overline{I}_{k+1}\left(\widehat{\eta}_{k+1,MR}-\eta_{k+1}\right)^2\middle|\mathcal{N}\right\}=o_p\left(1\right)$ by inductive hypothesis, so that $E_p\left\{\overline{I}_k\Pi^k\left[\widehat{\eta}_{k+1,MR}-\eta_{k+1}\right]^2\middle|\mathcal{N}\right\}=o_p\left(1\right)$. Also

$$E_p\left\{\overline{I}_k\Pi^k\left[e_{j,k}\right]^2\middle|\mathcal{N}\right\}=E_p\left\{\overline{I}_k\Pi^k\left[e_{j,k}\right]^2\middle|\mathcal{N}^k,\underline{\mathcal{N}}^{k+1}\right\}$$
$$\lesssim_P E_p\left\{\overline{I}_k e_{j,k}\left(\overline{L}_{k+1}\right)^2\middle|\underline{\mathcal{N}}^{k+1}\right\}$$

again by Lemma 21. Thus, we would arrived at the desired result if we show that

$$E_p\left\{\overline{I}_k e_{j,k}\left(\overline{L}_{k+1}\right)^2\middle|\underline{\mathcal{N}}^{k+1}\right\}=o_p\left(1\right)$$

for all $j=k+1,\ldots,K$.

Now,

$$e_{j,k}\left(\overline{L}_{k+1}\right)=E_p\left\{\frac{\overline{I}_{k+1}^j}{\widehat{\pi}_{k+1}^j}E_p\left(\widehat{\eta}_{j+1,MR}-\widehat{\eta}_{j,MR}|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j,\underline{\mathcal{N}}^{k+1}\right)\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}$$
$$=E_p\left[\frac{\overline{I}_{k+1}^j}{\widehat{\pi}_{k+1}^j}\left\{E_p\left(\widehat{\eta}_{j+1,MR}|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j,\underline{\mathcal{N}}^{j+1}\right)-\widehat{\eta}_{j,MR}\right\}\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right]$$
$$=E_p\left[\frac{\overline{I}_{k+1}^j}{\widehat{\pi}_{k+1}^j}\Delta_j\middle|\overline{A}_k=\overline{a}_k^*,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right],$$

with

$$\Delta_j \equiv E_p\left(\widehat{\eta}_{j+1,MR}|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j,\underline{\mathcal{N}}^{j+1}\right)-\widehat{\eta}_{j,MR}\left(\overline{L}_j\right)$$

So,

$$\begin{aligned}
E_p\left\{\overline{I}_k e_{j,k}\left(\overline{L}_{k+1}\right)^2\middle|\,\underline{\mathcal{N}}^{k+1}\right\} &= E_p\left[\overline{I}_k E_p\left(\frac{\overline{I}_{k+1}^j}{\widehat{\pi}_{k+1}^j}\Delta_j\middle|\,\overline{A}_k,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right)^2\middle|\,\underline{\mathcal{N}}^{k+1}\right]\\
&\leq E_p\left[\overline{I}_k E_p\left\{\left(\frac{\overline{I}_{k+1}^j}{\widehat{\pi}_{k+1}^j}\Delta_j\right)^2\middle|\,\overline{A}_k,\overline{L}_{k+1},\underline{\mathcal{N}}^{k+1}\right\}\middle|\,\underline{\mathcal{N}}^{k+1}\right]\\
&= E_p\left(\frac{\overline{I}_j}{\left(\widehat{\pi}_{k+1}^j\right)^2}\Delta_j^2\middle|\,\underline{\mathcal{N}}^{k+1}\right)\\
&\leq \left\|\left(\widehat{\pi}_{k+1}^j\right)^{-2}\right\|_\infty E_p\left(\overline{I}_j\Delta_j^2|\,\underline{\mathcal{N}}^{k+1}\right)\\
&\lesssim_P E_p\left(\overline{I}_j\Delta_j^2|\,\underline{\mathcal{N}}^{k+1}\right),
\end{aligned}$$

where the bound in the last row follows from 28. Now,

$$\begin{aligned}
\Delta_j &\equiv E_p\left(\widehat{\eta}_{j+1,MR}|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j,\underline{\mathcal{N}}^{j+1}\right)-\widehat{\eta}_{j,MR}\left(\overline{L}_j\right)\\
&= E_p\left(\widehat{\eta}_{j+1,MR}-\eta_{j+1}|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j,\underline{\mathcal{N}}^{j+1}\right)+\eta_j\left(\overline{L}_j\right)-\widehat{\eta}_{j,MR}\left(\overline{L}_j\right),
\end{aligned}$$

since, by definition

$$\eta_j\left(\overline{L}_j\right)\equiv E_p\left\{\eta_{j+1}\left(\overline{L}_{j+1}\right)\middle|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j\right\}.$$

Here, recall that $\widehat{\eta}_{K+1,MR}\left(\overline{L}_{K+1}\right)\equiv\eta_{K+1}\left(\overline{L}_{K+1}\right)\equiv\kappa\left(\overline{L}_{K+1}\right)$.

Hence,

$$E_p\left\{\overline{I}_k e_{j,k}^2|\,\underline{\mathcal{N}}^{k+1}\right\}\lesssim_P E_p\left\{\overline{I}_j E_p\left(\widehat{\eta}_{j+1,MR}-\eta_{j+1}|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j,\underline{\mathcal{N}}^{j+1}\right)^2\middle|\,\underline{\mathcal{N}}^{k+1}\right\}+E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,MR}-\eta_j\right)^2\middle|\,\mathcal{N}\right\}.$$

But,

$$E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j,MR}-\eta_j\right)^2\middle|\,\mathcal{N}\right\}=o_p(1)$$

for $j=k+1,\ldots,K$ by inductive hypothesis and

$$\begin{aligned}
&E_p\left\{\overline{I}_j E_p\left(\widehat{\eta}_{j+1,MR}-\eta_{j+1}|\,\overline{A}_j=\overline{a}_j^*,\overline{L}_j,\underline{\mathcal{N}}^{j+1}\right)^2\middle|\,\underline{\mathcal{N}}^{k+1}\right\}\\
&= E_p\left\{\overline{I}_j E_p\left(\widehat{\eta}_{j+1,MR}-\eta_{j+1}|\,\overline{A}_j,\overline{L}_j,\underline{\mathcal{N}}^{j+1}\right)^2\middle|\,\underline{\mathcal{N}}^{j+1}\right\}\\
&\leq E_p\left\{\overline{I}_j\left(\widehat{\eta}_{j+1,MR}-\eta_{j+1}\right)^2\middle|\,\mathcal{N}\right\}\\
&\lesssim E_p\left\{\overline{I}_{j+1}\left(\widehat{\eta}_{j+1,MR}-\eta_{j+1}\right)^2\middle|\,\mathcal{N}\right\}
\end{aligned}$$

by Lemma 33. But,

$$E_p\left\{\overline{I}_{j+1}\left(\widehat{\eta}_{j+1,MR}-\eta_{j+1}\right)^2\middle|\,\mathcal{N}\right\}=o_p(1)$$

for $j=k+1,\ldots,K$ by inductive hypothesis and by the fact that $\widehat{\eta}_{K+1,MR}\left(\overline{L}_{K+1}\right)\equiv\eta_{K+1}\left(\overline{L}_{K+1}\right)\equiv\kappa\left(\overline{L}_{K+1}\right)$. This concludes the proof.

Bibliography

[1] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[2] A. Belloni and V. Chernozhukov. High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation*, pages 121–156. Springer, 2011.

[3] A. Belloni, V. Chernozhukov, D. Chetverikov, and K. Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

[4] P. J. Bickel, Y. Ritov, A. B. Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[5] M. D. Cattaneo and M. H. Farrell. Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics*, 174(2):127–143, 2013.

[6] K. C. G. Chan. A simple multiply robust estimator for missing response problem. *Stat*, 2(1):143–149, 2013.

[7] X. Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

[8] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[9] S. R. Cole, M. A. Hernán, J. B. Margolick, M. H. Cohen, and J. M. Robins. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on cd4 cell count. *American journal of epidemiology*, 162(5):471–478, 2005.

[10] S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

[11] M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[12] M. H. Farrell, T. Liang, and S. Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.

[13] R. D. Gill, J. A. Wellner, and J. Præstgaard. Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics*, pages 97–128, 1989.

[14] S. Gruber and M. J. van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics*, 6(1), 2010.

[15] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[16] P. Han. A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference*, 148:101–110, 2014.

[17] P. Han. Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109(507):1159–1173, 2014.

[18] P. Han. Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, 103(3):683–700, 2016.

[19] P. Han and L. Wang. Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430, 2013.

[20] W. Härdle. *Applied nonparametric regression*. Number 19. Cambridge university press, 1990.

[21] W. Härdle. *Smoothing techniques: with implementation in S*. Springer Science & Business Media, 2012.

[22] A. R. Luedtke, O. Sofrygin, M. J. van der Laan, and M. Carone. Sequential double robustness in right-censored longitudinal models. *arXiv preprint arXiv:1705.02459*, 2017.

[23] Y. Luo and M. Spindler. High-dimensional l_2 boosting: Rate of convergence. *arXiv preprint arXiv:1602.08927*, 2016.

[24] J. Molina, A. Rotnitzky, M. Sued, and J. Robins. Multiple robustness in factorized likelihood models. *Biometrika*, 104(3):561–581, 2017.

[25] S. A. Murphy, M. J. van der Laan, J. M. Robins, and C. P. P. R. Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

[26] W. K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997.

[27] L. Orellana, A. Rotnitzky, and J. M. Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. *The international journal of biostatistics*, 6(2), 2010.

[28] M. L. Petersen, Y. Wang, M. J. Van Der Laan, D. Guzman, E. Riley, and D. R. Bangsberg. Pillbox organizers are associated with improved adherence to hiv antiretroviral therapy and viral suppression: a marginal structural model analysis. *Clinical Infectious Diseases*, 45(7):908–915, 2007.

[29] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.

[30] J. Robins. Marginal structural models. 1997 proceedings of the american statistical association, section on bayesian statistical science (pp. 1–10). *Retrieved from*, 1998.

[31] J. Robins, L. Li, E. Tchetgen, A. van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

[32] J. M. Robins. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure-period application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9-12):923–945, 1987.

[33] J. M. Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

[34] J. M. Robins. Association, causation, and marginal structural models. *Synthese*, 121(1):151–179, 1999.

[35] J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.

[36] J. M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

[37] J. M. Robins and M. A. Hernán. Estimation of the causal effects of time-varying exposures. In *Longitudinal data analysis*, pages 547–593. Chapman and Hall/CRC, 2008.

[38] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[39] J. M. Robins, L. Li, R. Mukherjee, E. T. Tchetgen, A. van der Vaart, et al. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

[40] J. M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pages 297–331. Springer, 1992.

[41] J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[42] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

[43] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[44] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121, 1995.

[45] J. M. Robins, P. Zhang, R. Ayyagari, R. Logan, E. T. Tchetgen, L. Li, T. Lumley, A. van der Vaart, H. H. R. Committee, et al. New statistical approaches to semiparametric regression with application to air pollution research. *Research report (Health Effects Institute)*, (175):3, 2013.

[46] A. Rotnitzky, D. Faraggi, and E. Schisterman. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, 101(475):1276–1288, 2006.

[47] D. Rubin and M. J. van der Laan. A doubly robust censoring unbiased transformation. *The international journal of biostatistics*, 3(1), 2007.

[48] M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

[49] A. Schick et al. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

[50] L. A. Stefanski and D. D. Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

[51] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[52] G. Tauchen. Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30(1-2):415–443, 1985.

[53] E. J. T. Tchetgen. A commentary on g. molenberghs's review of missing data methods. *Drug Information Journal*, 43(4):433–435, 2009.

[54] E. J. T. Tchetgen and I. Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.

[55] E. J. Tchetgen Tchetgen and I. Shpitser. Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864, 2014.

[56] A. A. Tsiatis, M. Davidian, and W. Cao. Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics*, 67(2):536–545, 2011.

[57] A. van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 29:679–686, 2014.

[58] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[59] T. J. VanderWeele, L. C. Hawkley, R. A. Thisted, and J. T. Cacioppo. A marginal structural model analysis for loneliness: implications for intervention trials and clinical practice. *Journal of consulting and clinical psychology*, 79(2):225, 2011.

[60] S. Vansteelandt, A. Rotnitzky, and J. Robins. Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94(4):841–860, 2007.

[61] S. Vansteelandt, T. J. VanderWeele, E. J. Tchetgen, and J. M. Robins. Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, 103(484):1693–1704, 2008.

[62] Z. Yu and M. van der Laan. Double robust estimation in longitudinal marginal structural models. *Journal of Statistical Planning and Inference*, 136(3):1061–1089, 2006.

[63] W. Zheng and M. J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.

[64] W. Zheng and M. J. van der Laan. Targeted maximum likelihood estimation of natural direct effects. *The international journal of biostatistics*, 8(1):1–40, 2012.