Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 1332–1357
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Robust discrimination under a hierarchy on the scatter matrices

Ana Bianco[a], Graciela Boente[a,*], Ana M. Pires[b], Isabel M. Rodrigues[b]

[a] *Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires C1428EHA, Argentina*

[b] *Departamento de Matemática and CEMAT, Instituto Superior Técnico, Technical University of Lisbon (TULisbon), Lisboa, Portugal*

Received 11 January 2007
Available online 7 September 2007

Abstract

Under normality, Flury and Schmid [Quadratic discriminant functions with constraints on the covariances matrices: some asymptotic results, J. Multivariate Anal. 40 (1992) 244–261] investigated the asymptotic properties of the quadratic discrimination procedure under hierarchical models for the scatter matrices, that is: (i) arbitrary scatter matrices, (ii) common principal components, (iii) proportional scatter matrices and (iv) identical matrices. In this paper, we study the properties of robust quadratic discrimination rules based on robust estimates of the involved parameters. Our analysis is based on the partial influence functions of the functionals related to these parameters and allows to derive the asymptotic variances of the estimated coefficients under models (i)–(iv). From them, we conclude that the asymptotic variances verify the same order relations as those obtained by Flury and Schmid [Quadratic discriminant functions with constraints on the covariances matrices: some asymptotic results, J. Multivariate Anal. 40 (1992) 244–261] for the classical estimators. We also perform a Monte Carlo study for different sample sizes and different hierarchies which shows the advantage of using robust procedures over classical ones, when anomalous data are present. It also confirms that better rates of misclassification can be achieved if a more parsimonious model among all the correct ones is used instead of the standard quadratic discrimination.

AMS 2000 subject classification: primary 62F35; secondary 62H30

Keywords: Common principal components; Outliers; Partial influence functions; Plug–in methods; Proportional scatter matrices; Quadratic discrimination; Robust estimation

* Corresponding author. Fax: +54 11 45763375.
E-mail addresses: abianco@dm.uba.ar (A. Bianco), gboente@dm.uba.ar (G. Boente), ana.pires@math.ist.utl.pt (A.M. Pires), isabel.rodrigues@math.ist.utl.pt (I.M. Rodrigues).

0047-259X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2007.08.008

1. Introduction

Assume that we are dealing with independent observations from two independent samples in $\mathbb{R}^p$ with location parameter $\boldsymbol{\mu}_i$ and dispersion/covariance matrix $\boldsymbol{\Sigma}_i$, $i = 1, 2$. It is usual in multivariate analysis to treat the dispersion/covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ as unrelated if an overall test of equality tells us that they are not identical. As mentioned in Flury [16] "In contrast to the univariate situation, inequality is not just inequality—there are indeed many ways in which covariance matrices can differ". He considered the following general relations among scatter matrices

- **Level 1.** $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.
- **Level 2.** The matrices satisfy a common principal component (CPC) model, i.e., $\boldsymbol{\Sigma}_i = \boldsymbol{\beta}\boldsymbol{\Lambda}_i\boldsymbol{\beta}^{\mathrm{T}}$, $i = 1, 2$, where $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p\right)$ is the orthogonal matrix of the common eigenvectors and $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$ are diagonal matrices containing the eigenvalues for each population.
- **Level 3.** The matrices are proportional to each other, i.e., $\boldsymbol{\Sigma}_2 = \rho_2\boldsymbol{\Sigma}_1$, with ρ_2 the proportionality constant.
- **Level 4.** $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

Without considering the location parameters, the number of parameters for each level is $p(p + 1)$, $2p + p(p - 1)/2$, $1 + p(p + 1)/2$ and $p(p + 1)/2$, respectively. The difference between the number of parameters in levels 1 and 4 is $p(p+1)/2$ which can be too large in practice, especially when dealing with high dimensional data.

As most classical estimators, which are optimal under normality assumptions, the linear and quadratic discriminant rules, i.e., the optimal classification rules under levels 4 and 1, respectively, are not robust due to the lack of robustness of the sample covariance matrix and so the misclassification rates can be affected by anomalous observations. To solve this problem robust alternatives to the sample mean and covariance matrix were plugged into the classification rule, see for instance, Campbell [8], Lachenbruch [22], Critchley and Vitiello [9], Fung [19], Fung [20], Croux and Dehon [10] and Croux and Joossens [13]. The aim when seeking for robust estimators of location and scatter is to estimate the location and the shape parameters, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, assuming that the distribution F of $\mathbf{x}$ is approximately known. To be more precise, it is often assumed that $\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}$ where the distribution G of $\mathbf{z}$ belongs to some neighborhood of a given distribution G_0. When $\mathbf{x}$ has an elliptically symmetric distribution F, i.e, when the distributions of the neighborhood are restricted to be spherically symmetric, the robust location functional equals $\boldsymbol{\mu}$ while the robust scatter functional is proportional to $\boldsymbol{\Sigma}$. Usually, these scatter functionals are calibrated so that under the central normal model they provide Fisher–consistent estimators of the covariance matrix. A discussion regarding the estimation of multivariate location and scatter can be found in Maronna, Martin and Yohai [24].

From now on, we will assume that $\left(\mathbf{x}_{ij}\right)_{1 \leqslant j \leqslant n_i, 1 \leqslant i \leqslant 2}$ are independent observations from two independent samples in $\mathbb{R}^p$, identically distributed within each sample, following a general multivariate location–dispersion distribution F_i with location parameter $\boldsymbol{\mu}_i$ and scatter matrix $\boldsymbol{\Sigma}_i$ that do not need to be equal to the population mean and covariance matrix, since we do not assume the existence of second moments as in the classical setting. Let us denote by $\mathbf{X}_i = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\right)$, $N = n_1 + n_2$ and $\tau_i = n_i/N$. As in the one–population setting, we want to include the situation in which the distribution F_i of $\mathbf{x}_{i1}$ is only approximately known. To be more precise, let $\boldsymbol{\Sigma}_i$ be symmetric positive definite matrices. As discussed above, we will thus assume that $\mathbf{x}_{i1} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{1/2}\mathbf{z}_{i1}$ where the distribution G_i of $\mathbf{z}_{i1}$ belongs to a neighborhood of the central target model G_0 that is often taken as the multivariate standard normal distribution.

As noted by Flury and Schmid [17], the reason for studying CPC discrimination and proportional discrimination is that if one of the restricted levels 2–4 holds, estimating $\mathbf{\Sigma}_i$ under suitable constraints should improve the estimation, leading to more stable estimates than those obtained under level 1. This suggests that better rates of misclassification can be achieved if the most parsimonious among all the correct models is used for discrimination. It is expected that the same lack of robustness observed for the linear and quadratic rules, will be inherited by CPC and proportional discrimination. For these reasons, in this paper, we go further and we will deal with robust discrimination involving levels 2 and 3.

This paper is organized as follows. In Section 2, we review different robust estimators leading to the robust discrimination rules under levels 1–4. In Section 3, we derive the partial influence functions of the coefficients under all the hierarchy levels, while in Section 4, we compute the asymptotic variances of the coefficients and we compare them across all correct models in a given situation. Finally, in Section 5 we present the results of a simulation study. Some proofs are given in the Appendix while the others can be found in Bianco et al. [1].

2. Robust discrimination

When the two populations have a normal distribution, optimal classification of a new observation $\mathbf{x}$ into one of the two populations is based on the quadratic function $Q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{\Delta}\mathbf{x} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} + \xi$, where $\mathbf{\Delta} = (1/2)\left(\mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1}\right)$, $\boldsymbol{\alpha} = \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2$ and $\xi = (1/2)\log(|\mathbf{\Sigma}_2|/|\mathbf{\Sigma}_1|) + (1/2)\left(\boldsymbol{\mu}_2^{\mathrm{T}}\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^{\mathrm{T}}\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right)$. Future observations are classified in the first population if $Q(\mathbf{x}) > \log(\pi_2/\pi_1)$ where π_1 and $\pi_2 = 1-\pi_1$ are the known prior probabilities that an observation belongs to group 1 or 2, respectively.

If the two populations have the same scatter matrix, the quadratic function becomes the Fisher's linear discrimination rule, which is optimal in the sense of minimizing the total probability of misclassification.

In practical situations, the parameters of the two populations are unknown and must be estimated, yielding to estimates of the quadratic, linear and constant coefficients $\mathbf{\Delta}$, $\boldsymbol{\alpha}$ and ξ, respectively. In this paper, we study some aspects of quadratic discrimination coefficients if $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are robustly estimated under the levels described above. In all four situations, the location parameters are estimated through robust equivariant location estimators $\widehat{\boldsymbol{\mu}}_i$, $i = 1, 2$.

Denote by $\mathbf{V}_i$ robust affine equivariant scatter estimators of $\mathbf{\Sigma}_i$, using only the observations of the ith sample. From these initial scatter matrices estimators, one can construct parsimonious robust estimators of $\mathbf{\Sigma}_i$, according to the assumed hierarchical model.

More precisely, under level 1, $\mathbf{\Delta}$, $\boldsymbol{\alpha}$ and ξ are estimated through

$$\begin{cases} \widehat{\mathbf{\Delta}}_{\mathrm{DIF}} = \dfrac{1}{2}\left(\mathbf{V}_2^{-1} - \mathbf{V}_1^{-1}\right), \\ \widehat{\boldsymbol{\alpha}}_{\mathrm{DIF}} = \mathbf{V}_1^{-1}\widehat{\boldsymbol{\mu}}_1 - \mathbf{V}_2^{-1}\widehat{\boldsymbol{\mu}}_2, \\ \widehat{\xi}_{\mathrm{DIF}} = \frac{1}{2}\log\left(\dfrac{|\mathbf{V}_2|}{|\mathbf{V}_1|}\right) + \dfrac{1}{2}\left(\widehat{\boldsymbol{\mu}}_2^{\mathrm{T}}\mathbf{V}_2^{-1}\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1^{\mathrm{T}}\mathbf{V}_1^{-1}\widehat{\boldsymbol{\mu}}_1\right). \end{cases} \tag{1}$$

On the other hand, a basic common structure, described as level 2 in the Introduction, assumes that the two scatter matrices have different eigenvalues but identical eigenvectors, i.e., $\mathbf{\Sigma}_i = \boldsymbol{\beta}\mathbf{\Lambda}_i\boldsymbol{\beta}^{\mathrm{T}}$, $i = 1, 2$, where $\mathbf{\Lambda}_i$ are diagonal matrices and $\boldsymbol{\beta}$ is the orthogonal matrix of the common eigenvectors. Denote $\widehat{\boldsymbol{\beta}}_{\mathrm{CPC},}$ and $\widehat{\mathbf{\Lambda}}_{\mathrm{CPC},i} = \mathrm{diag}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{CPC}}^{\mathrm{T}}\mathbf{V}_i\widehat{\boldsymbol{\beta}}_{\mathrm{CPC}}\right)$ the robust plug–in estimators of the

common directions and of the eigenvalue matrices $\mathbf{\Lambda}_i$ related to the scatter estimates $\mathbf{V}_i$, defined in Boente and Orellana [3] and studied in Boente, Pires and Rodrigues [5]. In this setting, define $\widehat{\mathbf{\Sigma}}_{\text{CPC},i} = \widehat{\boldsymbol{\beta}}_{\text{CPC}} \widehat{\mathbf{\Lambda}}_{\text{CPC},i} \widehat{\boldsymbol{\beta}}_{\text{CPC}}^{\text{T}}$, then, $\mathbf{\Delta}$, $\boldsymbol{\alpha}$ and ξ are estimated through

$$\begin{cases} \widehat{\mathbf{\Delta}}_{\text{CPC}} = \dfrac{1}{2}\left(\widehat{\mathbf{\Sigma}}_{\text{CPC},2}^{-1} - \widehat{\mathbf{\Sigma}}_{\text{CPC},1}^{-1}\right), \\ \widehat{\boldsymbol{\alpha}}_{\text{CPC}} = \widehat{\mathbf{\Sigma}}_{\text{CPC},1}^{-1}\widehat{\boldsymbol{\mu}}_1 - \widehat{\mathbf{\Sigma}}_{\text{CPC},2}^{-1}\widehat{\boldsymbol{\mu}}_2, \\ \widehat{\xi}_{\text{CPC}} = \dfrac{1}{2}\log\left(\dfrac{|\widehat{\mathbf{\Sigma}}_{\text{CPC},2}|}{|\widehat{\mathbf{\Sigma}}_{\text{CPC},1}|}\right) + \dfrac{1}{2}\left(\widehat{\boldsymbol{\mu}}_2^{\text{T}}\widehat{\mathbf{\Sigma}}_{\text{CPC},2}^{-1}\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1^{\text{T}}\widehat{\mathbf{\Sigma}}_{\text{CPC},1}^{-1}\widehat{\boldsymbol{\mu}}_1\right). \end{cases} \tag{2}$$

Under the proportional model described in level 3, the common eigenvalues, the proportionality constant and the eigenvalues of the first population can be robustly estimated as described in Boente and Orellana [4]. To be more precise, these authors extended to several populations the plug–in approach for principal component analysis, studied in Croux and Haesbroeck [12]. The plug–in procedure for the proportional model consists on solving a system of equations similar to that leading to the maximum likelihood estimators, for normally distributed observations, but with the sample covariance matrices replaced by robust scatter estimators. Denote $\widehat{\boldsymbol{\beta}}_{\text{PR}}$, $\widehat{\mathbf{\Lambda}}_{\text{PR},1}$ and $\widehat{\rho}_2$ the robust plug–in estimators of the parameters. Therefore, if we denote $\widehat{\mathbf{\Sigma}}_{\text{PR},1} = \widehat{\boldsymbol{\beta}}_{\text{PR}}\widehat{\mathbf{\Lambda}}_{\text{PR},1}\widehat{\boldsymbol{\beta}}_{\text{PR}}^{\text{T}}$ and $\widehat{\mathbf{\Sigma}}_{\text{PR},2} = \widehat{\rho}_2\widehat{\mathbf{\Sigma}}_{\text{PR},1}$, we have that $\mathbf{\Delta}$, $\boldsymbol{\alpha}$ and ξ are estimated through

$$\begin{cases} \widehat{\mathbf{\Delta}}_{\text{PR}} = \dfrac{1}{2}\left(\widehat{\mathbf{\Sigma}}_{\text{PR},2}^{-1} - \widehat{\mathbf{\Sigma}}_{\text{PR},1}^{-1}\right), \\ \widehat{\boldsymbol{\alpha}}_{\text{PR}} = \widehat{\mathbf{\Sigma}}_{\text{PR},1}^{-1}\widehat{\boldsymbol{\mu}}_1 - \widehat{\mathbf{\Sigma}}_{\text{PR},2}^{-1}\widehat{\boldsymbol{\mu}}_2, \\ \widehat{\xi}_{\text{PR}} = \dfrac{1}{2}\log\left(\dfrac{|\widehat{\mathbf{\Sigma}}_{\text{PR},2}|}{|\widehat{\mathbf{\Sigma}}_{\text{PR},1}|}\right) + \dfrac{1}{2}\left(\widehat{\boldsymbol{\mu}}_2^{\text{T}}\widehat{\mathbf{\Sigma}}_{\text{PR},2}^{-1}\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1^{\text{T}}\widehat{\mathbf{\Sigma}}_{\text{PR},1}^{-1}\widehat{\boldsymbol{\mu}}_1\right). \end{cases} \tag{3}$$

Finally, if the scatter matrices are assumed equal, the common scatter matrix can be estimated by $\widehat{\mathbf{\Sigma}}_{\text{EQ}} = \tau_1\mathbf{V}_1 + \tau_2\mathbf{V}_2$ leading to

$$\begin{cases} \widehat{\mathbf{\Delta}}_{\text{EQ}} = \mathbf{0}, \\ \widehat{\boldsymbol{\alpha}}_{\text{EQ}} = \widehat{\mathbf{\Sigma}}_{\text{EQ}}^{-1}\left(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2\right), \\ \widehat{\xi}_{\text{EQ}} = \dfrac{1}{2}\left(\widehat{\boldsymbol{\mu}}_2^{\text{T}}\widehat{\mathbf{\Sigma}}_{\text{EQ}}^{-1}\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1^{\text{T}}\widehat{\mathbf{\Sigma}}_{\text{EQ}}^{-1}\widehat{\boldsymbol{\mu}}_1\right). \end{cases} \tag{4}$$

A standard framework to derive the asymptotic behavior in robust principal component analysis is to assume that the estimators of the scatter matrix are asymptotically normally distributed and spherically invariant. For that reason, and since the samples of the two populations are independent, we will assume, throughout this paper, that for $i = 1, 2$, the estimators, $(\widehat{\boldsymbol{\mu}}_i, \mathbf{V}_i)$, of $(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$, are independent and satisfy the following assumptions:

A1. $\sqrt{n_i}\,(\mathbf{V}_i - \mathbf{\Sigma}_i) \xrightarrow{\mathscr{D}} \mathbf{Z}_i$, where, when dealing with random matrices, $\mathbf{W}_n \xrightarrow{\mathscr{D}} \mathbf{W}$ stands for $\text{vec}\,(\mathbf{W}_n) \xrightarrow{\mathscr{D}} \text{vec}\,(\mathbf{W})$, $\mathbf{Z}_i$ has a multivariate normal distribution with zero mean and covariance matrix $\mathbf{\Xi}_i$ such that $\mathbf{\Xi}_i = \sigma_1\left(\mathbf{I} + K_{pp}\right)(\mathbf{\Sigma}_i \otimes \mathbf{\Sigma}_i) + \sigma_2\,\text{vec}\,(\mathbf{\Sigma}_i)\ \text{vec}\,(\mathbf{\Sigma}_i)^{\text{T}}$, with K_{pp} the $p^2 \times p^2$ block matrix with the $(l, m)-$block equal to a $p \times p$ matrix with a 1 at entry (l, m) and 0 everywhere else.

A2. $\sqrt{n_i}\left(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\right) \xrightarrow{\mathscr{D}} \mathbf{z}_i$ where $\mathbf{z}_i \sim N(\mathbf{0}, \sigma_3\boldsymbol{\Sigma}_i)$. Moreover, we will assume that $\widehat{\boldsymbol{\mu}}_i$ is also asymptotically independent of the scatter estimator $\mathbf{V}_i$.

Remark 2.1. It is well known that, for elliptically distributed observations, M, S and τ–estimators are asymptotically normally distributed and spherically invariant. If the populations have ellipsoidal distributions that only differ on their location and scatter matrix and if the same robust location–scatter estimate is considered for each population, these estimators will satisfy **A1** and **A2** (see, [29]). Explicit forms for the constants σ_1 and σ_2 are given in Tyler [29], for $M-$estimators, and in Lopuhaä [23], for S and $\tau-$estimators.

It is worth noticing that **A1** and **A2** hold if the location and scatter estimates for both populations are related to the same functionals and if the populations have the same elliptical distribution, except for possible changes in the location and the scatter matrices. Thus, according to the discussion given in the Introduction, these assumptions hold if $\mathbf{x}_{i1} = \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{1/2}\mathbf{z}_{i1}$ where $\mathbf{z}_{i1} \sim G \in \mathscr{G}_\varepsilon$ such that $\mathscr{G}_\varepsilon = \{G = (1-\varepsilon)G_0 + \varepsilon H$, where H is an spherical distribution in $\mathbb{R}^p\}$.

3. Partial influence functions

When dealing with one population, the influence function is a measure of robustness with respect to modification of a single observation. It can be thought as the first derivative of the functional version of the estimator. Pires and Branco [27] introduced partial influence functions as an extension of this notion to the case in which the functional is related to more than one population. This generalization ensures that the usual properties of the influence function for the one population case are reached when dealing with several populations. Moreover, this definition measures resistance towards pointwise contaminations at each population.

Denote by F the product measure, $F = F_1 \times F_2$. Partial influence functions of a functional $T(F)$ are then defined as $\mathrm{PIF}_i(\mathbf{x}, T, F) = \lim_{\varepsilon\to 0}\left(T(F_{\varepsilon,\mathbf{x},i}) - T(F)\right)/\varepsilon$, $i = 1, 2$, where $F_{\varepsilon,\mathbf{x},1} = F_{1,\varepsilon,\mathbf{x}} \times F_2$, $F_{\varepsilon,\mathbf{x},2} = F_1 \times F_{2,\varepsilon,\mathbf{x}}$ and $F_{i,\varepsilon,\mathbf{x}} = (1-\varepsilon)F_i + \varepsilon\delta_{\mathbf{x}}$, $i = 1, 2$.

In this section, we will derive the partial influence functions of the functionals related to the discriminant coefficients defined in the previous section. Let $\mathbf{m}_i(G)$ and $\boldsymbol{\Upsilon}_i(G)$ be Fisher–consistent location and scatter functionals related to the estimates $\widehat{\boldsymbol{\mu}}_i$ and $\mathbf{V}_i$ considered in Section 2, such that $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$ and $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i$.

3.1. Level 1

Under level 1, the functionals related to the estimators of the coefficients $\boldsymbol{\Delta}$, $\boldsymbol{\alpha}$ and ξ defined in (1) are given by

$$\begin{cases} \mathbf{D}_{\text{DIF}}(F) = \dfrac{1}{2}\left(\boldsymbol{\Upsilon}_2^{-1}(F_2) - \boldsymbol{\Upsilon}_1^{-1}(F_1)\right), \\ \mathbf{a}_{\text{DIF}}(F) = \boldsymbol{\Upsilon}_1^{-1}(F_1)\mathbf{m}_1(F_1) - \boldsymbol{\Upsilon}_2^{-1}(F_2)\mathbf{m}_2(F_2), \\ c_{\text{DIF}}(F) = \dfrac{1}{2}\log\left(\dfrac{|\boldsymbol{\Upsilon}_2(F_2)|}{|\boldsymbol{\Upsilon}_1(F_1)|}\right) + \dfrac{1}{2}\Big[\mathbf{m}_2(F_2)^{\mathrm{T}}\boldsymbol{\Upsilon}_2^{-1}(F_2)\mathbf{m}_2(F_2) \\ \qquad\qquad -\mathbf{m}_1(F_1)^{\mathrm{T}}\boldsymbol{\Upsilon}_1^{-1}(F_1)\mathbf{m}_1(F_1)\Big]. \end{cases}$$

The following theorem gives the values of the partial influence functions of these coefficients which were derived in Croux and Joossens [13].

Theorem 3.1. *Let* $\mathbf{m}_i(G)$ *and* $\boldsymbol{\Upsilon}_i(G)$ *be Fisher–consistent location and scatter functionals such that* $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$ *and* $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i$. *Assume that the influence functions* $\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist. Then, the partial influence functions of* $\mathbf{D}_{\mathrm{DIF}}(F)$, $\mathbf{a}_{\mathrm{DIF}}(F)$ *and* $c_{\mathrm{DIF}}(F)$ *are*

$$\mathrm{PIF}_i(\mathbf{x}, \mathbf{D}_{\mathrm{DIF}}, F) = \frac{(-1)^{i+1}}{2}\boldsymbol{\Sigma}_i^{-1}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)\,\boldsymbol{\Sigma}_i^{-1}, \tag{5}$$

$$\mathrm{PIF}_i(\mathbf{x}, \mathbf{a}_{\mathrm{DIF}}, F) = (-1)^{i+1}\left(\boldsymbol{\Sigma}_i^{-1}\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i) - \boldsymbol{\Sigma}_i^{-1}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)\,\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right), \tag{6}$$

$$\begin{aligned}\mathrm{PIF}_i(\mathbf{x}, c_{\mathrm{DIF}}, F) = \frac{(-1)^i}{2}\Big\{&\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)\right) + 2\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)\\ &- \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)\,\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\Big\}.\end{aligned} \tag{7}$$

3.2. Level 2

Denote by $\boldsymbol{\beta}_{\mathrm{CPC}}(F)$ and $\lambda_{\mathrm{CPC},ij}$ the plug–in functionals related to the scatter functionals $\boldsymbol{\Upsilon}(F) = (\boldsymbol{\Upsilon}_1(F_1), \boldsymbol{\Upsilon}_2(F_2))$, i.e., the solution of

$$\begin{cases}\mathrm{diag}\left\{\boldsymbol{\beta}_{\mathrm{CPC}}(F)^{\mathrm{T}}\boldsymbol{\Upsilon}_i(F_i)\boldsymbol{\beta}_{\mathrm{CPC}}(F)\right\} = \boldsymbol{\Lambda}_{\mathrm{CPC},i}(F),\\ \boldsymbol{\beta}_{\mathrm{CPC},m}(F)^{\mathrm{T}}\left\{\sum_{i=1}^{2}\tau_i\frac{\lambda_{\mathrm{CPC},im}(F)-\lambda_{\mathrm{CPC},ij}(F)}{\lambda_{\mathrm{CPC},im}(F)\lambda_{\mathrm{CPC},ij}(F)}\boldsymbol{\Upsilon}_i(F_i)\right\}\boldsymbol{\beta}_{\mathrm{CPC},j}(F) = 0 \quad \text{for } m \neq j,\\ \boldsymbol{\beta}_{\mathrm{CPC},m}(F)^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{CPC},j}(F) = \delta_{mj},\end{cases} \tag{8}$$

where $\delta_{mj} = 0$ if $j \neq m$ and $\delta_{mj} = 1$ if $j = m$, while $\boldsymbol{\beta}_{\mathrm{CPC},j}$ denotes the jth column of the matrix $\boldsymbol{\beta}_{\mathrm{CPC}}$. The coefficient functionals obtained under the CPC model are given by

$$\begin{cases}\mathbf{D}_{\mathrm{CPC}}(F) = \frac{1}{2}\left(\mathbf{S}_{\mathrm{CPC},2}^{-1}(F) - \mathbf{S}_{\mathrm{CPC},1}^{-1}(F)\right),\\ \mathbf{a}_{\mathrm{CPC}}(F) = \mathbf{S}_{\mathrm{CPC},1}^{-1}(F)\mathbf{m}_1(F_1) - \mathbf{S}_{\mathrm{CPC},2}^{-1}(F)\mathbf{m}_2(F_2),\\ c_{\mathrm{CPC}}(F) = \frac{1}{2}\log\left(\frac{|\mathbf{S}_{\mathrm{CPC},2}(F)|}{|\mathbf{S}_{\mathrm{CPC},1}(F)|}\right) + \frac{1}{2}\Big[\mathbf{m}_2(F_2)^{\mathrm{T}}\mathbf{S}_{\mathrm{CPC},2}^{-1}(F)\mathbf{m}_2(F_2)\\ \qquad\qquad - \mathbf{m}_1(F_1)^{\mathrm{T}}\mathbf{S}_{\mathrm{CPC},1}^{-1}(F)\mathbf{m}_1(F_1)\Big],\end{cases} \tag{9}$$

where $\mathbf{S}_{\mathrm{CPC},i}(F) = \boldsymbol{\beta}_{\mathrm{CPC}}(F)\boldsymbol{\Lambda}_{\mathrm{CPC},i}(F)\boldsymbol{\beta}_{\mathrm{CPC}}(F)^{\mathrm{T}}$. The following theorem gives the values of their partial influence functions.

Theorem 3.2. *Let* $\mathbf{m}_i(G)$ *and* $\boldsymbol{\Upsilon}_i(G)$ *be Fisher–consistent location and scatter functionals such that* $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$ *and* $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i$. *Denote by* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ *and* $\lambda_{i1}, \ldots, \lambda_{ip}$ *the common eigenvectors and the eigenvalues of* $\boldsymbol{\Sigma}_i$. *Assume that the influence functions* $\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist and that* $\lambda_{11} > \ldots > \lambda_{1p}$. *Then, the partial influence functions of* $\mathbf{D}_{\mathrm{CPC}}(F)$, $\mathbf{a}_{\mathrm{CPC}}(F)$ *and* $c_{\mathrm{CPC}}(F)$ *are given by*

$$\begin{aligned}\mathrm{PIF}_i(\mathbf{x}, \mathbf{D}_{\mathrm{CPC}}, F) = -\frac{1}{2}\Big(&\boldsymbol{\Sigma}_2^{-1}\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},2}, F)\boldsymbol{\Sigma}_2^{-1}\\ &- \boldsymbol{\Sigma}_1^{-1}\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},1}, F)\boldsymbol{\Sigma}_1^{-1}\Big),\end{aligned} \tag{10}$$

$$\begin{aligned}\mathrm{PIF}_i(\mathbf{x}, \mathbf{a}_{\mathrm{CPC}}, F) = (-1)^{i+1}&\boldsymbol{\Sigma}_i^{-1}\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i) + \boldsymbol{\Sigma}_2^{-1}\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},2}, F)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\\ &-\boldsymbol{\Sigma}_1^{-1}\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},1}, F)\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1,\end{aligned} \tag{11}$$

$$\begin{aligned}\mathrm{PIF}_i(\mathbf{x}, c_{\mathrm{CPC}}, F) = (-1)^i \boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i) \\ + \frac{1}{2}\left\{\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\mathrm{PIF}_i\left(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},2}, F\right)\right) - \mathrm{tr}\left(\boldsymbol{\Sigma}_1^{-1}\mathrm{PIF}_i\left(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},1}, F\right)\right)\right\} \\ - \frac{1}{2}\left\{\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},2}, F)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},1}, F)\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right\}, \end{aligned} \tag{12}$$

where $\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},\ell}, F) = \sum_{j=1}^{p} \mathrm{PIF}_i(\mathbf{x}, \lambda_{\mathrm{CPC},\ell j}, F)\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\mathrm{T}}$. *Moreover, if* $\boldsymbol{\beta} = \mathbf{I}_p$, *then*

$$\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},\ell,jj}, F) = \delta_{\ell i}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_{i,jj}, F_i), \tag{13}$$

$$\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},\ell,js}, F) = \tau_i\left(\lambda_{\ell j} - \lambda_{\ell s}\right)\frac{\left(\lambda_{ij} - \lambda_{is}\right)}{\lambda_{ij}\lambda_{is}}\theta_{sj}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_{i,js}, F_i), \tag{14}$$

with $\mathbf{S}_{\mathrm{CPC},\ell,js}$ *the element* (j, s) *of the matrix* $\mathbf{S}_{\mathrm{CPC},\ell}$ *and* $\theta_{sj} = \left\{\sum_{\ell=1}^{2}\tau_\ell\left(\lambda_{\ell s} - \lambda_{\ell j}\right)^2/\left(\lambda_{\ell s}\lambda_{\ell j}\right)\right\}^{-1}$.

Remark 3.1. The expression given in Theorem 3.2 for $\mathrm{PIF}_i(\mathbf{x}, \mathbf{S}_{\mathrm{CPC},\ell}, F)$ allows to derive the partial influence functions of the discriminant coefficients, when using projection–pursuit estimates of the common directions and their size instead of plug–in estimators. The partial influence functions of the projection–pursuit functionals of the common eigenvectors and the eigenvalues of each population can be found in Boente et al. [5,7].

Note that if both scatter matrices are equal $\mathrm{PIF}_i(\mathbf{x}, \mathbf{D}_{\mathrm{CPC},js}, F) = 0$, for $j \neq s$, and so, as in Croux et al. [11], a second order analysis is necessary.

3.3. Level 3

Denote by $\boldsymbol{\beta}_{\mathrm{PR}}(F)$, $\lambda_{\mathrm{PR},j}$ and $\rho_{\mathrm{PR},2}$ the plug–in functionals related to the estimates of the common directions, the eigenvalues of the first population and the proportionality constant, under a proportional model, i.e., the solution of

$$\begin{cases} \frac{1}{p}\sum_{j=1}^{p}\frac{\boldsymbol{\beta}_{\mathrm{PR},j}(F)^{\mathrm{T}}\boldsymbol{\Upsilon}_2(F_2)\boldsymbol{\beta}_{\mathrm{PR},j}(F)}{\lambda_{\mathrm{PR},j}(F)} = \rho_{\mathrm{PR},2}(F), \\ \sum_{i=1}^{2}\frac{\tau_i}{\rho_{\mathrm{PR},i}(F)}\boldsymbol{\beta}_{\mathrm{PR},j}(F)^{\mathrm{T}}\boldsymbol{\Upsilon}_i(F_i)\boldsymbol{\beta}_{\mathrm{PR},j}(F) = \lambda_{\mathrm{PR},j}(F),\ 1 \leqslant j \leqslant p, \\ \boldsymbol{\beta}_{\mathrm{PR},m}(F)^{\mathrm{T}}\left[\sum_{i=1}^{2}\frac{\tau_i}{\rho_{\mathrm{PR},i}(F)}\boldsymbol{\Upsilon}_i(F_i)\right]\boldsymbol{\beta}_{\mathrm{PR},j}(F) = 0, \quad m \neq j, \\ \boldsymbol{\beta}_{\mathrm{PR},m}(F)^{\mathrm{T}}\boldsymbol{\beta}_{\mathrm{PR},j}(F) = \delta_{mj}, \end{cases} \tag{15}$$

where $\rho_{\mathrm{PR},1}(F) = 1$ and $\boldsymbol{\beta}_{\mathrm{PR},j}$ denotes the jth column of the matrix $\boldsymbol{\beta}_{\mathrm{PR}}$. The coefficient functionals obtained under a proportional model are given by

$$\begin{cases} \mathbf{D}_{\mathrm{PR}}(F) = \frac{1}{2}\left(\mathbf{S}_{\mathrm{PR},2}^{-1}(F) - \mathbf{S}_{\mathrm{PR},1}^{-1}(F)\right), \\ \mathbf{a}_{\mathrm{PR}}(F) = \mathbf{S}_{\mathrm{PR},1}^{-1}(F)\mathbf{m}_1(F_1) - \mathbf{S}_{\mathrm{PR},2}^{-1}(F)\mathbf{m}_2(F_2), \\ c_{\mathrm{PR}}(F) = \frac{1}{2}\log\left(\frac{|\mathbf{S}_{\mathrm{PR},2}(F)|}{|\mathbf{S}_{\mathrm{PR},1}(F)|}\right) \\ \qquad + \frac{1}{2}\left[\mathbf{m}_2(F_2)^{\mathrm{T}}\mathbf{S}_{\mathrm{PR},2}^{-1}(F)\mathbf{m}_2(F_2) - \mathbf{m}_1(F_1)^{\mathrm{T}}\mathbf{S}_{\mathrm{PR},1}^{-1}(F)\mathbf{m}_1(F_1)\right], \end{cases} \tag{16}$$

with $\mathbf{S}_{\text{PR},1}(F) = \boldsymbol{\beta}_{\text{PR}}(F)\boldsymbol{\Lambda}_{\text{PR},1}(F)\boldsymbol{\beta}_{\text{PR}}(F)^{\text{T}}$, $\boldsymbol{\Lambda}_{\text{PR},1}(F) = \text{diag}\left(\lambda_{\text{PR},1}, \ldots, \lambda_{\text{PR},p}\right)$ and $\mathbf{S}_{\text{PR},2}(F) = \rho_{\text{PR},2}(F)\mathbf{S}_{\text{PR},1}(F)$. Theorem 3.3 gives the values of their partial influence functions.

Theorem 3.3. *Let $\mathbf{m}_i(G)$ and $\boldsymbol{\Upsilon}_i(G)$ be Fisher–consistent location and scatter functionals such that $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$ and $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i = \rho_i\boldsymbol{\Sigma}_1$, $\rho_1 = 1$. Denote by $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ and $\lambda_1, \ldots, \lambda_p$ the common eigenvectors and the eigenvalues of $\boldsymbol{\Sigma}_1$. Assume that the influence functions* $\text{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\text{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist and that $\lambda_1 > \ldots > \lambda_p$. Then, the partial influence functions of $\mathbf{D}_{\text{PR}}(F)$, $\mathbf{a}_{\text{PR}}(F)$ and $c_{\text{PR}}(F)$ are given by*

$$\text{PIF}_i(\mathbf{x}, \mathbf{D}_{\text{PR}}, F) = -\frac{1}{2}\left(\boldsymbol{\Sigma}_2^{-1}\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},2}, F)\boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1}, F)\boldsymbol{\Sigma}_1^{-1}\right), \tag{17}$$

$$\begin{aligned}\text{PIF}_i(\mathbf{x}, \mathbf{a}_{\text{PR}}, F) = {} & (-1)^{i+1}\boldsymbol{\Sigma}_i^{-1}\text{IF}(\mathbf{x}, \mathbf{m}_i, F_i) + \boldsymbol{\Sigma}_2^{-1}\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},2}, F)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2, \\ & -\boldsymbol{\Sigma}_1^{-1}\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1}, F)\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1,\end{aligned} \tag{18}$$

$$\begin{aligned}\text{PIF}_i(\mathbf{x}, c_{\text{PR}}, F) = {} & (-1)^i\boldsymbol{\mu}_i^{\text{T}}\boldsymbol{\Sigma}_i^{-1}\text{IF}(\mathbf{x}, \mathbf{m}_i, F_i) \\ & +\frac{1}{2}\left\{tr\left(\boldsymbol{\Sigma}_2^{-1}\text{PIF}_i\left(\mathbf{x}, \mathbf{S}_{\text{PR},2}, F\right)\right) - tr\left(\boldsymbol{\Sigma}_1^{-1}\text{PIF}_i\left(\mathbf{x}, \mathbf{S}_{\text{PR},1}, F\right)\right)\right\} \\ & -\frac{1}{2}\left\{\boldsymbol{\mu}_2^{\text{T}}\boldsymbol{\Sigma}_2^{-1}\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},2}, F)\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right. \\ & \left.-\boldsymbol{\mu}_1^{\text{T}}\boldsymbol{\Sigma}_1^{-1}\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1}, F)\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right\},\end{aligned} \tag{19}$$

where $\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1}, F) = \sum_{j=1}^{p}\text{PIF}_i(\mathbf{x}, \lambda_{\text{PR},j}, F)\boldsymbol{\beta}_j\boldsymbol{\beta}_j^{\text{T}}$ *and* $\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},2}, F) = \text{PIF}_i(\mathbf{x}, \rho_{\text{PR},2}, F)\boldsymbol{\Sigma}_1 + \rho_2\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1}, F)$. *Moreover, if $\boldsymbol{\beta} = \mathbf{I}_p$, then*

$$\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1,jj}, F) = \frac{\tau_i}{\rho_i}\text{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_{i,jj}, F_i) - \frac{\tau_i}{\rho_i}\lambda_j A_i + \lambda_j A_1\delta_{i1}, \tag{20}$$

$$\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1,js}, F) = \frac{\tau_i}{\rho_i}\text{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_{i,js}, F_i), \tag{21}$$

$$\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},2}, F) = \left(A_2\delta_{i2} - \rho_2 A_1\delta_{i1}\right)\boldsymbol{\Sigma}_1 + \rho_2\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},1}, F), \tag{22}$$

with $A_i = (1/p)\sum_{j=1}^{p}\text{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_{i,jj}, F_i)/\lambda_j$.

As in level 2, if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, $\text{PIF}_i(\mathbf{x}, \mathbf{D}_{\text{PR},js}, F) = 0$, for $j \neq s$, and so a second order analysis is again necessary. Note also that, if the proportional model holds and $\boldsymbol{\Sigma}_1 = \text{diag}(\lambda_1, \ldots, \lambda_p)$, then $\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{CPC},\ell,js}, F) = \text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{PR},\ell,js}, F)$, for $\ell = 1, 2$.

3.4. Level 4

The coefficient functionals obtained under equality of the scatter matrices are given by

$$\begin{cases}\mathbf{a}_{\text{EQ}}(F) = \mathbf{S}_{\text{EQ}}^{-1}(F)\left[\mathbf{m}_1(F_1) - \mathbf{m}_2(F_2)\right], \\ c_{\text{EQ}}(F) = \dfrac{1}{2}\left[\mathbf{m}_2(F_2)^{\text{T}}\mathbf{S}_{\text{EQ}}^{-1}(F)\mathbf{m}_2(F_2) - \mathbf{m}_1(F_1)^{\text{T}}\mathbf{S}_{\text{EQ}}^{-1}(F)\mathbf{m}_1(F_1)\right],\end{cases} \tag{23}$$

with $\mathbf{S}_{\text{EQ}} = \tau_1\boldsymbol{\Upsilon}_1(F_1) + \tau_2\boldsymbol{\Upsilon}_2(F_2)$. The following result states the partial influence functions of the linear coefficient when using the robustified linear discrimination function and its proof can be found in Pires and Branco [27].

Theorem 3.4. *Let* $\mathbf{m}_i(G)$ *and* $\boldsymbol{\Upsilon}_i(G)$ *be Fisher–consistent location and scatter functionals such that* $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$ *and* $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$. *Assume that the influence functions* $\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist. Then, the partial influence function of* $\mathbf{a}_{\mathrm{EQ}}(F)$ *and* $c_{\mathrm{EQ}}(F)$ *are given by*

$$\mathrm{PIF}_i(\mathbf{x}, \mathbf{a}_{\mathrm{EQ}}, F) = (-1)^{i+1}\boldsymbol{\Sigma}^{-1}\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i) + \tau_i\boldsymbol{\Sigma}^{-1}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right), \quad (24)$$

$$\mathrm{PIF}_i(\mathbf{x}, c_{\mathrm{EQ}}, F) = (-1)^i \boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i) - \frac{1}{2}\tau_i \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1\right). \quad (25)$$

Remark 3.2. Croux and Joossens [13] studied how observations in the training sample affect the misclassification probability of the quadratic discriminant rule assuming level 1. Proposition 2 therein gives the partial influence functions of the total misclassification probability that is used to construct a diagnostic tool for detecting influential observations. Using it together with Theorems 3.2–3.4, one can derive an expression for the partial influence functions of the total misclassification probability under the restricted models described in levels 2–4, respectively.

Remark 3.3. Our approach based on partial influence functions assumes that prior probabilities are known. If π_i are unknown, they can be estimated by the empirical frequency of observations in the training data belonging to group i, for $i = 1, 2$ which makes it possible to attain the Bayes error rate asymptotically. In this case, influence functions can be derived as it was done by Croux, Filzmoser and Joossens [11] for the linear discriminant rule.

Figs. 1 and 2 give the plots of the partial influence function PIF_1 of the quadratic coefficients functionals $\mathbf{D}_{11}(F)$ and $\mathbf{D}_{12}(F)$. The behavior of $\mathbf{D}_{22}(F)$ is similar to that of $\mathbf{D}_{11}(F)$ except for a rotation. On the other hand, Fig. 3 shows the partial influence function PIF_1 of the norm of the linear coefficient functionals, $\mathbf{a}(F)$. In all figures, we have $p = 2$, $F = F_1 \times F_2$ with $F_1 = N_2(\mathbf{0}, \mathrm{diag}(2, 1))$ and $F_2 = N_2\left(\boldsymbol{\mu}_2, 4\,\mathrm{diag}(2, 1)\right)$ with $\boldsymbol{\mu}_2 = (4, 0)^{\mathrm{T}}$. The partial influence functions of $\xi(F)$ behave as the precedent ones and so we omit their graphs here. We have considered as scatter matrices estimators the sample covariance matrix, the S-estimator using as ρ function biweight Tukey's function calibrated to attain 25% breakdown point and the Donoho [14]–Stahel [28] estimator with weight function the Huber's function with constant $\sqrt{\chi^2_2(0.95)} = 2.4477$. For the last estimator, the univariate location and scale functionals are the median and the MAD (median of the absolute deviations with respect to the median). Expressions for the influence function of the Donoho–Stahel and the S-scatter functionals can be found in Gervini [21] and in Lopuhäa [30], respectively. Similar plots to those given by Croux and Joossens [13] assuming level 1, can be constructed for the total misclassification probability under levels 2–4, by using Proposition 2 therein and our results (see Remark 3.2).

In all cases, the shape of the partial influence functions of the robust estimates is comparable to that of their classical relatives at the center of the distribution. Besides, the influence at points further away is downweighted for the robust estimates, while it is much larger for the classical ones. However, it should be noticed that the robust functionals related to the Donoho–Stahel have a discontinuity at 0, due to the discontinuity of the influence function of the Donoho–Stahel scatter functional. On the other hand, the partial influence function of each robust functional follows the same behavior as the score function used to define them. To be more precise, in all cases, for the robust functionals, the partial influence function of $\mathbf{D}_{12}(F)$ is largest along the bisectors while that of $\mathbf{D}_{11}(F)$ attains large values only for smaller values of x_2 combined with moderate values

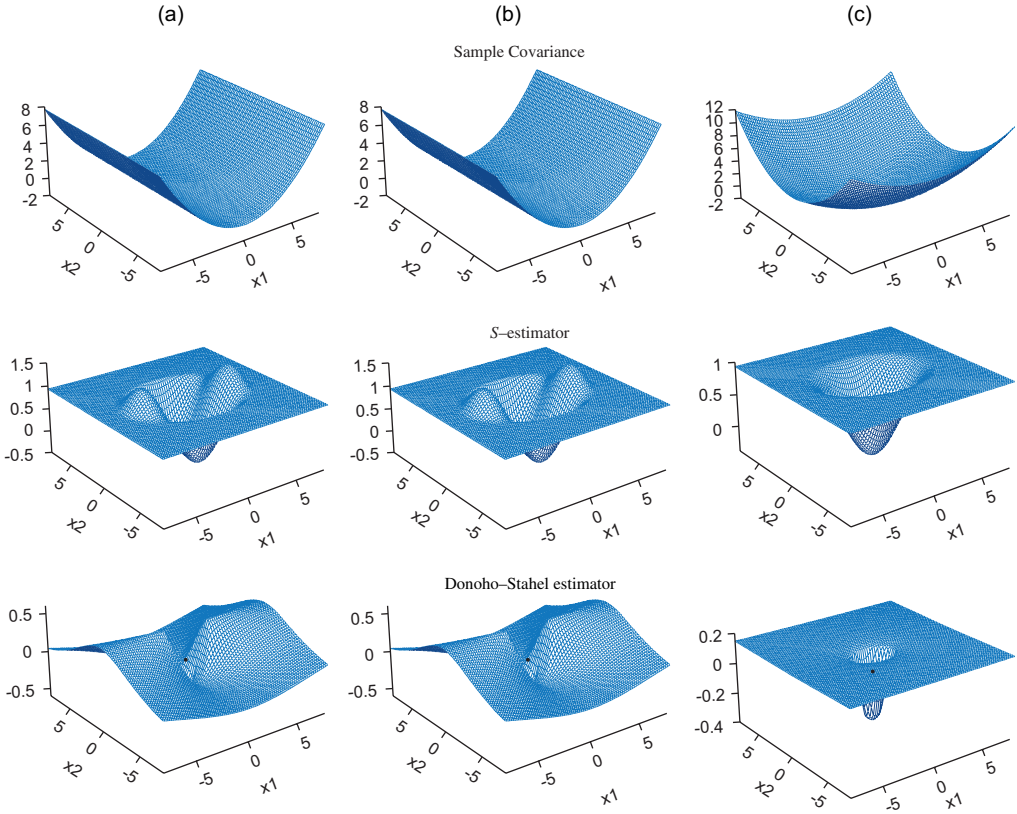

Fig. 1. (a) $\mathrm{PIF}_1(\mathbf{x}, D_{\mathrm{DIF},11}, F)$, (b) $\mathrm{PIF}_1(\mathbf{x}, D_{\mathrm{CPC},11}, F)$, and (c) $\mathrm{PIF}_1(\mathbf{x}, D_{\mathrm{PR},11}, F)$ at $F = F_1 \times F_2$ with $F_1 = N(\mathbf{0}, \mathrm{diag}(2, 1))$ and $F_2 = N\left(\boldsymbol{\mu}_2, 4\mathrm{diag}(2, 1)\right)$.

of x_1. Finally, the norm of $\mathrm{PIF}_1(\mathbf{x}, \mathbf{a}, F)$ has different shapes according to the model used for discrimination. When level 1 holds, it has a hat shape with the wings parallel to the axis x_1. Under level 2, the partial influence functions of each robust functionals show three modes while, under level 3, only two parallel bumps are present.

4. Asymptotic variances

Asymptotic variances can be derived heuristically, using partial influence functions. Let F_N denote the empirical distribution of the k independent samples $\mathbf{x}_{ij}$, $1 \leqslant j \leqslant n_i$, $1 \leqslant i \leqslant k$ and $T_N = T(F_N)$. In Pires and Branco [27], it is shown that if $N^{1/2}\{T_N - T(F)\} = \sum_{i=1}^{k} (\tau_i\, n_i)^{-1/2} \sum_{j=1}^{n_i} \mathrm{PIF}_i\left(\mathbf{x}_{ij}, T, F\right) + o_p(1)$, then the asymptotic variance of the estimates can be evaluated as

$$\mathrm{ASVAR}\,(T_N) = \mathrm{ASVAR}\,(T_N, F) = \sum_{i=1}^{k} \tau_i{}^{-1} E_{F_i}\left\{\mathrm{PIF}_i\,(\mathbf{x}_{i1}, T, F)\,\mathrm{PIF}_i\,(\mathbf{x}_{i1}, T, F)^{\mathrm{T}}\right\}. \quad (26)$$

Theorems 4.1–4.3 give the asymptotic variance of the quadratic, linear and constant coefficient estimators when the quadratic discriminant rule is used, under levels 1–3, respectively.

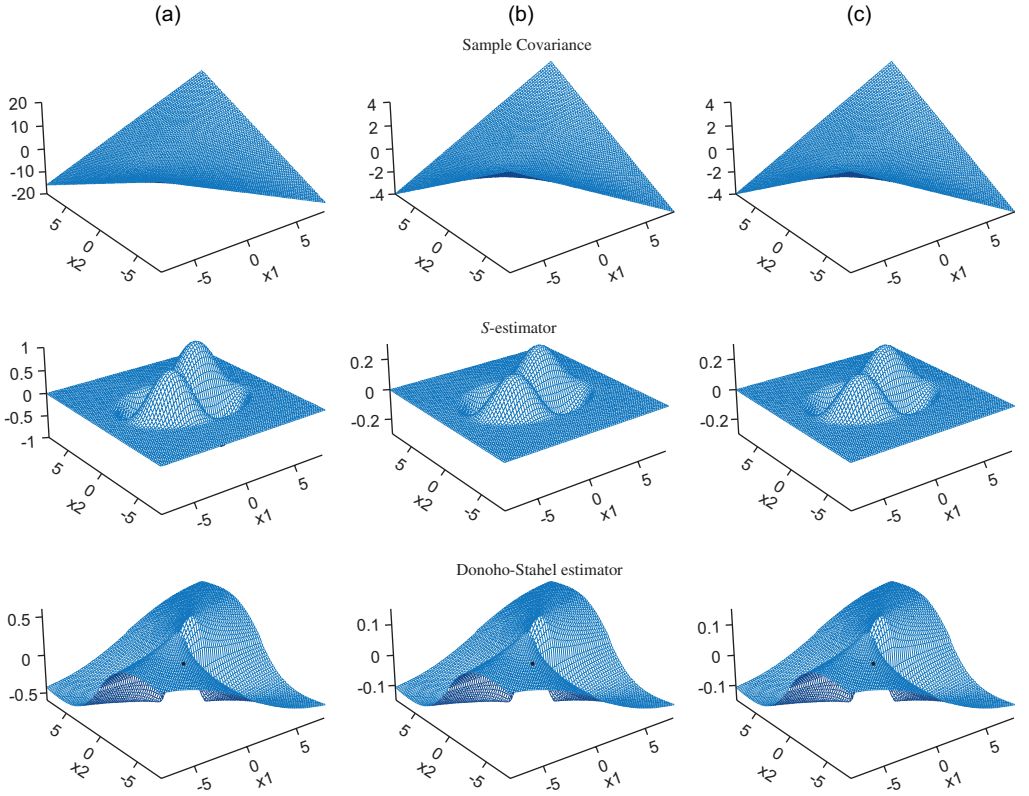

Fig. 2. (a) $\mathrm{PIF}_1(\mathbf{x}, D_{\mathrm{DIF},12}, F)$, (b) $\mathrm{PIF}_1(\mathbf{x}, D_{\mathrm{CPC},12}, F)$, and (c) $\mathrm{PIF}_1(\mathbf{x}, D_{\mathrm{PR},12}, F)$ at $F = F_1 \times F_2$ with $F_1 = N_2(\mathbf{0}, \mathrm{diag}(2,1))$ and $F_2 = N_2(\boldsymbol{\mu}_2, 4\,\mathrm{diag}(2,1))$.

Theorem 4.1. *Let* $\mathbf{m}_i(G)$ *and* $\boldsymbol{\Upsilon}_i(G)$ *be Fisher-consistent location and scatter functionals such that* $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$, $\mathbf{m}_i(F_{n_i}) = \widehat{\boldsymbol{\mu}}_i$, $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i$ *and* $\boldsymbol{\Upsilon}_i(F_{n_i}) = \mathbf{V}_i$, *with* F_{n_i} *the empirical distribution function of the* i*th population. Assume that the influence functions* $\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist and that* ***A1*** *and* ***A2*** *hold. Then, when* $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$, *i.e., when the* CPC *model holds with* $\boldsymbol{\beta} = \mathbf{I}_p$, *the asymptotic variances of the estimators* $\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF}}$ $\widehat{\boldsymbol{\alpha}}_{\mathrm{DIF}}$ *and* $\widehat{\xi}_{\mathrm{DIF}}$ *defined in* (1) *are given by*

$$\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF},js}) = \frac{1}{4}\left(\sigma_1 + [\sigma_1 + \sigma_2]\delta_{js}\right)\sum_{i=1}^{2}\frac{1}{\tau_i}\frac{1}{\lambda_{ij}\lambda_{is}}, \tag{27}$$

$$\mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DIF},j}) = \sum_{i=1}^{2}\frac{1}{\tau_i\lambda_{ij}}\left[\sigma_3 + \frac{\boldsymbol{\mu}_{ij}^2}{\lambda_{ij}}(\sigma_1+\sigma_2) + \sigma_1\sum_{s=1}^{p}\frac{\boldsymbol{\mu}_{is}^2}{\lambda_{is}}\right], \tag{28}$$

$$\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{DIF}}) = \sum_{i=1}^{2}\frac{1}{\tau_i}\left(v_{i1} + \frac{1}{4}v_{i2} + v_{i3}\right), \tag{29}$$

with $v_{i1} = \sigma_3\sum_{s=1}^{p}\boldsymbol{\mu}_{is}^2/\lambda_{is}$, $v_{i2} = \sigma_2\left[\sum_{s=1}^{p}\left(1-\boldsymbol{\mu}_{is}^2/\lambda_{is}\right)\right]^2 + 2\sigma_1\sum_{s=1}^{p}\left(1-\boldsymbol{\mu}_{is}^2/\lambda_{is}\right)^2$ *and* $v_{i3} = \sigma_1\sum_{j<s}\boldsymbol{\mu}_{is}^2\boldsymbol{\mu}_{ij}^2/\left(\lambda_{is}\lambda_{ij}\right)$.

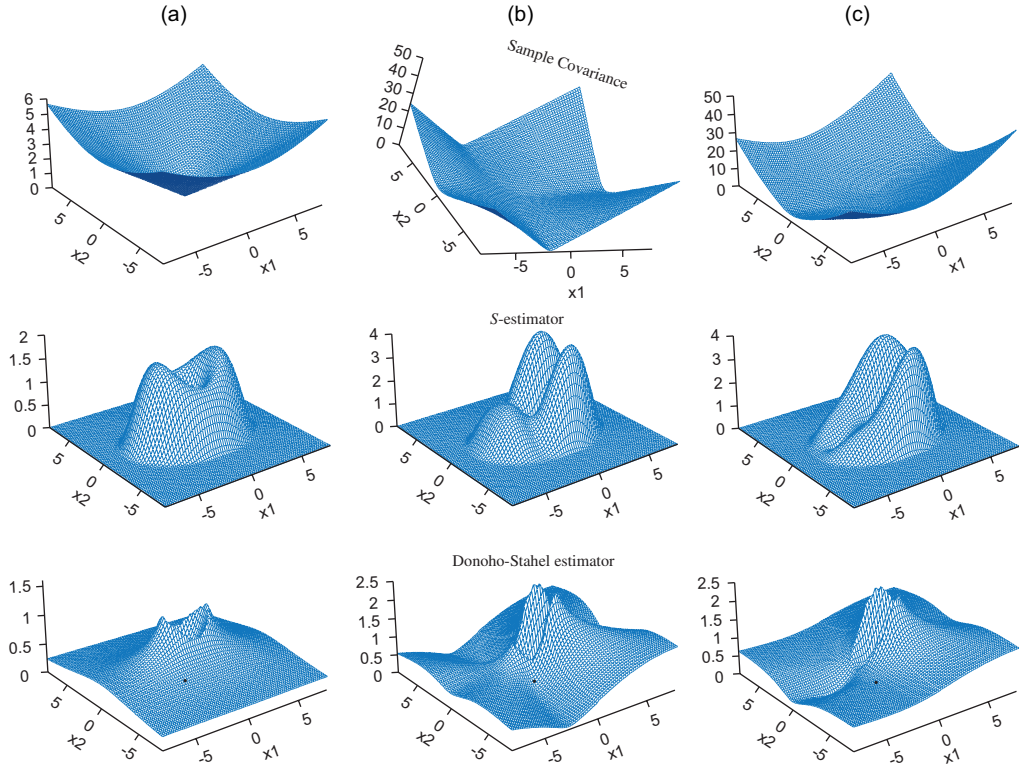

Fig. 3. (a) $\|\mathrm{PIF}_1(\mathbf{x}, \mathbf{a}_{\mathrm{DIF}}, F)\|$, (b) $\|\mathrm{PIF}_1(\mathbf{x}, \mathbf{a}_{\mathrm{CPC}}, F)\|$, and (c) $\|\mathrm{PIF}_1(\mathbf{x}, \mathbf{a}_{\mathrm{PR}}, F)\|$ at $F = F_1 \times F_2$ with $F_1 = N_2(\mathbf{0}, \mathrm{diag}(2, 1))$ and $F_2 = N_2\left(\boldsymbol{\mu}_2, 4\,\mathrm{diag}(2, 1)\right)$.

Theorem 4.2. *Let* $\mathbf{m}_i(G)$ *and* $\boldsymbol{\Upsilon}_i(G)$ *be Fisher-consistent location and scatter functionals such that* $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$, $\mathbf{m}_i(F_{n_i}) = \widehat{\boldsymbol{\mu}}_i$, $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i$ *and* $\boldsymbol{\Upsilon}_i(F_{n_i}) = \mathbf{V}_i$, *with* F_{n_i} *the empirical distribution function of the* i*th population. Moreover, assume that* $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i = \mathrm{diag}\left(\lambda_{i1}, \ldots, \lambda_{ip}\right)$, *i.e., the common principal components model holds with* $\boldsymbol{\beta} = \mathbf{I}_p$. *Assume that the influence functions* $\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist, that* $\lambda_{11} > \cdots > \lambda_{1p}$ *and that* ***A1*** *and* ***A2*** *hold. Then, the asymptotic variances of the estimators* $\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC}}$, $\widehat{\boldsymbol{\alpha}}_{\mathrm{CPC}}$ *and* $\widehat{\xi}_{\mathrm{CPC}}$ *defined in* (2) *are*

$$\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},js}) = \left(1-\delta_{js}\right)\frac{\sigma_1}{4}\theta_{sj}\left[\sum_{i=1}^{2}(-1)^i\frac{\lambda_{ij}-\lambda_{is}}{\lambda_{ij}\lambda_{is}}\right]^2 + \delta_{js}\frac{2\sigma_1+\sigma_2}{4}\sum_{i=1}^{2}\frac{1}{\tau_i\lambda_{ij}^2}, \tag{30}$$

$$\begin{aligned}\mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{CPC},j}) = {} & \sum_{i=1}^{2}\frac{1}{\tau_i\lambda_{ij}}\left[\sigma_3 + \frac{\boldsymbol{\mu}_{ij}^2}{\lambda_{ij}}(2\sigma_1+\sigma_2)\right] \\ & + \sigma_1\sum_{s=1}^{p}\theta_{sj}\left[\sum_{i=1}^{2}(-1)^{i+1}\frac{\boldsymbol{\mu}_{is}\left(\lambda_{ij}-\lambda_{is}\right)}{\lambda_{ij}\lambda_{is}}\right]^2,\end{aligned} \tag{31}$$

$$\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{CPC}}) = \sum_{i=1}^{2} \frac{1}{\tau_i} \left(v_{i1} + \frac{1}{4} v_{i2} + \tau_i^2 v_{i4} \right), \tag{32}$$

where $\theta_{sj} = \left\{ \sum_{\ell=1}^{2} \tau_\ell \left(\lambda_{\ell s} - \lambda_{\ell j}\right)^2 / \left(\lambda_{\ell s}\lambda_{\ell j}\right) \right\}^{-1}$, v_{i1} *and* v_{i2} *are defined in Theorem* 4.1 *and* $v_{i4} = \sigma_1 \sum_{j<s} \theta_{sj}^2 \left[\sum_{k=1}^{2} (-1)^k \boldsymbol{\mu}_{ks}\boldsymbol{\mu}_{kj} \left(\lambda_{kj} - \lambda_{ks}\right) / \left(\lambda_{kj}\lambda_{ks}\right) \right]^2 \left(\lambda_{ij} - \lambda_{is}\right)^2 / \left(\lambda_{ij}\lambda_{is}\right)$.

Theorem 4.3. *Let* $\mathbf{m}_i(G)$ *and* $\boldsymbol{\Upsilon}_i(G)$ *be Fisher-consistent location and scatter functionals such that* $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$, $\mathbf{m}_i(F_{n_i}) = \widehat{\boldsymbol{\mu}}_i$, $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i = \rho_i \boldsymbol{\Sigma}_1$, $\rho_1 = 1$ *and* $\boldsymbol{\Upsilon}_i(F_{n_i}) = \mathbf{V}_i$, *with* F_{n_i} *the empirical distribution function of the ith population. Moreover, assume that* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Lambda}_1 = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_p\right)$, *i.e., the proportional model holds with* $\boldsymbol{\beta} = \mathbf{I}_p$. *Assume that the influence functions* $\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist,* $\lambda_1 > \cdots > \lambda_p$ *and that* ***A1*** *and* ***A2*** *hold. Then, the asymptotic variances of the estimators* $\widehat{\boldsymbol{\Delta}}_{\mathrm{PR}}$, $\widehat{\boldsymbol{\alpha}}_{\mathrm{PR}}$ *and* $\widehat{\xi}_{\mathrm{PR}}$ *defined in* (3) *are given by*

$$\begin{aligned} &\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{PR},js}) \\ &= \frac{1}{4\lambda_s\lambda_j} \left\{ \left(1 - \delta_{js}\right) \sigma_1\gamma_2 + \delta_{js} \frac{1}{p} \left[2(p-1)\sigma_1\gamma_2 + (2\sigma_1 + p\sigma_2) \sum_{i=1}^{2} \frac{1}{\rho_i^2 \tau_i} \right] \right\}, \end{aligned} \tag{33}$$

$$\begin{aligned} &\mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{PR},j}) \\ &= \sum_{i=1}^{2} \frac{1}{\tau_i \rho_i \lambda_j} \left[\sigma_3 + \frac{\boldsymbol{\mu}_{ij}^2 (2\sigma_1 + p\sigma_2)}{p\rho_i\lambda_j} \right] + \sigma_1 \left[\frac{p-2}{p\lambda_j^2} v_j + \sum_{s=1}^{p} \frac{1}{\lambda_j\lambda_s} v_s \right], \end{aligned} \tag{34}$$

$$\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{PR}}) = \sum_{i=1}^{2} \frac{1}{\tau_i} \left(v_{i1} + \frac{1}{4} v_{i5} + \tau_i^2 v_{i4} \right), \tag{35}$$

where $\gamma_2 = \left(\rho_2^{-1} - 1\right)^2$ *and* $v_s = \left(\boldsymbol{\mu}_{2s}\rho_2^{-1} - \boldsymbol{\mu}_{1s}\right)^2$, v_{i1} *is defined in Theorem* 4.1, v_{i4} *defined in Theorem* 4.2 *equals* $\sigma_1 \sum_{s<j} \lambda_j^{-1}\lambda_s^{-1} \left(\sum_{i=1}^{2} (-1)^i \boldsymbol{\mu}_{is}\boldsymbol{\mu}_{ij}/\rho_i \right)^2$ *and*

$$\begin{aligned} v_{i5} = 2\sigma_1 \sum_{j=1}^{p} &\left[(-1)^i \tau_i \left(\frac{\boldsymbol{\mu}_{1j}^2}{\lambda_j} - \frac{\boldsymbol{\mu}_{2j}^2}{\rho_2 \lambda_j} \right) + 1 - \frac{\tau_1}{\rho_2\, p} \sum_{s=1}^{p} \frac{\boldsymbol{\mu}_{2s}^2}{\lambda_s} - \frac{\tau_2}{p} \sum_{s=1}^{p} \frac{\boldsymbol{\mu}_{1s}^2}{\lambda_s} \right]^2 \\ &+ \sigma_2 \left[\sum_{j=1}^{p} \left(1 - \frac{\boldsymbol{\mu}_{ij}^2}{\rho_i\, \lambda_j} \right) \right]^2. \end{aligned}$$

Note that if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{PR},js}) = 0$, for $j \neq s$, and so a higher order expansion is needed.

Theorem 4.4 states the asymptotic variance of the linear and constant coefficient estimators when using the robustified discrimination function.

Theorem 4.4. *Let* $\mathbf{m}_i(G)$ *and* $\boldsymbol{\Upsilon}_i(G)$ *be Fisher-consistent location and scatter functionals such that* $\mathbf{m}_i(F_i) = \boldsymbol{\mu}_i$, $\mathbf{m}_i(F_{n_i}) = \widehat{\boldsymbol{\mu}}_i$, $\boldsymbol{\Upsilon}_i(F_i) = \boldsymbol{\Sigma}_i$ *and* $\boldsymbol{\Upsilon}_i(F_{n_i}) = \mathbf{V}_i$, *with* F_{n_i} *the empirical distribution function of the ith population. Moreover, assume that* $\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Lambda}_1 = \mathrm{diag}\left(\lambda_1, \ldots, \lambda_p\right)$,

i.e., level 4 *holds with* $\boldsymbol{\beta} = \mathbf{I}_p$. *Assume that the influence functions* $\mathrm{IF}(\mathbf{x}, \mathbf{m}_i, F_i)$ *and* $\mathrm{IF}(\mathbf{x}, \boldsymbol{\Upsilon}_i, F_i)$ *exist and that* ***A*1** *and* ***A*2** *hold. Then, the asymptotic variances of the estimators* $\widehat{\boldsymbol{\alpha}}_{\mathrm{EQ}}$ *and* $\widehat{\xi}_{\mathrm{EQ}}$ *defined in* (4) *are given by*

$$\mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{EQ},j}) = \sigma_3 \sum_{i=1}^{2} \frac{1}{\tau_i \lambda_j} + \sigma_1 \sum_{s=1}^{p} \frac{1}{\lambda_j \lambda_s} v_s + (\sigma_1 + \sigma_2) \frac{1}{\lambda_j^2} v_j, \tag{36}$$

$$\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{EQ}}) = \sum_{i=1}^{2} \frac{1}{\tau_i} \left(v_{i1} + \frac{1}{4} v_{i6} + \tau_i^2 v_{i4} \right), \tag{37}$$

where $v_s = \left[\boldsymbol{\mu}_{2s} - \boldsymbol{\mu}_{1s}\right]^2$, v_{i1} *is defined in Theorem* 4.1, v_{i4} *are defined in Theorem* 4.2 *and* $v_{i6} = \tau_i^2 \left\{ 2\sigma_1 \sum_{j=1}^{p} \left(\boldsymbol{\mu}_{1j}^2/\lambda_j - \boldsymbol{\mu}_{2j}^2/\lambda_j \right)^2 + \sigma_2 \left[\sum_{j=1}^{p} \left(\boldsymbol{\mu}_{1j}^2 - \boldsymbol{\mu}_{2j}^2 \right)/\lambda_j \right]^2 \right\}$.

4.1. Variance comparisons across the different levels

In this section we compare the asymptotic variances of the estimated coefficients under the different hierarchies considered. Without loss of generality, we will assume that $\boldsymbol{\mu}_1 = \mathbf{0}$.

When the CPC model holds, Theorems 4.1 and 4.2 entail that

- $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF},jj}) = \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},jj})$: As with the classical rule, the gain achieved by using the CPC instead of ordinary quadratic discrimination may not be large, at least in the two sample case.
- $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF},js}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},js})$, $j \neq s$: Moreover, as noted by Flury and Schmid [17] for the classical estimators, equality holds for the robust quadratic coefficients if $\lambda_{1s} - \lambda_{1j} = \lambda_{2s} - \lambda_{2j}$. On the other hand, as in the classical case, if $\lambda_{1s}^{-1} - \lambda_{1j}^{-1} = \lambda_{2s}^{-1} - \lambda_{2j}^{-1}$, the coefficient $\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},js}$ obtained under a CPC model tends to zero at a rate faster than $n^{-1/2}$.
- $\mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DIF},j}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{CPC},j})$ and $\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{DIF}}) \geqslant \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{CPC}})$: In both cases equality is attained if $\boldsymbol{\mu}_{2s} = 0$, for $s \neq j$.

For the classical estimators, Flury and Schmid [17] noticed that in the particular case of the O'Neill [25] model, only the off-diagonal quadratic coefficients $\widehat{\boldsymbol{\Delta}}_{js}$ have smaller asymptotic variances under the CPC model, while identical results are obtained for the linear coefficients. This property also holds for our robust proposals.

When the underlying model is a proportional one, from Theorems 4.1 to 4.3 we have that

- $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF},jj}) = \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},jj}) > \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{PR},jj})$: As in the classical setting, CPC discrimination and ordinary quadratic discrimination yield the same asymptotic variances. For ρ_2 close to 1, $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{PR},jj})$ can become considerably smaller than $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},jj})$. On the other hand, when $\tau_1 = \tau_2 = 1/2$ and p is large, $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{PR},jj})$ can also become considerably smaller than $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},jj})$.
- $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF},js}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},js}) = \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{PR},js})$, for $j \neq s$: Moreover, when $\tau_1 = \tau_2 = 1/2$, we have that $\frac{1}{2}\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF},js}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{CPC},js}) = \mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{PR},js})$. On the other hand, for ρ_2 close to 1, the last two variances may become considerably smaller than $\mathrm{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\mathrm{DIF},js})$ and so, for these coefficients, using the more parsimonious model appears to have considerable advantage over ordinary quadratic discrimination.

Table 1
Relationship among the asymptotic variances under the different hierarchical models

Estimated coefficient	True Model		
	Level 2	Level 3	Level 4
$\widehat{\mathbf{\Delta}}_{jj}$	DIF = CPC	DIF = CPC > PR	DIF = CPC > PR > 0
$\widehat{\mathbf{\Delta}}_{js}$	DIF $\geqslant$ CPC	DIF $\geqslant$ CPC = PR	DIF > CPC = PR = 0
$\widehat{\boldsymbol{\alpha}}_j$	DIF $\geqslant$ CPC	DIF $\geqslant$ CPC $\geqslant$ PR	DIF $\geqslant$ CPC $\geqslant$ PR $\geqslant$ EQ
$\widehat{\xi}$	DIF $\geqslant$ CPC	DIF $\geqslant$ CPC $\geqslant$ PR	DIF $\geqslant$ CPC $\geqslant$ PR $\geqslant$ EQ

DIF, CPC, PR, and EQ indicate the model used to estimate the parameters, i.e., the model used for discrimination.

- $\mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DIF},j}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{CPC},j}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{PR},j})$: If $\boldsymbol{\mu}_2 = \mathbf{0}$ equality holds in all cases, otherwise some improvement may be expected.
- $\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{DIF}}) \geqslant \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{CPC}}) \geqslant \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{PR}})$: Moreover, $\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{CPC}}) = \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{PR}})$ if for some constant c, $\boldsymbol{\mu}_2 = c\left(\lambda_1^{1/2}, \ldots, \lambda_p^{1/2}\right)^{\mathrm{T}}$. In particular, if $\boldsymbol{\mu}_2 = \mathbf{0}$ we have $\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{DIF}}) = \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{CPC}}) = \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{PR}})$.

As in the classical case, these results suggest that using the proportional model, provided it is true, may be advantageous, particularly for large dimensions. Furthermore, CPC discrimination can also be expected to perform better than quadratic discrimination under these circumstances.

When the scatter matrices are equal, $\widehat{\mathbf{\Delta}}_{\mathrm{EQ},js} = 0$ for any j, s and

- $\mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{DIF},jj}) = \mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{CPC},jj}) > \mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{PR},jj}) > 0$: Note that in the robust setting, since σ_2 can be different from 0, we do not obtain the inequality $\mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{DIF},jj}) = \mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{CPC},jj}) > p\mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{PR},jj})$ as in the classical case.
- $\mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{DIF},js}) > \mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{CPC},js}) = \mathrm{ASVAR}(\widehat{\mathbf{\Delta}}_{\mathrm{PR},js}) = 0$ for $j \neq s$: Under both CPC and proportional discrimination, the variance of $\widehat{\mathbf{\Delta}}_{js}$ converges to zero at a rate faster than n^{-1}.
- $\mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{DIF},j}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{CPC},j}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{PR},j}) \geqslant \mathrm{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\mathrm{EQ},j})$: If $\boldsymbol{\mu}_2 = \mathbf{0}$ equality holds in all cases, otherwise a reduction of the variance can be attained by using one of the constrained models. As for the classical rule, the advantage of proportional and linear discrimination increases with the dimension.
- $\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{DIF}}) \geqslant \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{CPC}}) \geqslant \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{PR}}) \geqslant \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{EQ}})$. Note that, for this parameter, if $\boldsymbol{\mu}_2 = \mathbf{0}$, we have that $\mathrm{ASVAR}(\widehat{\xi}_{\mathrm{DIF}}) = \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{CPC}}) = \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{PR}}) > \mathrm{ASVAR}(\widehat{\xi}_{\mathrm{EQ}})$.

In Table 1 we summarize the above results concerning the relationships among the asymptotic variances along the hierarchical models on the scatter matrices. It is worth noticing that even when the asymptotic variances of the robust estimators are not proportional to their classical relatives, the relationships shown in Table 1 coincide with those obtained by Flury and Schmid [17], that is, the order is preserved.

5. Monte Carlo study

Up till now, we have considered asymptotic variances, but in the context of discrimination misclassification error rates are also important, especially for moderate to small sample sizes. In order to have a deeper insight into misclassification rates, we have performed a simulation study.

We have considered two populations of sizes $n_1 = n_2 = n = 20, 30, 40, 50, 75$ and 100 in dimension $p = 4$. The classification rules to be compared are:

- the ordinary quadratic rule denoted Q_{DIF};
- the quadratic classification rule under level 2, i.e., Q_{CPC};
- the quadratic classification rule under level 3, i.e., Q_{PR}; and
- the linear classification rule denoted Q_{EQ},

indicated as diamonds, squares, inverted triangles and circles combined with solid lines in all figures, respectively. The horizontal dashed line indicates the optimal error rate. All of them were computed using the sample mean and the sample covariance matrix and also using as robust estimators the Donoho–Stahel estimators with weight function the Huber's function with constant $\sqrt{\chi^2_p(0.95)}$ and the S-estimators using as ρ function the biweight Tukey's function calibrated to attain 25% breakdown point. The S-estimators were computed using the MATLAB programs provided in Christophe Croux's personal web site taking 1000 random p-subsets. To obtain approximately the worst direction for the Donoho–Stahel estimator, we have combined a search over 1000 random directions on the p-dimensional sphere together with 1000 directions using random p-subsets.

Since, $\pi_1 = \pi_2 = \frac{1}{2}$, the total misclassification error of a given rule Q, under the central model, equals

$$TPM(Q) = \frac{P_1\left(Q(\mathbf{y}) < 0\right) + P_2\left(Q(\mathbf{y}) > 0\right)}{2},$$

where P_i is the probability related to a $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. In order to estimate it, we have considered validation samples of size $m = 10\,000$. To be more precise, we have generated independent random variables $\mathbf{y}_{i1}, \ldots, \mathbf{y}_{im}$ with $\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, for $i = 1, 2$. For each observation $\mathbf{y}_{ij}$ we evaluated $Q_{ij} = Q(\mathbf{y}_{ij})$ and we have computed

$$\widehat{TPM}(Q) = \frac{\#\left\{Q_{1j} < 0\right\} + \#\left\{Q_{2j} > 0\right\}}{2m}.$$

We have performed 1000 replications and the mean of the estimated misclassification error over replications, $\widehat{TPM}(Q)$, was computed in order to compare the discrimination rules under different models and different contaminations.

We give a detailed description of the five designs considered. In all cases and without loss of generality, we have assumed that $\boldsymbol{\mu}_1 = \mathbf{0}$.

- **Design 1** (Efron's model): An optimal model for linear discrimination with $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ (see [15,18]). Under this model the advantage of using the linear rule over the CPC discrimination and the ordinary quadratic discrimination increases with the dimension p. On the other hand, the variances of the estimators of the linear and constant coefficients of Q_{PR} approach those obtained in linear discrimination when p increases. With respect to the quadratic coefficients, the same argument holds for the non-diagonal elements when using the classical methods while for the robust one, a term involving the coefficient σ_2 is always present. One expects that proportional discrimination will do as well as linear discrimination, for the classical rule and assessing the effect of using robust estimators is one of the goals of this simulation study. The parameters were chosen as $\boldsymbol{\mu}_2 = (3, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \text{diag}\,(1, 2, 8, 16)$, yielding an optimal error rate of 0.0668. The eigenvalues were chosen to be different to

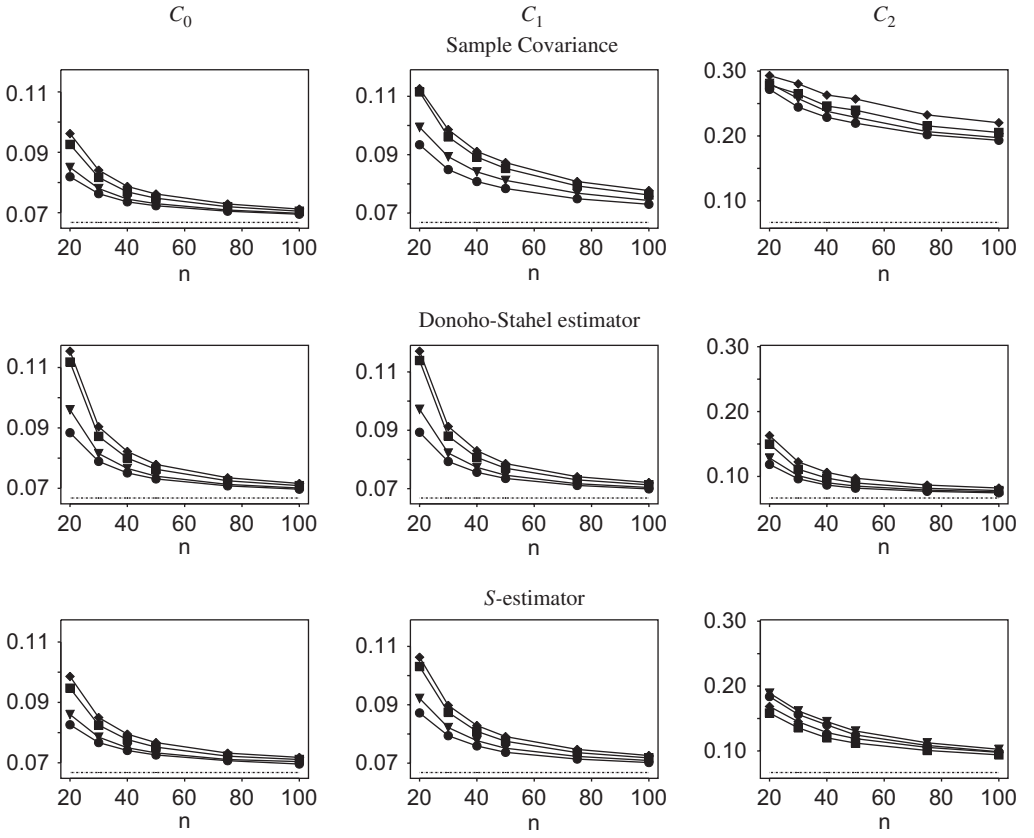

Fig. 4. Estimated misclassification rates under **Design 1**. Diamonds correspond to Q_{DIF}, squares to Q_{CPC}, inverted triangles to Q_{PR} and circles to the linear discrimination rule. The horizontal dashed line indicates the optimal error rate.

avoid convergence problems when solving the equations leading to the estimators under a CPC model.

- **Design 2** (A proportional model): In this case we have considered a design similar to design 1, but including a proportionality constant, i.e., we have chosen $\boldsymbol{\mu}_2 = (3, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\Sigma}_1 = \text{diag}\,(1, 2, 8, 16)$, $\boldsymbol{\Sigma}_2 = 4\boldsymbol{\Sigma}_1$. The optimal rate is now 0.0885.
- **Design 3** (O'Neill's model): This is a design based on a particular model studied by O'Neill [25] and considered in Flury et al. [18], for the purpose of comparing the performance of linear and ordinary quadratic classification rules. In this design $\boldsymbol{\mu}_2 = (\Delta, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\Sigma}_1 = \mathbf{I}_4$, $\boldsymbol{\Sigma}_2 = \text{diag}\left(\sigma^2, 1, 1, 1\right)$. We have chosen $\Delta = 4.5$ and $\sigma^2 = 9$ leading to an optimal error rate of 0.1073. O'Neill's model is a CPC model but not a proportional one and thus, both ordinary quadratic discrimination and CPC discrimination are theoretically correct. Optimal classification is quadratic only in the first variable. However, as the calculations given above suggest, the CPC discrimination is not expected to do much better than ordinary quadratic discrimination. On the other hand, in the classical case, O'Neill [25,26] noticed that it took a very large sample size for quadratic discrimination to improve linear discrimination, even if σ is so different than 1 as in our example.

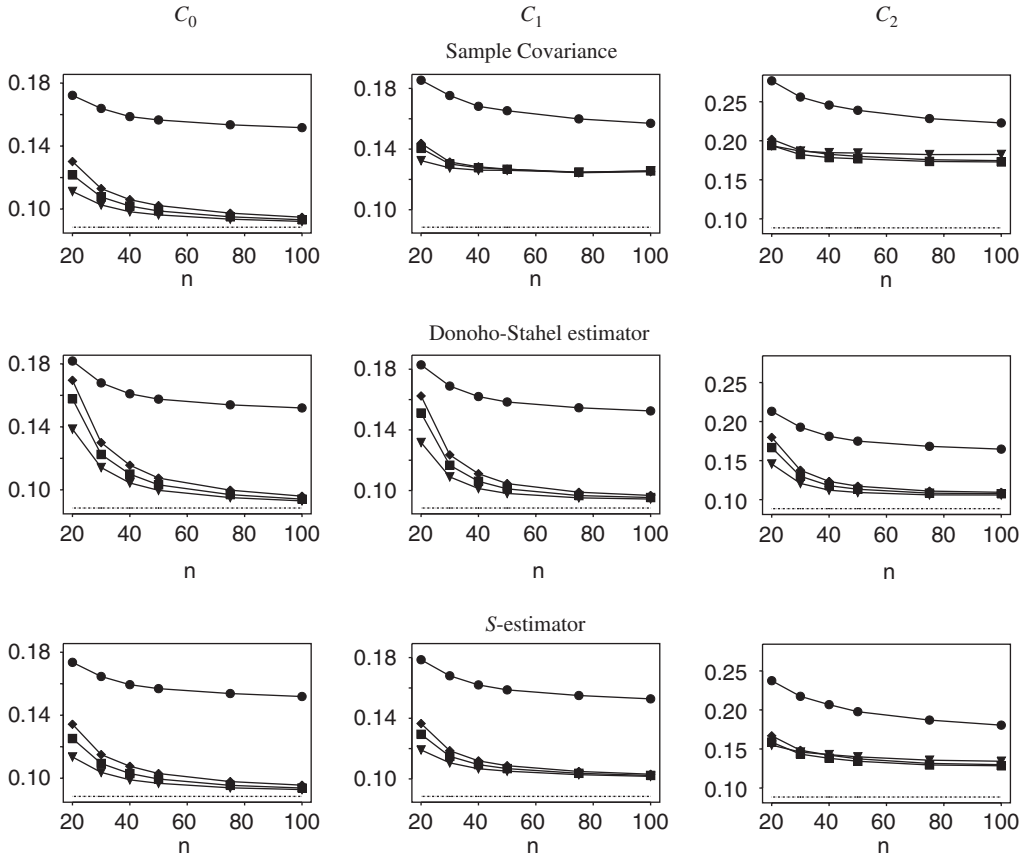

Fig. 5. Estimated misclassification rates under **Design 2**. Diamonds correspond to Q_{DIF}, squares to Q_{CPC}, inverted triangles to Q_{PR} and circles to the linear discrimination rule. The horizontal dashed line indicates the optimal error rate.

- **Design 4** (A CPC model): When the CPC model holds, both Q_{DIF} and Q_{CPC} are theoretically correct. As mentioned above, and as discussed in Flury and Smith [17], the asymptotic variances indicate that using Q_{CPC} does not necessarily yield estimates of the discriminant function coefficients with smaller variances than Q_{DIF}. The advantage of the method depends on the eigenvalues. For instance, if $\lambda_{1s} - \lambda_{1j} = \lambda_{2s} - \lambda_{2j}$, for all (s, j), then CPC discrimination and ordinary quadratic discrimination should do about equally well. On the other hand, if $\lambda_{1s}^{-1} - \lambda_{1j}^{-1} = \lambda_{2s}^{-1} - \lambda_{2j}^{-1}$, for all (s, j), then some quadratic coefficients have smaller asymptotic variances if estimated using the CPC model. Our parameter setup for the simulation study was taken as $\boldsymbol{\Sigma}_1 = \text{diag}\,(1/5, 1/2, 2/3, 5/6)$, $\boldsymbol{\Sigma}_2 = \text{diag}\,(1/4, 1, 2, 5)$ and $\boldsymbol{\mu}_2 = (1, 0, 0, 0)^{\text{T}}$ to study the improvement obtained with the robust CPC rule. The optimal error rate equals 0.0991.
- **Design 5** (A quadratic model): We have considered the same design as in Flury et al. [18] in which none of the CPC, proportional or linear discrimination rules are correct. We wanted the CPC model to be far from correct. A particular way to generate such models is to take

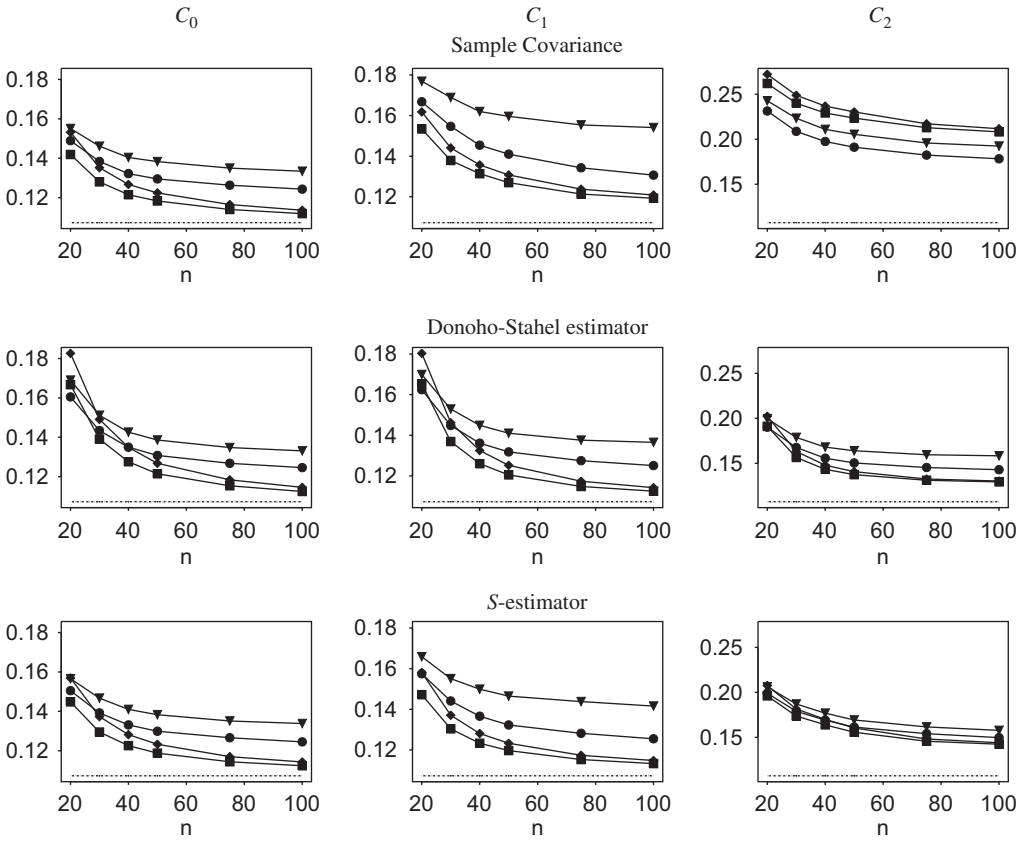

Fig. 6. Estimated misclassification rates under **Design 3**. Diamonds correspond to Q_{DIF}, squares to Q_{CPC}, inverted triangles to Q_{PR} and circles to the linear discrimination rule. The horizontal dashed line indicates the optimal error rate.

$$\boldsymbol{\mu}_2 = (2,0,0,0)^{\mathrm{T}},$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1/2 & 1/2 \\ 0 & 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1/2 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1/2 \end{pmatrix}$$

yielding an optimal error rate of 0.1278.

The results for normal data will be indicated by C_0, while two contaminations were studied

- C_1: $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in}$ are i.i.d. $0.9N_4(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + 0.1N_4(\boldsymbol{\mu}_i, 9\boldsymbol{\Sigma}_i)$.
- C_2: $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in}$ are i.i.d. $0.9N_4(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + 0.1N_4(\boldsymbol{\mu}_i + \boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ with $\boldsymbol{\mu} = (10,0,0,0)^{\mathrm{T}}$. The aim of this contamination is to see how the bias of parameter estimates affects the probability of misclassification.

Figs. 4–8 summarize the results of the simulation study. The results show the advantage of using robust procedures when contamination is present. For instance, under C_2, in most cases, the error

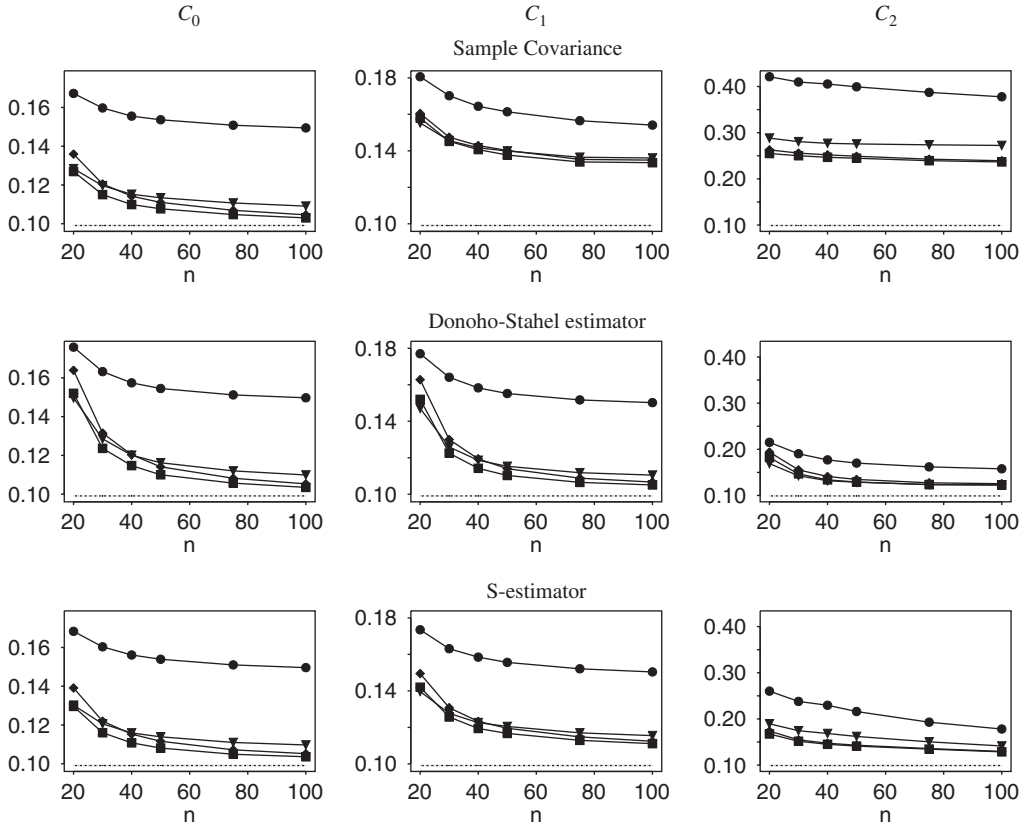

Fig. 7. Estimated misclassification rates under **Design 4**. Diamonds correspond to Q_{DIF}, squares to Q_{CPC}, inverted triangles to Q_{PR} and circles to the linear discrimination rule. The horizontal dashed line indicates the optimal error rate.

rates with the classical rules are over twice those of the uncontaminated situations. The robust procedures behave quite similarly under normal errors and the two contaminations considered. In general, contamination C_2 seems to be more harmful than contamination C_1. However, the error rates related to the Donoho–Stahel estimator are slightly smaller than those related to the S-estimator under C_2. Besides, for small sample sizes ($n = 20, 30$), the Donoho–Stahel rule shows in all cases larger rates under C_0 than that derived from the S-estimator. This performance of the Donoho–Stahel rule may be due to the difficulty to obtain the optimal direction for small sample sizes and also to the larger bias of the estimator for small sample sizes. On the other hand, as expected, under C_0, the advantage of using the classical rule over the robust ones decreases as the sample size increases. Moreover, under C_0, the conclusions obtained in Flury et al. [18] hold for both the classical and robust discrimination rules. To summarize,

- **Design 1**: If equality holds, under C_0, then the linear discrimination is the best one, but not much is lost if proportional discrimination is used. If CPC or ordinary quadratic discrimination are used, approximately twice the observations are needed to obtain the same error rate $\widehat{TPM}$. These conclusions remain valid for the robust procedure based on the Donoho–Stahel estimator even under both contaminations and for that based on the S-scatter under C_0 and C_1. Surprisingly,

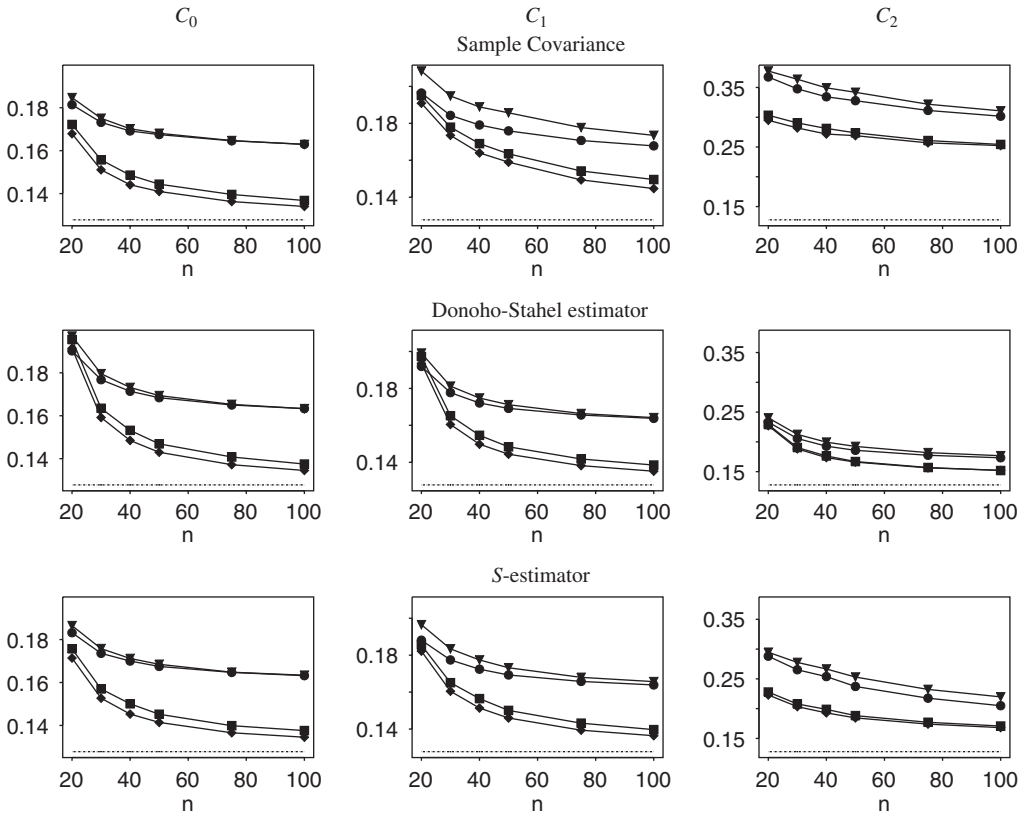

Fig. 8. Estimated misclassification rates under **Design 5**. Diamonds correspond to Q_{DIF}, squares to Q_{CPC}, inverted triangles to Q_{PR} and circles to the linear discrimination rule. The horizontal dashed line indicates the optimal error rate.

under C_2 the rule based on the S-estimator gives better rates with the CPC and quadratic discrimination than for the linear one.

- **Design 2**: All methods except the linear discrimination are theoretically correct, but clearly, under C_0, Q_{PR} performs much better than either Q_{CPC} or Q_{DIF}. This conclusion remains valid for the robust Donoho–Stahel procedure under both contaminations and for the S-rule under C_1, while for the classical procedure all rules perform similarly bad for data faraway from normality. It is worth noticing that, under C_2, when using the S-estimator, all the quadratic discrimination rates are quite similar, however, the lower rates are attained by CPC discrimination.
- **Design 3**: One would expect that the two theoretically correct rules, Q_{CPC} and Q_{DIF} performed considerably better than the inappropriate proportional and linear methods. However, under C_0, all methods perform quite similarly, and only for sample sizes larger than 40, Q_{DIF} starts to perform better than the linear rule. On the other hand, Q_{CPC} appears to have a noticeable advantage over Q_{DIF} while the linear rule performs better than Q_{PR}. As mentioned in Flury et al. [18], a possible explanation for this unexpected phenomenon is that Q_{PR} introduces the wrong flexibility, compared to the linear discriminant rule. Proportional discrimination forces the boundary of the classification regions to be genuinely quadratic, which is undesirable in this case. This model corresponds to the situation where the direction of the mean difference is

identical to the direction of the difference in variance. More precisely, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is proportional to the eigenvector of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$ related to the single eigenvalue that is different from 1. These comments remain valid for both robust procedures, not only under C_0 but also under C_1 and C_2. However, the classical rule, under C_2 reverses the conclusions since the best error rate is attained by the linear discrimination rule followed by Q_{PR} and by both the CPC and the ordinary quadratic rule that performed quite similarly. In this design, none of the methods seems suitable to handle the situation due to their slow convergence to the optimal rate, under C_0.

- **Design 4**: This design was tailored to favor CPC discrimination. Under C_0, CPC beats the ordinary discrimination rule for small sample sizes. Note that for $n = 20$ the proportional rule performs better than the theoretically correct CPC rule, when using the robust procedure based on the Donoho–Stahel estimator. In both cases, robust and non-robust, the proportional discrimination rule performs surprisingly well. In particular, for sample sizes lower than 40 it gives better rates than ordinary quadratic discrimination. This is quite outstanding in view of the fact that the variance ratios range from 1.25 to 6 and thus, the two scatter matrices are far from being proportional. This behavior underlines the usefulness of proportional discrimination due to its flexibility when introducing only a single parameter for each additional group. Under C_1, the same behavior is observed for all procedures, robust or not, while under C_2, the classical proportional discrimination rule performs much worse than the other two quadratic rules. The conclusions described for the behavior of the robust proposal under C_0, also hold under C_2.
- **Design 5**: In this case, the appropriate rule is the ordinary quadratic method. All three constrained methods are theoretically wrong. However, the linear discrimination rule shows its advantage over the proportional one under the three distributions considered. The ordinary quadratic rule performs better than the CPC rule for both the classical and robust procedures and all contaminations. Besides, it should be noticed that, under C_2, the error rates of Q_{DIF} and Q_{CPC} are almost the same. Note that Q_{DIF} and Q_{CPC} perform much better than proportional and linear discrimination.

6. Final comments

In this paper we have studied robust methods for discriminating between two groups of elliptical observations, considering several levels of dissimilarities of the scatter matrices.

We have shown, both theoretically and by means of a simulation study, the advantage of using robust procedures over classical ones, especially if the data deviate from multivariate normality.

Our results have also shown that, in some cases, better rates of misclassification can be achieved if a more parsimonious model among all the correct ones is used. Therefore, an important issue is to assess the adequacy of each of the different hierarchical levels. Classical tests for selecting a level within this hierarchy are presented in Flury [16]. Robust versions of those tests have been proposed recently by Boente et al. [6].

Acknowledgments

The authors would like to thank two Referees for their valuable comments and suggestions that lead to improve the presentation of the paper. This research was partially supported by Grants X-094 from the Universidad de Buenos Aires, PID 5505 from CONICET and PAV 120 and PICT 21407 from ANPCYT, Argentina and by *Programa Operacional* "*Ciência*, *Tecnologia*, *Inovação*" (POCTI) of the *Fundação para a Ciência e a Tecnologia* (FCT), cofinanced by the European

Community fund FEDER. It was also supported by the joint cooperation program ANPCYT–GRICES PO/PA04-EIII/020.

Appendix

Note that (10), (11), (12), (17), (18), (24) and (25) follow immediately from (9), (16) and (23), respectively. In order to prove Theorems 3.2 and 3.3 it will be enough to derive the expressions for the partial influence functions of the matrices $\mathbf{S}_{\text{CPC},\ell}$ and $\mathbf{S}_{\text{R},\ell}$, $\ell = 1, 2$. These partial influence functions follow immediately from their definitions and from the partial influence functions of the common eigenvectors and the eigenvalues given in Boente et al. [5] under a CPC model and in Boente et al. [2] under a proportional model.

Note that **A1** and **A2** imply that

$$\begin{cases} \text{ASCOV}\left(\mathbf{V}_{i,js}, \mathbf{V}_{i,m\ell}\right) = \text{ASCOV}\left(\mathbf{V}_{i,js}, \mathbf{V}_{i,mm}\right) = 0 & \text{for } j < s,\ m < \ell \text{ and } (j,s) \neq (m,\ell), \\ \text{ASCOV}\left(\mathbf{V}_{i,jj}, \mathbf{V}_{i,ss}\right) = \sigma_2 \lambda_{ij}\lambda_{is} & \text{for } j < s, \\ \text{ASVAR}\left(\mathbf{V}_{i,jj}\right) = (2\sigma_1 + \sigma_2)\lambda_{ij}^2, & \\ \text{ASVAR}\left(\mathbf{V}_{i,js}\right) = \sigma_1 \lambda_{ij}\lambda_{is}, & \\ \text{ASCOV}\left(\mathbf{V}_{i,s\ell}, \widehat{\boldsymbol{\mu}}_{ij}\right) = \text{ASCOV}\left(\widehat{\boldsymbol{\mu}}_{ir}, \widehat{\boldsymbol{\mu}}_{ij}\right) = 0 & \text{for } j \neq r, \\ \text{ASVAR}\left(\widehat{\boldsymbol{\mu}}_{ij}\right) = \sigma_3 \lambda_{ij}. & \end{cases} \tag{38}$$

Proof of Theorem 4.1. Its proof follows immediately, using (26), (38) and that

$$\begin{aligned}
\text{ASVAR}(\widehat{\boldsymbol{\Delta}}_{\text{DIF},js}) &= \frac{1}{4}\sum_{i=1}^{2} \frac{1}{\tau_i} \frac{1}{\lambda_{ij}^2 \lambda_{is}^2} \text{ASVAR}(\mathbf{V}_{i,js}), \\
\text{ASVAR}(\widehat{\boldsymbol{\alpha}}_{\text{DIF},j}) &= \sum_{i=1}^{2} \frac{1}{\tau_i \lambda_{ij}} \sigma_3 + \sum_{i=1}^{2}\sum_{s=1}^{p} \frac{1}{\tau_i} \frac{\boldsymbol{\mu}_{is}^2}{\lambda_{ij}^2 \lambda_{is}^2} \text{ASVAR}(\mathbf{V}_{i,js}) \\
&\quad + \sum_{i=1}^{2} \frac{1}{\tau_i \lambda_{ij}^2} \sum_{s \neq \ell} \frac{\boldsymbol{\mu}_{is}\boldsymbol{\mu}_{i\ell}}{\lambda_{is}\lambda_{i\ell}} \text{ASCOV}(\mathbf{V}_{i,js}, \mathbf{V}_{i,j\ell}), \\
\text{ASVAR}(\widehat{\xi}_{\text{DIF}}) &= \sum_{i=1}^{2} \frac{1}{\tau_i} \sum_{j=1}^{p} \frac{\boldsymbol{\mu}_{ij}^2}{\lambda_{ij}^2} \text{ASVAR}(\widehat{\boldsymbol{\mu}}_{ij}) + \frac{1}{4}\sum_{i=1}^{2} \frac{1}{\tau_i} \sum_{j=1}^{p} \left(1 - \frac{\boldsymbol{\mu}_{ij}^2}{\lambda_{ij}}\right)^2 \frac{1}{\lambda_{ij}^2} \text{ASVAR}(\mathbf{V}_{i,jj}) \\
&\quad + \frac{1}{4}\sum_{i=1}^{2} \frac{1}{\tau_i} \sum_{j \neq s} \left(1 - \frac{\boldsymbol{\mu}_{ij}^2}{\lambda_{ij}}\right)\left(1 - \frac{\boldsymbol{\mu}_{is}^2}{\lambda_{is}}\right) \frac{1}{\lambda_{ij}\lambda_{is}} \text{ASCOV}\left(\mathbf{V}_{i,jj}, \mathbf{V}_{i,ss}\right) \\
&\quad + \sum_{i=1}^{2} \frac{1}{\tau_i} \sum_{j<s} \frac{\boldsymbol{\mu}_{ij}^2 \boldsymbol{\mu}_{is}^2}{\lambda_{ij}^2 \lambda_{is}^2} \text{ASVAR}\left(\mathbf{V}_{i,js}\right) \\
&= \sum_{i=1}^{2} \frac{1}{\tau_i} v_{i1} + \frac{1}{4}\sum_{i=1}^{2} \frac{1}{\tau_i} v_{i2} + \sum_{i=1}^{2} \frac{1}{\tau_i} v_{i3}. \qquad \square
\end{aligned}$$

Proof of Theorem 4.2. Using (26), we get that the asymptotic variance of $\widehat{\mathbf{\Lambda}}_{\text{CPC}}$ is given by $\sum_{i=1}^{2}(1/\tau_i)E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{D}_{\text{CPC},js}, F)\right]^2$. For any $1 \leqslant j, s \leqslant p$, we have that

$$
\begin{aligned}
&E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{D}_{\text{CPC},js}, F)\right]^2 \\
&= \frac{1}{4}\left\{\frac{1}{\lambda_{2j}^2\lambda_{2s}^2}E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{CPC},2,js}, F)\right]^2\right. \\
&\left.+\frac{1}{\lambda_{1j}^2\lambda_{1s}^2}E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{CPC},1,js}, F)\right]^2\right\} \\
&-\frac{1}{2}\frac{1}{\lambda_{1j}\lambda_{1s}}\frac{1}{\lambda_{2j}\lambda_{2s}}E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{CPC},2,js}, F)\text{PIF}_i(\mathbf{x}, \mathbf{S}_{\text{CPC},1,js}, F)\right].
\end{aligned}
$$

When $j = s$, the above expression, (13) and (38) entail (30). Let us consider now the case when $j \neq s$. From (14), we derive that

$$
\begin{aligned}
&E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{D}_{\text{CPC},js}, F)\right]^2 \\
&= \frac{\tau_i^2}{4}\theta_{js}^2\frac{\left(\lambda_{ij}-\lambda_{is}\right)^2}{\lambda_{ij}^2\lambda_{is}^2}\left(\frac{\lambda_{2j}-\lambda_{2s}}{\lambda_{2j}\lambda_{2s}}-\frac{\lambda_{1j}-\lambda_{1s}}{\lambda_{1j}\lambda_{1s}}\right)^2 E_{F_i}\left[\text{IF}(\mathbf{x}, \mathbf{\Upsilon}_{i,js}, F_i)\right]^2.
\end{aligned}
$$

Hence, using again (38), straightforward calculations allow to derive (30). In order to prove (31) we have to compute $E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{a}_{\text{CPC},j}, F)\right]^2$. From the expressions of the partial influence functions of the functional given in (11) and from (38), we obtain easily that

$$
E_{F_i}\left[\text{PIF}_i(\mathbf{x}, \mathbf{a}_{\text{CPC},j}, F)\right]^2 = \frac{1}{\lambda_{ij}^2}\left\{\lambda_{ij}\sigma_3 + \boldsymbol{\mu}_{ij}^2(2\sigma_1+\sigma_2) + \sigma_1\sum_{s\neq j}\boldsymbol{\mu}_{is}^2\frac{\lambda_{ij}}{\lambda_{is}}\right\},
$$

which entails (31).

Finally, the expression for $\text{ASVAR}(\widehat{\xi}_{\text{CPC}})$, follows easily using that $\text{PIF}_i(\mathbf{x}, c_{\text{CPC}}, F) = (-1)^i\,(P_{i1} + P_{i2}/2) - \tau_i P_{i4}$, where

$$
P_{i1} = \sum_{j=1}^{p}\frac{\boldsymbol{\mu}_{ij}}{\lambda_{ij}}IF(\mathbf{x}, \mathbf{m}_{ij}, F_i), \tag{39}
$$

$$
P_{i2} = \sum_{j=1}^{p}\frac{1}{\lambda_{ij}}IF(\mathbf{x}, \mathbf{\Upsilon}_{i,jj}, F_i)\left(1-\frac{\boldsymbol{\mu}_{ij}^2}{\lambda_{ij}}\right), \tag{40}
$$

$$
P_{i4} = \sum_{j<s}\theta_{sj}\eta_{sj}\frac{\lambda_{ij}-\lambda_{is}}{\lambda_{ij}\lambda_{is}}IF(\mathbf{x}, \mathbf{\Upsilon}_{i,js}, F_i), \tag{41}
$$

with θ_{sj} defined in Theorem 3.2, and $\eta_{js} = \sum_{i=1}^{2}(-1)^i \boldsymbol{\mu}_{ij}\boldsymbol{\mu}_{is}(\lambda_{ij} - \lambda_{is})/(\lambda_{ij}\lambda_{is})$. Details can be found in Bianco et al. [1]. □

The proofs of Theorems 4.3 and 4.4 follow using similar arguments. Details can be found in Bianco et al. [1].

References

[1] A. Bianco, G. Boente, A.M. Pires, I.M. Rodrigues, Robust discrimination under a hierarchy on the scatter matrices, Working Paper, 2007, Universidad de Buenos Aires. Available at ⟨http://www.ic.fcen.uba.ar/preprints/biancoboentepiresrodrigues.pdf⟩.

[2] G. Boente, F. Critchley, L. Orellana, Influence functions for robust estimators under proportional scatter matrices, Statist. Methods Appl. 15 (2007) 295–327.

[3] G. Boente, L. Orellana, A robust approach to common principal components, in: L.T. Fernholz, S. Morgenthaler, W. Stahel (Eds.), Statistics in Genetics and in the Environmental Sciences, Birkhauser, Basel, 2001, pp. 117–147.

[4] G. Boente, L. Orellana, Robust plug-in estimators in proportional scatter models, J. Statist. Plann. Inference 122 (2004) 95–110.

[5] G. Boente, A.M. Pires, I.M. Rodrigues, Influence functions and outlier detection under the common principal components model: a robust approach, Biometrika 89 (2002) 861–875.

[6] G. Boente, A.M. Pires, I.M. Rodrigues, Robust tests for the common principal components model, Working Paper, 2005, Universidad de Buenos Aires. Available at ⟨http://www.ic.fcen.uba.ar/preprints/boentepiresrodrigues2.pdf⟩.

[7] G. Boente, A.M. Pires, I.M. Rodrigues, General projection-pursuit estimators for the common principal components model: influence functions and Monte Carlo study, J. Multivariate Anal. 97 (2006) 124–147.

[8] N.A. Campbell, The influence function as an aid in outlier detection in discriminant analysis, Appl. Statist. 27 (1978) 251–258.

[9] F. Critchley, C. Vitiello, The influence of observations on misclassification probability estimates in linear discriminant analysis, Biometrika 78 (1991) 677–690.

[10] C. Croux, C. Dehon, Robust linear discriminant analysis using S-estimators, Canad. J. Statist. 29 (2001) 473–492.

[11] C. Croux, P. Filzmoser, K. Joossens, Classification efficiencies for robust linear discriminant analysis, Statist. Sinica (2007), to appear.

[12] C. Croux, G. Haesbroeck, Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, Biometrika 87 (2000) 603–618.

[13] C. Croux, K. Joossens, Influence of observations on the misclassification probability in quadratic discriminant analysis, J. Multivariate Anal. 96 (2005) 384–403.

[14] D.L. Donoho, Breakdown properties of multivariate location estimators, Ph.D. Qualifying Paper, Harvard University, 1982.

[15] B. Efron, The efficiency of logistic regression compared to normal discrimination analysis, J. Amer. Statist. Assoc. 70 (1975) 892–898.

[16] B.K. Flury, Common Principal Components and Related Multivariate Models, Wiley, New York, 1988.

[17] B.K. Flury, M.J. Schmid, Quadratic discriminant functions with constraints on the covariances matrices: some asymptotic results, J. Multivariate Anal. 40 (1992) 244–261.

[18] B.K. Flury, M.J. Schmid, A. Narayanan, Error rates in quadratic discrimination with constraints on the covariance matrices, J. Classification 11 (1994) 101–120.

[19] W.K. Fung, Diagnostics in linear discriminant analysis, J. Amer. Statist. Assoc. 90 (1995) 952–956.

[20] W.K. Fung, Diagnosing influential observations in quadratic discriminant analysis, Biometrics 52 (1996) 1235–1241.

[21] D. Gervini, The influence function of the Donoho–Stahel estimator of multivariate location and scale, Statist. Probab. Lett. 60 (2002) 425–435.

[22] P.A. Lachenbruch, Note on initial misclassification effects on the quadratic discriminant function, Technometrics 21 (1979) 129–132.

[23] H.P. Lopuhaä, Breakdown points and asymptotic properties of multivariate S-estimators and τ-estimators: a summary, in: W. Stahel, S. Weisberg (Eds.), Directions in Robust Statistics and Diagnostics, Part I, Springer, New York, 1991, pp. 167–182.

[24] R. Maronna, D. Martin, V. Yohai, Robust Statistics: Theory and Methods, Wiley, New York, 2006.

[25] T.J. O'Neill, A theoretical method of comparing classification rules under non-optimal conditions with application to the estimates of Fisher's linear and quadratic discriminant rules under unequal covariances matrices, Technical Report No. 217, 1984, Department of Statistics, Stanford University.
[26] T.J. O'Neill, Error rates of non-Bayes classification rules and robustness of Fisher's linear discriminant function, Biometrika 79 (1992) 177–184.
[27] A.M. Pires, J. Branco, Partial influence functions, J. Multivariate Anal. 83 (2002) 451–468.
[28] W. Stahel, Robust estimation: infinitesimal optimality and covariance matrix estimation, Thesis, ETH, Zurich, 1981, (in German).
[29] D.E. Tyler, Radial estimates and the test for sphericity, Biometrika 69 (1982) 429–436.
[30] H.P. Lopuhaä (1989). On the relation between *S*-estimators and *M*-estimators of multivariate location and covariance. *Annals of Statistics*, **17**, 1662–1683.