Journal of Multivariate Analysis 133 (2015) 356–376

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A robust predictive approach for canonical correlation analysis

Jorge G. Adrover [a,*], Stella M. Donato [b]

[a] FAMAF, Universidad Nacional de Córdoba, CIEM and CONICET, Argentina
[b] Instituto de Cálculo, Universidad de Buenos Aires and CONICET, Argentina

ARTICLE INFO

Article history:
Received 10 June 2013
Available online 5 November 2014

AMS 1991 subject classifications:
primary 62F35
secondary 62H12.0

Keywords:
Canonical correlation analysis
S-estimation
M-scales
Mean relative prediction error

ABSTRACT

Canonical correlation analysis (CCA) is a dimension-reduction technique in which two random vectors from high dimensional spaces are reduced to a new pair of low dimensional vectors after applying linear transformations to each of them, retaining as much information as possible. The components of the transformed vectors are called canonical variables. One seeks linear combinations of the original vectors maximizing the correlation subject to the constraint that they are to be uncorrelated with the previous canonical variables within each vector. By these means one actually gets two transformed random vectors of lower dimension whose expected square distance has been minimized subject to have uncorrelated components of unit variance within each vector. Since the closeness between the two transformed vectors is evaluated through a highly sensitive measure to outlying observations as the mean square loss, the linear transformations we are seeking are also affected. In this paper we use a robust univariate dispersion measure (like an M-scale) based on the distance of the transformed vectors to derive robust S-estimators for canonical vectors and correlations. An iterative algorithm is performed by exploiting the existence of efficient algorithms for S-estimation in the context of Principal Component Analysis. Some convergence properties are analyzed for the iterative algorithm. A simulation study is conducted to compare the new procedure with some other robust competitors available in the literature, showing a remarkable performance. We also prove that the proposal is Fisher consistent.

1. Introduction

Principal component analysis (PCA) and canonical correlation analysis (CCA) are two dimension-reduction techniques of widespread use in statistics. Though the principal component analysis relates to an internal analysis, i.e. within-group spectral decomposition for the study of dispersion, and the canonical correlations to an external analysis, i.e. between-group interrelations or correlations, conceptually they are interrelated. We will further explore this relationship. For a random vector $\mathbf{x}$ in the Euclidean space of dimension q, with positive definite dispersion matrix Σ, PCA looks for the spectral decomposition of Σ, the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_q$ associated with the corresponding eigenvalues in decreasing order $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_q > 0$, that is,

$$\Sigma = \sum_{i=1}^{q} \delta_i \mathbf{v}_i \mathbf{v}_i^t. \tag{1}$$

[*] Corresponding author.
E-mail addresses: adrover@famaf.unc.edu.ar (J.G. Adrover), stelladonato@yahoo.com.ar (S.M. Donato).

http://dx.doi.org/10.1016/j.jmva.2014.09.007
0047-259X/© 2014 Elsevier Inc. All rights reserved.

The variables $\mathbf{v}_1^t(\mathbf{x}-E\mathbf{x}),\ldots,\mathbf{v}_q^t(\mathbf{x}-E\mathbf{x})$ are usually referred as principal components. The spectral decomposition gives the orthonormal directions of maximum dispersion for $\mathbf{x}$, where the eigenvalues and eigenvectors can be defined through an optimization scheme,

$$\begin{aligned}\delta_1 &= \max_{\mathbf{a}\in\mathbb{R}^q,\ \|\mathbf{a}\|=1} \mathrm{Var}(\mathbf{a}^t(\mathbf{x}-E\mathbf{x})), \qquad \mathbf{v}_1 = \arg\max_{\mathbf{a}\in\mathbb{R}^q,\ \|\mathbf{a}\|=1} \mathrm{Var}(\mathbf{a}^t(\mathbf{x}-E\mathbf{x})) \qquad (2)\\ \delta_j &= \max_{\|\mathbf{a}\|=1,\ \mathrm{Cov}(\mathbf{a}^t(\mathbf{x}-E\mathbf{x}),\mathbf{v}_k^t(\mathbf{x}-E\mathbf{x}))=0,\ k=1,\ldots,j-1} \mathrm{Var}(\mathbf{a}^t(\mathbf{x}-E\mathbf{x})), \quad j>1\\ \mathbf{v}_j &= \arg\max_{\|\mathbf{a}\|=1,\ \mathrm{Cov}(\mathbf{a}^t(\mathbf{x}-E\mathbf{x}),\mathbf{v}_k^t(\mathbf{x}-E\mathbf{x}))=0,\ k=1,\ldots,j-1} \mathrm{Var}(\mathbf{a}^t(\mathbf{x}-E\mathbf{x})),\end{aligned}$$

where Var and Cov stand for the variance and the covariance operators for random variables. On the other hand, the principal components are the best linear predictors for $\mathbf{z}=\mathbf{x}-E\mathbf{x}$ when looking for linear combinations $\sum_{k=1}^p(\mathbf{a}_k^t\mathbf{z})\mathbf{a}_k$ based on an orthonormal set $\left\{\mathbf{a}_1,\ldots,\mathbf{a}_p,\mathbf{a}_{p+1},\ldots,\mathbf{a}_q\right\},\ p<q$. More precisely, principal components solve the optimization problem

$$\begin{aligned}(\boldsymbol{\mu}_\mathbf{x},V_p) &= \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p,V} E\,\|(\mathbf{x}-\boldsymbol{\mu})-P_V(\mathbf{x}-\boldsymbol{\mu})\|^2\\ &= \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^p,V} E\left\|P_{V^\perp}(\mathbf{x}-\boldsymbol{\mu})\right\|^2, \qquad (3)\end{aligned}$$

where P_V stands for the orthogonal projection on a subspace V of dimension $p<q$, $V=\langle\mathbf{a}_1,\ldots,\mathbf{a}_p\rangle$ means that V is generated by the orthonormal set $\left\{\mathbf{a}_1,\ldots,\mathbf{a}_p\right\}$ and $V^\perp=\langle\mathbf{a}_{p+1},\ldots,\mathbf{a}_q\rangle$ denotes the orthogonal complement of V. Then, the solutions $(\boldsymbol{\mu}_\mathbf{x},V_p)$ for (3) are given by

$$\boldsymbol{\mu}_\mathbf{x}=E\mathbf{x}, \qquad V_p=\langle\mathbf{v}_1,\ldots,\mathbf{v}_p\rangle \quad \text{and} \quad P_{V_p}(\mathbf{z})=\sum_{k=1}^p(\mathbf{v}_k^t\mathbf{z})\mathbf{v}_k.$$

CCA was proposed by Hotelling [10] to determine the relationship between two sets of variables obtained by transforming the vectors $\mathbf{x}$ and $\mathbf{y}$ into two vectors $\mathbf{z}$ and $\mathbf{w}$ in lower dimensions whose association has been greatly strengthened (see Das and Sen [5] for a very thorough account on CCA and their wide variety of applications). In recent years, CCA has also gained popularity as a method for the analysis of genomic data, since CCA has the potential to be a powerful tool for identifying relationships between genotype and gene expression. It has also been used in geostatistical applications (see Furrer and Genton [8]). CCA is closely related to multivariate regression when the vectors $\mathbf{x}$ and $\mathbf{y}$ are not treated symmetrically (see Yohai and García Ben [20]). Given the two random vectors $\mathbf{x}$ and $\mathbf{y}$ of dimensions p and q respectively, with dispersion matrix given by

$$\Sigma=\begin{pmatrix}E(\mathbf{x}-E\mathbf{x})(\mathbf{x}-E\mathbf{x})^t & E(\mathbf{x}-E\mathbf{x})(\mathbf{y}-E\mathbf{y})^t\\ E(\mathbf{y}-E\mathbf{y})(\mathbf{x}-E\mathbf{x})^t & E(\mathbf{y}-E\mathbf{y})(\mathbf{y}-E\mathbf{y})^t\end{pmatrix}=\begin{pmatrix}\Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}}\\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}}\end{pmatrix}, \qquad (4)$$

$$\det(\Sigma_{\mathbf{xx}})>0<\det(\Sigma_{\mathbf{yy}}), \qquad 0<r=\mathrm{rank}(\Sigma_{\mathbf{xy}})\le\min(p,q)=s. \qquad (5)$$

CCA seeks linear combinations of the variables in $\mathbf{x}$ and the variables in $\mathbf{y}$ that are maximally correlated with each other, that is, the first canonical vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ are defined (except for the signs) as

$$\left(\boldsymbol{\alpha}_1,\boldsymbol{\beta}_1\right)=\mathop{\arg\max}_{(\mathbf{a},\mathbf{b})\in(\mathbb{R}^p-\{\mathbf{0}\})\times(\mathbb{R}^q-\{\mathbf{0}\})}\mathrm{Corr}\left(\mathbf{a}^t\mathbf{x},\mathbf{b}^t\mathbf{y}\right). \qquad (6)$$

Since the correlation measure is scale invariant, we can define the first canonical vectors $\boldsymbol{\alpha}_1,\boldsymbol{\beta}_1$ as solutions to the optimization problem,

$$\left(\boldsymbol{\alpha}_1,\boldsymbol{\beta}_1\right)=\mathop{\arg\max}_{(\mathbf{a},\mathbf{b})\in\mathcal{A}_1}\mathrm{Corr}\left(\mathbf{a}^t\mathbf{x},\mathbf{b}^t\mathbf{y}\right), \qquad (7)$$

with

$$\mathcal{A}_1=\left\{(\mathbf{a},\mathbf{b})\in\mathbb{R}^p\times\mathbb{R}^q:\mathrm{Var}\left(\mathbf{a}^t\mathbf{x}\right)=\mathbf{a}^t\Sigma_{\mathbf{xx}}\mathbf{a}=1,\mathrm{Var}\left(\mathbf{b}^t\mathbf{y}\right)=\mathbf{b}^t\Sigma_{\mathbf{yy}}\mathbf{b}=1\right\}. \qquad (8)$$

The variables $\boldsymbol{\alpha}_1^t(\mathbf{x}-E\mathbf{x})$ and $\boldsymbol{\beta}_1^t(\mathbf{y}-E\mathbf{y})$ are called the first canonical variables and its positive correlation $\rho_1=\mathrm{Corr}(\boldsymbol{\alpha}_1^t\mathbf{x},\boldsymbol{\beta}_1^t\mathbf{y})$ is called the first canonical correlation. Canonical vectors and variables of higher order are defined recursively. Given $k>1$, let us take the first $k-1$ canonical variables $\boldsymbol{\alpha}_1^t(\mathbf{x}-E\mathbf{x}),\ldots,\boldsymbol{\alpha}_{k-1}^t(\mathbf{x}-E\mathbf{x})$ and $\boldsymbol{\beta}_1^t(\mathbf{y}-E\mathbf{y}),\ldots,\boldsymbol{\beta}_{k-1}^t(\mathbf{y}-E\mathbf{y})$ based on canonical vectors $\{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_{k-1}\}\subset\mathbb{R}^p$ and $\left\{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{k-1}\right\}\subset\mathbb{R}^q$. Then, the kth canonical variables $\boldsymbol{\alpha}_k^t(\mathbf{x}-E\mathbf{x})$ and $\boldsymbol{\beta}_k^t(\mathbf{y}-E\mathbf{y})$ can be obtained by seeking the vectors $\boldsymbol{\alpha}_k\in\mathbb{R}^p$ and $\boldsymbol{\beta}_k\in\mathbb{R}^q$ so that the linear combinations $\boldsymbol{\alpha}_k^t\mathbf{x}$ and $\boldsymbol{\beta}_k^t\mathbf{y}$ with unit variance, uncorrelated to $\boldsymbol{\alpha}_1^t\mathbf{x},\ldots,\boldsymbol{\alpha}_{k-1}^t\mathbf{x}$ and $\boldsymbol{\beta}_1^t\mathbf{y},\ldots,\boldsymbol{\beta}_{k-1}^t\mathbf{y}$, maximize the correlation coefficient between them. More precisely, we look for vectors defined as

$$\left(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\right)=\mathop{\arg\max}_{(\mathbf{a},\mathbf{b})\in\mathcal{A}_k}\mathrm{Corr}\left(\mathbf{a}^t\mathbf{x},\mathbf{b}^t\mathbf{y}\right), \qquad (9)$$

with

$$\mathcal{A}_k = \left\{ (\mathbf{a},\mathbf{b}) \in \mathbb{R}^p \times \mathbb{R}^q : \begin{array}{cc} \mathrm{Var}(\mathbf{a}^t\mathbf{x}) = 1, & \mathrm{Corr}(\mathbf{a}^t\mathbf{x}, \boldsymbol{\alpha}_j^t\mathbf{x}) = 0, \\ \mathrm{Corr}(\mathbf{b}^t\mathbf{y}, \boldsymbol{\beta}_j^t\mathbf{y}) = 0, & \mathrm{Var}(\mathbf{b}^t\mathbf{y}) = 1, \\ j = 1, 2, \ldots, k-1 & \end{array} \right\}. \tag{10}$$

If ρ_k stands for the positive correlation between $\boldsymbol{\alpha}_k^t\mathbf{x}$ and $\boldsymbol{\beta}_k^t\mathbf{y}$ (the kth canonical correlation), then $\rho_k^2 = \left(\mathrm{Corr}(\boldsymbol{\alpha}_k^t\mathbf{x}, \boldsymbol{\beta}_k^t\mathbf{y})\right)^2$ and one gets a decreasing sequence of squared canonical correlations, $\rho_1^2 \geq \cdots \geq \rho_r^2$ for $r = \mathrm{rank}(\Sigma_{\mathbf{xy}}) \leq s = \min(p, q)$. The vectors $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ will be unique (apart from signs) if the canonical correlations are distinct. It is well known that the optimization problem given by (9) and (10) is equivalent to solving the eigensystem

$$\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\boldsymbol{\alpha}_k = \rho_k^2\boldsymbol{\alpha}_k, \quad k = 1, \ldots, r \tag{11}$$

$$\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\boldsymbol{\beta}_k = \rho_k^2\boldsymbol{\beta}_k, \quad k = 1, \ldots, r, \tag{12}$$

which makes the search computationally more tractable. Classical estimators are obtained by replacing in (11) and (12) by the sample covariance matrix. A robust counterpart can be easily performed by solving the linear system

$$\begin{aligned} \Sigma_{\mathbf{xx}}^{(R)^{-1}}\Sigma_{\mathbf{xy}}^{(R)}\Sigma_{\mathbf{yy}}^{(R)-1}\Sigma_{\mathbf{yx}}^{(R)}\boldsymbol{\alpha}_k^{(R)} &= \left(\rho_k^{(R)}\right)^2\boldsymbol{\alpha}_k^{(R)}, \quad k = 1, \ldots, r \\ \Sigma_{\mathbf{yy}}^{(R)^{-1}}\Sigma_{\mathbf{yx}}^{(R)}\Sigma_{\mathbf{xx}}^{(R)^{-1}}\Sigma_{\mathbf{xy}}^{(R)}\boldsymbol{\beta}_k^{(R)} &= \left(\rho_k^{(R)}\right)^2\boldsymbol{\beta}_k^{(R)}, \quad k = 1, \ldots, r, \end{aligned} \tag{13}$$

with $\Sigma^{(R)}$ a robust dispersion estimator partitioned as in (4).

CCA can also be seen from a predictive point of view as it was done in (3) for PCA (see Seber [16], p. 260). The canonical variables

$$\mathbf{z} = (\boldsymbol{\alpha}_1^t(\mathbf{x} - E\mathbf{x}), \ldots, \boldsymbol{\alpha}_r^t(\mathbf{x} - E\mathbf{x}))^t \quad \text{and} \quad \mathbf{w} = (\boldsymbol{\beta}_1^t(\mathbf{y} - E\mathbf{y}), \ldots, \boldsymbol{\beta}_r^t(\mathbf{y} - E\mathbf{y}))^t$$

are the best linear combinations to predict each other by making the mean squared loss $E(\|\mathbf{z} - \mathbf{w}\|^2)$ as small as possible since they solve the optimization problem

$$\left(A_C, B_C, \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\mu}_{\mathbf{y}}\right) = \underset{(\bar{A},\bar{B},\boldsymbol{\mu},\boldsymbol{\nu})\in\mathcal{C}}{\arg\min}\; E\left(\left\|\bar{A}\,(\mathbf{x} - \boldsymbol{\mu}) - \bar{B}(\mathbf{y} - \boldsymbol{\nu})\right\|^2\right) \tag{14}$$

with

$$\mathcal{C} = \left\{\left(\bar{A}, \bar{B}, \boldsymbol{\mu}, \boldsymbol{\nu}\right) : \bar{A} \in \mathbb{R}^{r\times p}, \bar{B} \in \mathbb{R}^{r\times q}, \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\nu} \in \mathbb{R}^q, \bar{A}\Sigma_{\mathbf{xx}}\bar{A}^t = I_r = \bar{B}\Sigma_{\mathbf{yy}}\bar{B}^t\right\}, \tag{15}$$

I_r an $r \times r$ identity matrix and $\left(A_C, B_C, \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\mu}_{\mathbf{y}}\right)$ given by the canonical and expected vectors (the subscript C stands for Classical), $A_C = \begin{pmatrix}\boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_r\end{pmatrix}^t$, $B_C = \begin{pmatrix}\boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \cdots & \boldsymbol{\beta}_r\end{pmatrix}^t$, $\boldsymbol{\mu}_{\mathbf{x}} = E\mathbf{x}$ and $\boldsymbol{\mu}_{\mathbf{y}} = E\mathbf{y}$.

The robust proposals for CCA parallel the development of robust procedures for PCA. Romanazzi [14] showed that the classical canonical analysis is sensitive to outlying observations. Karnel [11] considered robust CCA estimators by using M-estimators of multivariate scatter in (13). Because of its low breakdown point for high dimensions, Croux and Dehon [4] used instead the MCD estimator proposed by Rousseeuw [15], which has high breakdown point to estimate the same covariances. Taskinen et al. [17] stated asymptotic properties for CCA based on robust estimators of the covariance matrix. Filzmoser et al. [7] derived a robust method for obtaining the first canonical variables using robust alternating regressions (RAR) following the approach suggested by Wold [19]. Branco et al. [3] extended the method introduced in Filzmoser et al. [7] and they proposed a robust method for obtaining all the canonical variables using RAR.

Branco et al. [3] also dealt with a robust projection pursuit procedure along the lines given by the Eqs. (7)–(10), in which the measure of association given by the correlation $\mathrm{Corr}\left(\mathbf{a}^t\mathbf{x}, \mathbf{b}^t\mathbf{y}\right)$ is replaced by a robust correlation index $I_R\left(\mathbf{a}^t\mathbf{x}, \mathbf{b}^t\mathbf{y}\right)$. I_R can be defined either using a rank correlation measure or taking a 2 × 2 robust dispersion matrix $\Sigma^{(R)}$ based on $(\mathbf{a}^t\mathbf{x}, \mathbf{b}^t\mathbf{y})$, in which $\Sigma^{(R)}$ can be computed using many robust options for multivariate scatter matrix and location available in the statistical literature and the robust correlation index is given by $I_R\left(\mathbf{a}^t\mathbf{x}, \mathbf{b}^t\mathbf{y}\right) = \sigma_{12}/(\sigma_{11}\sigma_{22})$.

Our proposal follows the predictive approach proposed in the context of PCA by Maronna [12], by taking (14) and (15) and measuring the mean squared loss through a robust scale. An efficient algorithm is implemented to compute CCA making use of the connection between PCA and CCA: the PCA algorithm implemented in Maronna [12] is adapted to compute the robust canonical vectors and correlations.

In Section 2 we define a robust predictive approach for CCA by considering a robust scale rather than the mean squared loss in Eq. (14). In Section 3 we establish the Fisher consistency under elliptical distributions and we discuss briefly the concept of breakdown point in this setting. In Section 4 the computing algorithm is established with some convergence properties. Section 5 includes a simulation study to analyze the performance of several proposals for robust CCA. In Section 6 we include some concluding remarks. Proofs are deferred to the Appendix.

2. A robust proposal for CCA

The optimization problems (14) and (15) shed light on the relationship between CCA and PCA. Given the matrices $\bar{A} \in \mathbb{R}^{r\times p}$ and $\bar{B} \in \mathbb{R}^{r\times q}$, let us take $A = \bar{A}\Sigma_{\mathbf{xx}}^{1/2},\ B = \bar{B}\Sigma_{\mathbf{yy}}^{1/2},\ D = \begin{pmatrix} A & -B \end{pmatrix} \in \mathbb{R}^{r\times m},\ m = p+q$ and the random vector $\mathbf{z} = (\mathbf{x}^t\Sigma_{\mathbf{xx}}^{-1/2}, \mathbf{y}^t\Sigma_{\mathbf{yy}}^{-1/2})^t$. By reformulating (14) and (15) for the standardized vectors $\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x}$ and $\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y}$, we have

$$\min_{(\bar{A},\bar{B},\boldsymbol{\mu},\boldsymbol{\nu})\in\mathcal{C}} E\left(\left\|\bar{A}\mathbf{x}-\bar{B}\mathbf{y}-(\bar{A}\boldsymbol{\mu}-\bar{B}\boldsymbol{\nu})\right\|^2\right) = \min_{(D,\mathbf{a})\in\mathcal{B}_{r,m}} E\left(\|D\mathbf{z}-\mathbf{a}\|^2\right) \tag{16}$$

with

$$\mathcal{B}_{r,m} = \left\{(D,\mathbf{a}) : \mathbf{a}\in\mathbb{R}^r, D = \begin{pmatrix} A & -B \end{pmatrix} \in \mathbb{R}^{r\times m},\ AA^t = I_r = BB^t\right\},$$

which is the optimization problem given by (3), except for a missing normalizing constant $\frac{1}{\sqrt{2}}$ in $\begin{pmatrix} A & -B \end{pmatrix}$. Since the covariance matrix for the standardized random vector $\mathbf{z}$ is given by

$$M = \begin{pmatrix} I_p & \Sigma_{\mathbf{xx}}^{-1/2}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2} \\ \Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1/2} & I_q \end{pmatrix}, \tag{17}$$

(3) and (16) give more insight into the relationship between CCA and PCA based on the covariance matrix M (see ten Berge [18]). We next introduce a prediction-based approach to construct robust canonical vectors and correlations, mimicking the ideas given in Maronna [12] for PCA. If we take a look at (16) the nonrobust mean squared loss might be replaced for a robust loss to evaluate the "largeness" of the "residuals" $\left\|A\tilde{\mathbf{x}}-B\tilde{\mathbf{y}}-\mathbf{a}\right\|^2$, with $\Sigma_{\mathbf{xx}}^{(R)}$ and $\Sigma_{\mathbf{yy}}^{(R)}$ robust dispersion estimators for $\Sigma_{\mathbf{xx}}$ and $\Sigma_{\mathbf{yy}}$ respectively, $\tilde{\mathbf{x}} = \left(\Sigma_{\mathbf{xx}}^{(R)}\right)^{-1/2}\mathbf{x}$ and $\tilde{\mathbf{y}} = \left(\Sigma_{\mathbf{yy}}^{(R)}\right)^{-1/2}\mathbf{y}$, and $\left(\begin{pmatrix} A & -B \end{pmatrix}, \mathbf{a}\right) \in \mathcal{B}_{r,m}$. Therefore, to assess the "largeness" of $\left\|A\tilde{\mathbf{x}}-B\tilde{\mathbf{y}}-\mathbf{a}\right\|^2$ we compute an M-scale $\sigma = \sigma(A,B,\mathbf{a})$ implicitly defined through

$$E\rho\left(\frac{\left\|A\tilde{\mathbf{x}}-B\tilde{\mathbf{y}}-\mathbf{a}\right\|^2}{\sigma}\right) = \delta, \tag{18}$$

where $0<\delta<1$ and $\rho : [0,\infty) \to [0,1]$ is a nondecreasing, left-continuous function such that $\rho(0) = 0$ and $\lim_{x\to\infty}\rho(x) = 1$. Then, the robust standardized SM-canonical vectors are defined through the equation

$$(A_o, B_o, \mathbf{a}_o) = \arg\min_{(A,B,\mathbf{a})\in\mathcal{B}_{r,m}} \sigma(A,B,\mathbf{a}), \tag{19}$$

and the final SM-canonical vectors are defined as

$$A_{\text{SM}} = A_o\left(\Sigma_{\mathbf{xx}}^{(R)}\right)^{-1/2}, \qquad B_{\text{SM}} = B_o\left(\Sigma_{\mathbf{yy}}^{(R)}\right)^{-1/2}.$$

The sample version of the estimates is simply obtained by replacing the population expectation by the empirical expectation. The algorithm is easily derived from the fact that we have a constrained minimization and the Lagrange multipliers method applies.

Either robust PCA in [12] or the proposal in (19) are reminiscent of the S-estimators for multivariate scatter and location, in which a scale of the squared Mahalanobis distances

$$d^2(\mathbf{z},\boldsymbol{\mu},\Sigma) = (\mathbf{z}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{z}-\boldsymbol{\mu}) \tag{20}$$

is minimized to yield robust estimates for multivariate dispersion and location. A common procedure is used in the three cases (multivariate scatter, PCA or CCA respectively), in which the smallness of "residuals" is assessed by means of a robust M-scale. Even though the multivariate S-estimators encompass all the information relative to principal and canonical vectors, the two proposals either for PCA or CCA gain remarkably in accuracy as the simulation study reveals.

3. Fisher-consistency of the proposal

In the multivariate location and dispersion model (MLDM) we observe an m-dimensional random vector $\mathbf{z} = (z_1,\ldots,z_m)^t$ with distribution $F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(B) = F_{\mathbf{0}}\left(\boldsymbol{\Sigma}^{-1/2}(B-\boldsymbol{\mu})\right)$, where $F_{\mathbf{0}}$ is a known distribution in $\mathbb{R}^m$, B is a Borel set in $\mathbb{R}^m$, $\boldsymbol{\mu}\in\mathbb{R}^m$ and $\Sigma\in S_m$, the set of $m\times m$ positive definite matrices. An important case is the family of elliptical distributions. We say that an m-dimensional random vector has an elliptical distribution if it has a density of the form

$$f(\mathbf{z},\boldsymbol{\mu}_0,\Sigma_0) = \frac{1}{(\det\Sigma_0)^{1/2}} f_0((\mathbf{z}-\boldsymbol{\mu}_0)^t\Sigma_0^{-1}(\mathbf{z}-\boldsymbol{\mu}_0)), \tag{21}$$

where $f_0 : \mathbb{R}^+ \to \mathbb{R}^+$ is decreasing. If $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\Sigma_0 = I_m$ in (21), then $\mathbf{a}^t\mathbf{z}$ has the same distribution for all $\mathbf{a}\in S^{m-1} = \{\mathbf{a}\in\mathbb{R}^m : \|\mathbf{a}\| = 1\}$. Let us denote $O_{r,m} = \left\{A\in\mathbb{R}^{r\times m} : AA^t = I_r\right\}$. If $\mathbf{w}\in\mathbb{R}^m$, and $\mathcal{D}_m$ stands for the set of distributions of

$\mathbf{w}$ denoted by $\mathcal{D}(\mathbf{w})$, we call a multivariate location and dispersion functional to an application $(\mathbf{T}, S) : \mathcal{D}_m \to \mathbb{R}^m \times \mathbb{R}^{m\times m}$ such that (i) $S(\mathcal{D}(\mathbf{w})) \in S_m$, if m stands for the dimension of the random vector $\mathbf{w}$, (ii) it is affine equivariant, i.e., given a nonsingular matrix $G \in \mathbb{R}^{m\times m}$ and a vector $\mathbf{b} \in \mathbb{R}^m$,

$$\mathbf{T}(\mathcal{D}(G\mathbf{w} + \mathbf{b})) = G\mathbf{T}(\mathcal{D}(\mathbf{w})) + \mathbf{b},$$
$$S(\mathcal{D}(G\mathbf{w} + \mathbf{b})) = GS(\mathcal{D}(\mathbf{w}))G^t.$$

Let us take location and dispersion functionals $(\mathbf{T}, S)$ such that $(\mathbf{T}_\mathbf{x}, S_\mathbf{x}) = (\mathbf{T}(\mathcal{D}(\mathbf{x})), S(\mathcal{D}(\mathbf{x})))$ and $(\mathbf{T}_\mathbf{y}, S_\mathbf{y}) = (\mathbf{T}(\mathcal{D}(\mathbf{y})), S(\mathcal{D}(\mathbf{y})))$ respectively. Take $\bar{\mathbf{x}} = S_\mathbf{x}^{-1/2}(\mathbf{x} - \mathbf{T}_\mathbf{x})$ and $\bar{\mathbf{y}} = S_\mathbf{y}^{-1/2}(\mathbf{y} - \mathbf{T}_\mathbf{y})$. Let Δ_d be the set of $d \times d$ diagonal matrices. Then, a standardized CCA functional $(A_o, B_o, \Lambda_o) : \mathcal{D}_o \to O_{r,p} \times O_{r,q} \times (\Delta_r \cap S_r)$, with $\mathcal{D}_o \subset \mathcal{D}_{p+q}$, is defined to be the solution to an optimization problem, that is,

$$(A_o, B_o) = \arg \min_{D\in O_{r,p}, E\in O_{r,q}} \gamma \left(\mathcal{D} \begin{pmatrix} D\bar{\mathbf{x}} \\ E\bar{\mathbf{y}} \end{pmatrix} \right)$$
$$\Lambda_o = \gamma_0 \left(\mathcal{D} \begin{pmatrix} A_o\bar{\mathbf{x}} \\ B_o\bar{\mathbf{y}} \end{pmatrix} \right),$$

for some functions $\gamma : \tilde{\mathcal{D}} \to \mathbb{R}$ and $\gamma_0 : \tilde{\mathcal{D}} \to \Delta_r \cap S_r$, with $\tilde{\mathcal{D}} \subset \mathcal{D}_{2r}$, and a CCA functional is taken as $\big(A(\mathcal{D}(\mathbf{x}, \mathbf{y})), B(\mathcal{D}(\mathbf{x}, \mathbf{y})), \Lambda(\mathcal{D}(\mathbf{x}, \mathbf{y}))\big) = (A_o S_\mathbf{x}^{-1/2}, B_o S_\mathbf{y}^{-1/2}, \Lambda_o)$. This general concept of CCA functionals includes the functionals associated to the proposals for CCA mentioned in Section 1 as well as the SM-estimation. The functional $(\mathbf{T}, S)$ for the location and dispersion parameters at MLDM is said to be Fisher consistent if $\mathbf{T}(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \boldsymbol{\mu}$ and $S(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) = \Sigma$. If $\mathbf{z}$ is elliptically contoured with finite second moments, then the covariance matrix Σ and Σ_0 are equal up to a constant, that is, $\Sigma = c\Sigma_0$ for some positive constant c. If $\mathbf{z} = (\mathbf{x}^t, \mathbf{y}^t)^t$ is elliptically distributed, then $\mathbf{x}$ and $\mathbf{y}$ are also elliptical, with the location parameter partitioned as $\boldsymbol{\mu}_0 = \left(\boldsymbol{\mu}_\mathbf{x}^t, \boldsymbol{\mu}_\mathbf{y}^t\right)^t$ and the dispersion parameter Σ_0 as in (4). Then, if the target model is elliptical with finite second moments, we say that a CCA functional (A, B, Λ) is Fisher consistent if (A, B, Λ) solves

$$\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}} \left(A^t(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\right) = A^t(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\Lambda(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})$$
$$\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}} \left(B^t(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\right) = B^t(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}})\Lambda(F_{\boldsymbol{\mu},\boldsymbol{\Sigma}}).$$

Then, let us take dispersion functionals $S_\mathbf{x} = S(\mathcal{D}(\mathbf{x}))$ and $S_\mathbf{y} = S(\mathcal{D}(\mathbf{y}))$ which are Fisher consistent for $\Sigma_{\mathbf{xx}}$ and $\Sigma_{\mathbf{yy}}$. Then, we take $\widetilde{\mathbf{x}} = S_\mathbf{x}^{-1/2}\mathbf{x}$ and $\widetilde{\mathbf{y}} = S_\mathbf{y}^{-1/2}\mathbf{y}$ and we look for the solutions σ for (18). Let us take the spectral decomposition for M in (17) given by $M = \sum_{i=1}^{p+q} \gamma_i \mathbf{t}_i \mathbf{t}_i^t$, with $\gamma_1 > \gamma_2 > \cdots > \gamma_r \geq \cdots \geq \gamma_{p+q-r+1} > \cdots > \gamma_{p+q}$, $\mathbf{t}_i^t \mathbf{t}_j = \delta_{ij}$, $1 \leq i, j \leq p + q$, with δ_{ij} the Kronecker delta. Since $\mathbf{t}_i = (\mathbf{v}_i^t, \mathbf{w}_i^t)^t$, $\mathbf{v}_i \in \mathbb{R}^p$, $\mathbf{w}_i \in \mathbb{R}^q$, $i = 1, \ldots, p + q$, let us call

$$A_o = \left(\frac{\mathbf{v}_{p+q-r+1}}{\left\|\mathbf{v}_{p+q-r+1}\right\|}, \ldots, \frac{\mathbf{v}_{p+q}}{\left\|\mathbf{v}_{p+q}\right\|} \right)^t \in \mathbb{R}^{r\times p},$$
$$B_o = \left(\frac{\mathbf{w}_{p+q-r+1}}{\left\|\mathbf{w}_{p+q-r+1}\right\|}, \ldots, \frac{\mathbf{w}_{p+q}}{\left\|\mathbf{w}_{p+q}\right\|} \right)^t \in \mathbb{R}^{r\times q},$$
$$\mathbf{a}_o = A_o \Sigma_{\mathbf{xx}}^{-1/2}\boldsymbol{\mu}_\mathbf{x} - B_o \Sigma_{\mathbf{yy}}^{-1/2}\boldsymbol{\mu}_\mathbf{y}.$$

In the next theorem we state the Fisher consistency of the SM-estimator defined in (19) for elliptical families.

Theorem 1. *Let $\mathbf{z}$ be a random vector with elliptical density given by (21). Then, the SM-estimator defined in (19) is a Fisher consistent estimating functional, that is,*

$$(A_o, B_o, \mathbf{a}_o) = \arg \min_{\mathbf{a}\in\mathbb{R}^r, AA^t = I_r = BB^t} \sigma\left(A, B, \mathbf{a}\right).$$

Corollary 1. *Let $\mathbf{z}$ be a random vector with elliptical distribution F and density given by (21). If ρ is differentiable, with $\rho' = \psi$, with Σ given in (4), then, we have for some constant $c > 0$,*

$$E_F\psi \left(\frac{\left\| A_o\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{x} - B_o\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{y} - \mathbf{a}_o \right\|^2}{\sigma\left(A_o, B_o, \mathbf{a}_o\right)} \right) (\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^t = c\Sigma.$$

One measure to quantify the robustness of an statistical procedure is given by its breakdown point. Roughly speaking, the breakdown point quantifies the minimum amount of contamination for which there are outlying observations which make the estimator move away from the natural parameter space within it is supposed to belong to. Let us formalize this concept for our problem. Given $0 \leq \varepsilon < 0.5$, we assume that the random vector $\mathbf{z} = (\mathbf{x}^t, \mathbf{y}^t)^t \in \mathbb{R}^m$ has a distribution which belongs to an ε-contamination neighborhood, that is,

$$\mathcal{V}_\varepsilon(F_{\mu,\Sigma}) = \left\{F = (1 - \epsilon)F_{\mu,\Sigma} + \epsilon G : G \text{ an arbitrary distribution function in } \mathbb{R}^m\right\}.$$

Given a standardized CCA functional $(A_o(.), B_o(.), \Lambda(.)) : \mathcal{V}_\varepsilon(F_{\mu,\Sigma}) \to O_{r,p} \times O_{r,q} \times (\Delta_r \cap S_r)$ such that $(A_o^0, B_o^0, \Lambda^0) = (A_o(F_{\mu,\Sigma}), B_o(F_{\mu,\Sigma}), \Lambda(F_{\mu,\Sigma}))$ and $\Lambda^0 = \operatorname{diag}(\left(\rho_1^0\right)^2, \dots, \left(\rho_r^0\right)^2)$ under the target model $F_{\mu,\Sigma}$ then, the bias of the jth canonical vector $A_{j,o}$ (respectively $B_{j,o}$) at $F \in \mathcal{V}_\varepsilon(F_{\mu,\Sigma})$, $j = 1, \dots, r$ is defined as $b_{1,j}(\varepsilon, F, A_o) = 1 - \left|A_{j,o}(F)^t A_o^0\right|$ (respectively $b_{2,j}(\varepsilon, F, B_o) = 1 - \left|B_{j,o}(F)^t B_o^0\right|$). The bias for the jth squared canonical correlation ρ_j^2 at $F \in \mathcal{V}_\varepsilon(F_{\mu,\Sigma})$, $j = 1, \dots, r$ is defined as $b_{c,j}(\varepsilon, F, \Lambda_o) = \left|\rho_j^2(F) - \left(\rho_j^0\right)^2\right|$. Then, the maximum asymptotic biases are defined as

$$B_{1,j}(\varepsilon, A_o) = \sup_{F \in \mathcal{V}_\varepsilon(F_{\mu,\Sigma})} b_{1,j}(\varepsilon, F, A_o), \qquad B_{2,j}(\varepsilon, B_o) = \sup_{F \in \mathcal{V}_\varepsilon(F_{\mu,\Sigma})} b_{2,j}(\varepsilon, F, B_o), \tag{22}$$

$$B_{c,j}(\varepsilon, \Lambda_o) = \sup_{F \in \mathcal{V}_\varepsilon(F_{\mu,\Sigma})} b_{c,j}(\varepsilon, F, \Lambda_o), \quad j = 1, \dots, r. \tag{23}$$

The asymptotic breakdown points for these functionals are given by

$$\begin{aligned} &\varepsilon_{1,j}^*(A_o) = \inf\left\{\varepsilon > 0 : B_{1,j}(\varepsilon, A_o) = 1\right\}, \qquad \varepsilon_{2,j}^*(B_o) = \inf\left\{\varepsilon > 0 : B_{2,j}(\varepsilon, B_o) = 1\right\}, \\ &\varepsilon_{1,j}^*(\Lambda_o) = \inf\left\{\varepsilon > 0 : B_{c,j}(\varepsilon, \Lambda) = \max\left(1 - \left(\rho_j^0\right)^2, \left(\rho_j^0\right)^2\right)\right\}, \quad j = 1, \dots, r. \end{aligned} \tag{24}$$

In the case of SM-estimates it is quite straightforward to verify that the M-scale has a breakdown point $\varepsilon^* = \min(\delta, 1 - \delta)$. On the other hand, dealing with the computation of maximum bias and breakdown point as in (22) and (24) seems to be quite intractable. This fact had been also noticed in Maronna [12] for SM-estimation in PCA. Only a few papers can be found in the literature treating this kind of derivation. Zamar [21] treated maximum bias and breakdown point for M-estimators in orthogonal regression. Berrendero [1] and Boente and Orellana [2] dealt with maximum bias and breakdown point of robust projection pursuit estimators for PCA and Common PCA respectively in the case of two-dimensional random vectors.

4. Computing algorithm

The canonical vectors given by Eqs. (14) and (15) have a robust version through (19). Given the sample $\mathbf{z}_i = \left(\mathbf{x}_i^t, \mathbf{y}_i^t\right)^t$, $i = 1, \dots, n$, we firstly take robust initial location estimators for $\mathbf{x}$ and $\mathbf{y}$, $\hat{\boldsymbol{\mu}}_{\mathbf{x}}^{(0)}$ and $\hat{\boldsymbol{\mu}}_{\mathbf{y}}^{(0)}$ respectively, and we get initial robust estimated dispersion matrices $\hat{\Sigma}_{\mathbf{xx}}^{(R)}$ and $\hat{\Sigma}_{\mathbf{yy}}^{(R)}$ for the random vectors $\mathbf{x}$ and $\mathbf{y}$ respectively. Thus, let us take the standardized random vectors $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \left(\left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2} \mathbf{x}, \left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2} \mathbf{y}\right)$. Then, according to (19) we have to solve

$$\min_{(A,B,\mathbf{a}) \in \mathscr{B}_{r,m}} \sigma(A, B, \mathbf{a}),$$

with σ implicitly defined through the equation

$$g(A, B, \mathbf{a}, \sigma) = \frac{1}{n} \sum_{i=1}^n \rho\left(\frac{\left\|A\tilde{\mathbf{x}}_i - B\tilde{\mathbf{y}}_i - \mathbf{a}\right\|^2}{\sigma}\right) = \delta. \tag{25}$$

After a few calculations we get that $\left\|A\tilde{\mathbf{x}} - B\tilde{\mathbf{y}} - \mathbf{a}\right\|^2$ turns out to be

$$\left\|A\tilde{\mathbf{x}} - B\tilde{\mathbf{y}} - \mathbf{a}\right\|^2 = \operatorname{tr}(A\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t A^t) + \operatorname{tr}(B\tilde{\mathbf{y}}\tilde{\mathbf{y}}^t B^t) - 2\operatorname{tr}(B\tilde{\mathbf{y}}\tilde{\mathbf{x}}^t A^t) - 2\operatorname{tr}\left(\mathbf{a}\tilde{\mathbf{x}}^t A^t\right) + 2\operatorname{tr}\left(\mathbf{a}\tilde{\mathbf{y}}^t B^t\right) + \operatorname{tr}\left(\mathbf{a}\mathbf{a}^t\right).$$

Given the multipliers matrices $\Theta = (\theta_{ij}) \in \mathbb{R}^{r\times r}$ and $\Xi = (\xi_{ij}) \in \mathbb{R}^{r\times r}$, and the canonical basis $\{\mathbf{e}_i\}$, $i = 1, \dots, r$ in $\mathbb{R}^r$, the side restrictions of the minimization problem are given by

$$\begin{aligned} &\sum_{i,j} \xi_{ij}\mathbf{e}_i^t(AA^t - I)\mathbf{e}_j = \operatorname{tr}(AA^t\Xi^t) - \operatorname{tr}(\Xi^t), \\ &\sum_{i,j} \theta_{ij}\mathbf{e}_i^t(BB^t - I)\mathbf{e}_j = \operatorname{tr}(BB^t\Theta^t) - \operatorname{tr}(\Theta^t). \end{aligned}$$

We then get the critical points of the augmented function

$$h(A, B, \mathbf{a}, \Theta, \Xi) = \sigma(A, B, \mathbf{a}) + \operatorname{tr}(AA^t\Xi^t) + \operatorname{tr}(BB^t\Theta^t) - \operatorname{tr}(\Theta^t) - \operatorname{tr}(\Xi^t). \tag{26}$$

After differentiating h with respect to $\mathbf{a}$, A, B, Θ and Ξ (see Theorem A.95, p. 386, [13] for differentiation rules of the trace function), if $\psi = \rho'$ (the derivative of ρ), we get that

$$\mathbf{a} = A\tilde{\boldsymbol{\mu}}_{\mathbf{x}} - B\tilde{\boldsymbol{\mu}}_{\mathbf{y}}, \tag{27}$$

with

$$\tilde{\boldsymbol{\mu}}_{\mathbf{x}} = \frac{\sum_{i=1}^{n} \psi_i \tilde{\mathbf{x}}_i}{\sum_{i=1}^{n} \psi_i} \quad \text{and} \quad \tilde{\boldsymbol{\mu}}_{\mathbf{y}} = \frac{\sum_{i=1}^{n} \psi_i \tilde{\mathbf{y}}_i}{\sum_{i=1}^{n} \psi_i},$$

$$\psi_i = \psi\left(\frac{\left\| A\tilde{\mathbf{x}}_i - B\tilde{\mathbf{y}}_i - \mathbf{a}\right\|^2}{\sigma}\right), \quad i = 1, \ldots, n,$$

and the linear system,

$$\begin{aligned} A\tilde{M}_{11} - B\tilde{M}_{21} &= -\sigma\left[t(A,B,\sigma)\right]\left(\frac{\Xi + \Xi^t}{2}\right)A \\ B\tilde{M}_{22} - A\tilde{M}_{12} &= -\sigma\left[t(A,B,\sigma)\right]\left(\frac{\Theta + \Theta^t}{2}\right)B, \end{aligned} \tag{28}$$

with

$$t(A,B,\sigma) = \frac{1}{\sigma^2}\left[\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{\left\| A\tilde{\mathbf{x}}_i - B\tilde{\mathbf{y}}_i - \mathbf{a}\right\|^2}{\sigma}\right)\left\| A\tilde{\mathbf{x}}_i - B\tilde{\mathbf{y}}_i - \mathbf{a}\right\|^2\right],$$

where the matrices $\tilde{M}_{11}, \tilde{M}_{21}, \tilde{M}_{12}$ and $\tilde{M}_{22}$ are defined as

$$\begin{aligned} \tilde{M}_{11} &= \frac{1}{n}\sum_{i=1}^{n}\psi_i\left(\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{x}}\right)\left(\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{x}}\right)^t, &\quad \tilde{M}_{21} &= \frac{1}{n}\sum_{i=1}^{n}\psi_i\left(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{y}}\right)\left(\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{x}}\right)^t, \\ \tilde{M}_{12} &= \frac{1}{n}\sum_{i=1}^{n}\psi_i\left(\tilde{\mathbf{x}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{x}}\right)\left(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{y}}\right)^t, &\quad \tilde{M}_{22} &= \frac{1}{n}\sum_{i=1}^{n}\psi_i\left(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{y}}\right)\left(\tilde{\mathbf{y}}_i - \tilde{\boldsymbol{\mu}}_{\mathbf{y}}\right)^t. \end{aligned} \tag{29}$$

To ease the computation it seems reasonable to make the approximation $\tilde{M}_{11} \approx I_p$ and $\tilde{M}_{22} \approx I_q$ since we started with a standardized data set. Then, the system of equations in (28) turns out to be

$$\begin{aligned} A - B\tilde{M}_{21} &= -\sigma\left[t(A,B,\sigma)\right]\left(\frac{\Xi + \Xi^t}{2}\right)A \\ B - A\tilde{M}_{12} &= -\sigma\left[t(A,B,\sigma)\right]\left(\frac{\Theta + \Theta^t}{2}\right)B, \end{aligned} \tag{30}$$

and, after some algebraic manipulation, we get the eigensystem, with the diagonal matrix $\Lambda \in \mathbb{R}^{r\times r}$,

$$\begin{pmatrix} I_p & \tilde{M}_{12} \\ \tilde{M}_{21} & I_q \end{pmatrix}\begin{pmatrix} A^t \\ -B^t \end{pmatrix} = \begin{pmatrix} A^t \\ -B^t \end{pmatrix}\Lambda. \tag{31}$$

If $(\hat{A}_o, \hat{B}_o, \hat{\Lambda}_o)$ solves the eigensystem (31), we can set

$$\hat{\boldsymbol{\mu}}_{\mathbf{x}} = \frac{\sum_{i=1}^{n} \psi_i \tilde{\mathbf{x}}_i}{\sum_{i=1}^{n} \psi_i} \quad \text{and} \quad \hat{\boldsymbol{\mu}}_{\mathbf{y}} = \frac{\sum_{i=1}^{n} \psi_i \tilde{\mathbf{y}}_i}{\sum_{i=1}^{n} \psi_i},$$

with $\hat{\sigma} = \sigma(\hat{A}_o, \hat{B}_o, \hat{A}_o\hat{\boldsymbol{\mu}}_{\mathbf{x}} - \hat{B}_0\hat{\boldsymbol{\mu}}_{\mathbf{y}})$, and the r first robust estimated canonical vectors and squared canonical correlations are given by

$$\hat{A}_{\text{SM}} = \hat{A}_o\left(\hat{\Sigma}^{(R)}_{\mathbf{xx}}\right)^{-1/2}, \qquad \hat{B}_{\text{SM}} = \hat{B}_o\left(\hat{\Sigma}^{(R)}_{\mathbf{yy}}\right)^{-1/2} \quad \text{and} \quad \hat{\Lambda}^2_{\text{SM}} = \left(\hat{\Lambda}_o - I_r\right)^2. \tag{32}$$

The reasonability for the estimator $\hat{\Lambda}^2_{\text{SM}}$ derives from a fact noticed previously in ten Berge [18] as a byproduct to the equivalence of two oblique congruence rotation methods. More precisely, it is observed that every principal vector $(\mathbf{v}^t, \mathbf{w}^t)^t$ for the matrix M with eigenvalue $\lambda \in (0,2) - \{1\}$ induces canonical vectors $\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{v}$ and $\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{w}$ with squared canonical correlation $(\lambda - 1)^2$. For the sake of clarity, we have included Lemma 3 in the Appendix, to make this relationship between PCA and CCA more evident.

An iterative algorithm can be easily derived, following the numerical proposal given by Maronna [12]. To search for the global minimum of $\sigma(A, B, \mathbf{a})$, the iterative procedure generates a set of candidates from different random initial points and the minimum is chosen within this finite set. Let us take an initial matrix $D^{(0)} = \begin{pmatrix} A_o^{(0)} & -B_o^{(0)} \end{pmatrix} \in \mathbb{R}^{r\times(p+q)}$ such that $A_o^{(0)}\left(A_o^{(0)}\right)^t = I_r = B_o^{(0)}\left(B_o^{(0)}\right)^t$ and parameters $N_1 \in \mathbb{N}$, $N_2 \in \mathbb{N}$, $\delta > 0$ and $tol > 0$. The superscript (s) will stand for the sth step in the iterative procedure. Put $D^{(s)} = \begin{pmatrix} A_o^{(s)} & -B_o^{(s)} \end{pmatrix} \in \mathbb{R}^{r\times(p+q)}$as the updated matrix and the iterative procedure is next introduced.

Step 1

a0. Take preliminary robust covariance estimators $\hat{\Sigma}_{\mathbf{xx}}^{(R)}$ and $\hat{\Sigma}_{\mathbf{yy}}^{(R)}$, set

$$\tilde{\mathbf{z}}_j = (\tilde{\mathbf{x}}_j^t, \tilde{\mathbf{y}}_j^t)^t = \left(\mathbf{x}_j^t\left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2}, \mathbf{y}_j^t\left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2}\right)^t, \quad j = 1, \dots, n.$$

b0. Compute initial location and scale estimates as

$$\begin{aligned}\mathbf{a}^{(0)} &= \operatorname{med}\left(D^{(0)}\tilde{\mathbf{z}}_1, \dots, D^{(0)}\tilde{\mathbf{z}}_n\right)\\ &= \operatorname{med}\left(A^{(0)}\tilde{\mathbf{x}}_1 - B^{(0)}\tilde{\mathbf{y}}_1, \dots, A^{(0)}\tilde{\mathbf{x}}_n - B^{(0)}\tilde{\mathbf{y}}_n\right).\\ \sigma^{(0)} &= \operatorname{MAD}\left(\left\|D^{(0)}\tilde{\mathbf{z}}_1 - \mathbf{a}^{(0)}\right\|^2, \dots, \left\|D^{(0)}\tilde{\mathbf{z}}_n - \mathbf{a}^{(0)}\right\|^2\right).\end{aligned}$$

c0. Put $\Delta = 0$.

Step 2 (local search for the minimum). For $s = 1$ to $N_1 + N_2$, while $\Delta \le tol$,

a1. Compute the residuals $r_i^2\left(D^{(s-1)}, \mathbf{a}^{(s-1)}\right) = \left\|D^{(s-1)}\tilde{\mathbf{z}}_i - \mathbf{a}^{(s-1)}\right\|^2$, $i = 1, \dots, n$.

b1. Compute the M-scale $\sigma^{(s)} = \sigma(A^{(s-1)}, B^{(s-1)}, \mathbf{a}^{(s-1)})$ defined as in (25) and the standardized residuals

$$r_i^{(s)}\left(D^{(s-1)}, \mathbf{a}^{(s-1)}\right) = \left[r_i^2\left(D^{(s-1)}, \mathbf{a}^{(s-1)}\right)\right]/\sigma^{(s)}, \quad i = 1, \dots, n.$$

c1. Compute the weights $\psi_i^{(s)} = \psi\left(r_i^{(s)}\left(D^{(s-1)}, \mathbf{a}^{(s-1)}\right)\right)$, $i = 1, \dots, n$.

d1. Calculate the vectors $\boldsymbol{\mu}_{\mathbf{x}}^{(s)} = \left[\sum_{i=1}^n \psi_i^{(s)}\mathbf{x}_i\right]/\left[\sum_{i=1}^n \psi_i^{(s)}\right]$, $\boldsymbol{\mu}_{\mathbf{y}}^{(s)} = \left[\sum_{i=1}^n \psi_i^{(s)}\mathbf{y}_i\right]/\left[\sim_{i=1}^n \psi_i^{(s)}\right]$, $\boldsymbol{\mu}^{(s)} = \left(\left(\boldsymbol{\mu}_{\mathbf{x}}^{(s)}\right)^t \quad \left(\boldsymbol{\mu}_{\mathbf{y}}^{(s)}\right)^t\right)^t$ and $\tilde{\boldsymbol{\mu}}^{(s)} = \left(\left(\boldsymbol{\mu}_{\mathbf{x}}^{(s)}\right)^t\left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2} \quad \left(\boldsymbol{\mu}_{\mathbf{y}}^{(s)}\right)^t\left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2}\right)^t$. Take $\mathbf{a}^{(s)} = D^{(s-1)}\tilde{\boldsymbol{\mu}}^{(s)}$.

e1. Calculate $\Delta = 1 - \left(\sigma^{(s)}/\sigma^{(s-1)}\right)$.

f1. If $s > N_1$ then

a2. Calculate the matrices

$$M_{21}^{(s)} = \frac{1}{n}\sum_{i=1}^n \psi_i\left(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}}^{(s)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}}^{(s)}\right)^t, \qquad M_{12}^{(s)} = \left(M_{21}^{(s)}\right)^t. \tag{33}$$

Given $\tilde{M}_{12}^{(s)} = \left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2} M_{12}^{(s)}\left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2}$ and $\tilde{M}_{21}^{(s)} = \left(\tilde{M}_{12}^{(s)}\right)^t$, solve the eigenvalues and eigenvectors for the linear system,

$$M^{(s)}\begin{pmatrix}\mathbf{v}_k^{(s)}\\ \mathbf{w}_k^{(s)}\end{pmatrix} = \begin{pmatrix} I_p & \tilde{M}_{12}^{(s)}\\ \tilde{M}_{21}^{(s)} & I_q\end{pmatrix}\begin{pmatrix}\mathbf{v}_k^{(s)}\\ \mathbf{w}_k^{(s)}\end{pmatrix} = \lambda_k^{(s)}\begin{pmatrix}\mathbf{v}_k^{(s)}\\ \mathbf{w}_k^{(s)}\end{pmatrix}, \tag{34}$$

with $\mathbf{v}_k^{(s)} \in \mathbb{R}^p$ and $\mathbf{w}_k^{(s)} \in \mathbb{R}^q$ associated to the kth least eigenvalues $\lambda_k^{(s)}$, with $k = 1, \dots, r$.

b2. Define $D^{(s)} = \begin{pmatrix} A_o^{(s)} & -B_o^{(s)}\end{pmatrix} \in \mathbb{R}^{r\times(p+q)}$ by putting $\left(\mathbf{v}_k^{(s)}\right)^t/\left\|\mathbf{v}_k^{(s)}\right\|$ and $\left(\mathbf{w}_k^{(s)}\right)^t/\left\|\mathbf{w}_k^{(s)}\right\|$ obtained in (34) as the kth row of $A_o^{(s)}$ and $B_o^{(s)}$ respectively, and $\Lambda_o^{(s)} = \operatorname{diag}(\lambda_1^{(s)}, \dots, \lambda_r^{(s)}) \in \mathbb{R}^{r\times r}$.

c2. Set $A^{(s)} = A_o^{(s)}\left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2}$ and $B^{(s)} = B_o^{(s)}\left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2}$.

Step 3 (global search for the minimum)

a3. For N initial matrices $D^{(0)} = \begin{pmatrix} A_o^{(0)} & -B_o^{(0)}\end{pmatrix} \in \mathbb{R}^{r\times(p+q)}$ whose entries are random variables uniformly distributed on $(0, 1)$, a Gram–Schmidt procedure is performed to ensure that $A_o^{(0)}\left(A_o^{(0)}\right)^t = I_r = B_o^{(0)}\left(B_o^{(0)}\right)^t$.

b3. The iterative algorithm given in Step 1 and Step 2 is conducted for each $D^{(0)}$ as input yielding the N outputs

$$\left\{A_{o,h}, B_{o,h}, \Lambda_{o,h}, \mathbf{a}_h\right\}_{h=1}^N$$

and the corresponding scales $\left\{\sigma\left(A_{o,h}, B_{o,h}, \mathbf{a}_h\right)\right\}_{h=1}^N$.

c3. After sorting the N scales we keep the K smallest scales

$$\sigma(A_{o,(h_1)}, B_{o,(h_1)}, \mathbf{a}_{(h_1)}) \leq \cdots \leq \sigma(A_{o,(h_K)}, B_{o,(h_K)}, \mathbf{a}_{(h_K)})$$

and the estimates $A_{o,(h_1)}, B_{o,(h_1)}, \ldots, A_{o,(h_K)}, B_{o,(h_K)}$ associated with them. These K estimates are used as initial estimators to run the iterative procedure again.

d3. We get K candidates,

$$\mathcal{J} = \left\{A^{(\text{out})}_{o(h_j)}, B^{(\text{out})}_{o(h_j)}, \Lambda^{(\text{out})}_{o(h_j)}, \mathbf{a}^{(\text{out})}_{(h_j)}\right\}_{j=1}^{K}$$

and the final output is singled out as

$$(\hat{A}_o, \hat{B}_o, \hat{\mathbf{a}}) = \arg \min_{j\in\{1,\ldots,K\}} \sigma(A^{(\text{out})}_{o(h_j)}, B^{(\text{out})}_{o(h_j)}, \mathbf{a}^{(\text{out})}_{(h_j)}),$$

$$\hat{\Lambda}_o = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_r)$$

with $\hat{\Lambda}_o$ the matrix whose diagonal elements correspond to the eigenvalues associated with $\hat{A}_o$ and $\hat{B}_o$ in increasing order. Then, the SM-estimates for canonical vectors and squared correlations are given by $\hat{A}_{\text{SM}} = \hat{A}_o \left(\hat{\Sigma}^{(R)}_{\mathbf{xx}}\right)^{-1/2}$, $\hat{B}_{\text{SM}} = \hat{B}_o \left(\hat{\Sigma}^{(R)}_{\mathbf{yy}}\right)^{-1/2}$ and

$$\hat{\Lambda}^2_{\text{SM}} = \text{diag}(\widehat{\rho}^2_1, \ldots, \widehat{\rho}^2_r) = \left(\hat{\Lambda}_o - I_r\right)^2. \tag{35}$$

We also include a proposal given in Branco et al. [3] to estimate robustly the canonical correlations based on the SM-estimates for canonical variates $(\hat{\mathbf{v}}^t_k\mathbf{x}, \hat{\mathbf{w}}^t_k\mathbf{y})$, where $\hat{\mathbf{v}}_k$ and $\hat{\mathbf{w}}_k$ are the kth rows of $\hat{A}_{\text{SM}}$ and $\hat{B}_{\text{SM}}$ respectively. In general, given a robust estimator for bivariate dispersion, $\hat{\Sigma}^{(R)}$, a robust correlation RC is defined by replacing the covariance and the standard deviations in the classical sample correlation by their robust counterparts $\hat{\sigma}_{12}, \hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$ respectively, that is, $RC = \widehat{\sigma}_{12}/(\widehat{\sigma}_{11}\widehat{\sigma}_{22})$. Therefore, a robust estimator for the squared canonical correlations is provided by computing $\hat{\Sigma}^{(R)}$ using an MCD or S-estimate based on $\left\{\left(\hat{\mathbf{v}}^t_k\mathbf{x}_i, \hat{\mathbf{w}}^t_k\mathbf{y}_i\right)\right\}_{i=1}^{n}$,

$$\widehat{\rho}^2_k = RC^2\left(\hat{\mathbf{v}}^t_k\mathbf{x}, \hat{\mathbf{w}}^t_k\mathbf{y}\right) = \frac{\hat{\sigma}^2_{12}}{\hat{\sigma}^2_{11}\hat{\sigma}^2_{22}}. \tag{36}$$

Since the ρ function used to define the M-scale is redescending, the iterative algorithm yields only a local minimum of σ, and hence the starting values are essential. That is why the iterative procedure is initialized with a bunch of random candidates for the canonical vectors and then the problem of minimizing σ is replaced by the finite problem of minimizing $\sigma(A, B, \mathbf{a})$ over $\mathcal{J}$. Regarding the convergence properties of the computing algorithm, it can be proved that the scale descends at each iteration of the algorithm described in Step 2 along with a sequence of estimators whose accumulation points are local minimum for the scale function.

Lemma 1. *Let ρ be a nondecreasing, differentiable and concave function in (18). Then σ decreases at each iteration of the algorithm in Step 2.*

Lemma 2. *Assume that ρ is a nondecreasing, twice differentiable and concave function in (18). Then, any accumulation point of the sequence $\left\{\left(D^{(k)}, \mathbf{a}^{(k)}\right)\right\}_{k=1}^{\infty}$ obtained by applying the iterative algorithm is a local minimum of $\sigma(D, \mathbf{a})$.*

5. Simulation study

To assess the performance of different proposals for CCA some measures have been considered. Given $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^q$ identically distributed random vectors, we take $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^t \in \mathbb{R}^{n\times p}$ and $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^t \in \mathbb{R}^{n\times q}$. Let $k \leq \text{rank}(\Sigma_{\mathbf{xy}}) \leq \min(p, q)$, take $\hat{A}_k \in \mathbb{R}^{k\times p}$ and $\hat{B}_k \in \mathbb{R}^{k\times q}$ as the estimators of the k first canonical vectors based on the sample (XY). We assume that $\mathbf{z}_i = (\mathbf{x}^t_i, \mathbf{y}^t_i)^t \sim (1-\varepsilon)F_0 + \varepsilon G_0,\ i = 1, \ldots, n,\ 0 < \varepsilon < 0.5,\ G_0$ a distribution function on $\mathbb{R}^{p+q}$, F_0 the core distribution such that $E_{F_0}\mathbf{z}_1 = \mathbf{0}_{p+q}$ and its covariance matrix is given in (4).

5.1. Relative prediction error

Maronna [12] defines a prediction measure to evaluate the performance of PCA estimators. The concept can be easily adapted to the context of CCA. Let us take another random vector $\mathbf{z} = \left(\mathbf{x}^t, \mathbf{y}^t\right)^t \in \mathbb{R}^{p+q}$ independent of (XY) such that $\mathbf{z} \sim F_0$, $A^0_k \in \mathbb{R}^{k\times p}$ and $B^0_k \in \mathbb{R}^{k\times q}$ are the matrices with the population canonical vectors associated with the largest k squared canonical correlations based on the distribution F_0. Then the Relative Prediction Error (RPE) is defined to be

$$RPE\left(\hat{A}_k, \hat{B}_k\right) = \frac{E\left(\left\|\hat{A}_k\mathbf{x} - \hat{B}_k\mathbf{y}\right\|^2 |X, Y\right)}{E\left(\left\|A_k^0\mathbf{x} - B_k^0\mathbf{y}\right\|^2 |X, Y\right)} - 1 = \frac{E\left(\left\|\hat{A}_k\mathbf{x} - \hat{B}_k\mathbf{y}\right\|^2\right)}{E\left(\left\|A_k^0\mathbf{x} - B_k^0\mathbf{y}\right\|^2\right)} - 1 \tag{37}$$

$$= \frac{\operatorname{tr}\left(\hat{A}_k\Sigma_{\mathbf{xx}}\hat{A}_k^t\right) + \operatorname{tr}\left(\hat{B}_k\Sigma_{\mathbf{yy}}\hat{B}_k^t\right) - 2\operatorname{tr}\left(\hat{A}_k\Sigma_{\mathbf{xy}}\hat{B}_k^t\right)}{\operatorname{tr}\left(A_k^0\Sigma_{\mathbf{xx}}\left(A_k^0\right)^t\right) + \operatorname{tr}\left(B_k^0\Sigma_{\mathbf{yy}}\left(B_k^0\right)^t\right) - 2\operatorname{tr}\left(A_k^0\Sigma_{\mathbf{xy}}\left(B_k^0\right)^t\right)} - 1. \tag{38}$$

Since the true A_k^0 and B_k^0 minimize the mean squared error $E\left(\|A\mathbf{x} - B\mathbf{y}\|^2\right)$, the formula (37) compares the fit yielded by the estimators $\hat{A}_k$ and $\hat{B}_k$ with the fit which gives the smallest expected error. Consequently, $E\left(\left\|\hat{A}_k\mathbf{x} - \hat{B}_k\mathbf{y}\right\|^2\right) \geq E\left(\left\|A_k^0\mathbf{x} - B_k^0\mathbf{y}\right\|^2\right)$ and $RPE\left(\hat{A}_k, \hat{B}_k\right) \geq 0$. The smaller the value of RPE, the better the estimation provided, since the mean squared error is closer to the minimum attainable.

In the simulation study, RPE is calculated for each replicated sample and then a mean prediction error is computed. Therefore, we can take the Mean Relative Prediction Error (MRPE) as

$$\text{MRPE}\left(\hat{A}_k, \hat{B}_k\right) = \frac{1}{n_r}\sum_{j=1}^{n_r} RPE^{(j)}\left(\hat{A}_k^{(j)}, \hat{B}_k^{(j)}\right), \tag{39}$$

where n_r stands for the number of replications in the simulation study, and the superscript (j) denotes the replication we are using to compute the performance measures $RPE^{(j)},\ j = 1, \ldots, n_r$.

5.2. Mean squared error for the estimated canonical vectors

Given $k \in \{1, \ldots, r\}$, let us take $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ the population canonical vectors solving (11) and (12) respectively, with the subscript k referring to the kth largest canonical correlation and $\widehat{\boldsymbol{\alpha}}_k^{(j)}$ and $\widehat{\boldsymbol{\beta}}_k^{(j)}$ estimated vectors corresponding to the jth replication, $j = 1, \ldots, n_r$, based on the sample $\left(X^{(j)} \quad Y^{(j)}\right)$. Branco et al. [3] dealt with the Mean Squared Error (MSE) in the following invariant manner

$$\text{MSE}\left(\widehat{\boldsymbol{\alpha}}_k\right) = \frac{1}{n_r}\sum_{j=1}^{n_r}\cos^{-1}\left(\frac{\left|\boldsymbol{\alpha}_k^t\widehat{\boldsymbol{\alpha}}_k^{(j)}\right|}{\|\boldsymbol{\alpha}_k\|\left\|\widehat{\boldsymbol{\alpha}}_k^{(j)}\right\|}\right), \tag{40}$$

(similarly for $\widehat{\boldsymbol{\beta}}_k$), where $\cos^{-1} : [0, 1] \to [0, \pi/2]$. The measure does not depend on the norm of the canonical vectors, either the population or the estimated ones.

5.3. Mean squared error for the estimated canonical correlations

Given $k \in \{1, \ldots, r\},\ r = \operatorname{rank}(\Sigma_{\mathbf{xy}})$, let ρ_k be the kth population canonical correlation in decreasing order and $\widehat{\rho}_k^{(j)}$ the estimated value based on the sample $\left(X^{(j)} \quad Y^{(j)}\right)$corresponding to the jth replication, $j = 1, \ldots, n_r$. Branco et al. [3] take the Mean Squared Error for canonical correlation as

$$\text{MSE}\left(\widehat{\rho}_k\right) = \frac{1}{n_r}\sum_{j=1}^{n_r}\left(\phi\left(\widehat{\rho}_k^{(j)}\right) - \phi\left(\rho_k\right)\right)^2, \tag{41}$$

where $\phi\,(.) = \tanh^{-1}(.)$ is the Fisher transformation to make the classical canonical correlation estimator asymptotically normal.

To evaluate the performance of several estimators for CCA, we have conducted a simulation study which follows closely the analysis considered in Branco et al. [3] to make a fair comparison. Three possible structures for the covariance between $\mathbf{x}$ and $\mathbf{y}$ are considered,

$$\Sigma_l = \begin{pmatrix} \Sigma_{\mathbf{xx}}^{(l)} & \Sigma_{\mathbf{xy}}^{(l)} \\ \Sigma_{\mathbf{yx}}^{(l)} & \Sigma_{\mathbf{yy}}^{(l)} \end{pmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}, \quad l = 1, 2, 3,$$

with $\Sigma_{\mathbf{xx}}^{(l)} = I_p$, $\Sigma_{\mathbf{yy}}^{(l)} = I_q$, and $\Sigma_{\mathbf{xy}}^{(l)}$ displayed in Table 1. The population canonical vectors and correlations for these matrices are introduced in Table 2.

Table 1
Three covariance structures for the core model in (42).

Configurations		
1 $p = 2\; q = 2$ $\Sigma_{xy}^{(1)} = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.5 \end{pmatrix}$	2 $p = 2\; q = 4$ $\Sigma_{xy}^{(2)} = \begin{pmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \end{pmatrix}$	3 $p = 4\; q = 4$ $\Sigma_{xy}^{(3)} = \begin{pmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}$

Table 2
Canonical correlations and vectors for the population covariance matrices given in Table 1.

Covariance matrix	Canonical vectors associated with x	Canonical vectors associated with y	Canonical correlations
Σ_1	$\alpha_1 = (1, 0)^t$ $\alpha_2 = (0, 1)^t$	$\beta_1 = (1, 0)^t$ $\beta_2 = (0, 1)^t$	$\rho_1 = 0.90$ $\rho_2 = 0.50$
Σ_2	$\alpha_1 = (1, 0)^t$ $\alpha_2 = (0, 1)^t$	$\beta_1 = (1, 0, 0, 0)^t$ $\beta_2 = (0, 1, 0, 0)^t$	$\rho_1 = 0.90$ $\rho_2 = 0.50$
Σ_3	$\alpha_1 = (1, 0, 0, 0)^t$ $\alpha_2 = (0, 1, 0, 0)^t$ $\alpha_3 = (0, 0, 1, 0)^t$ $\alpha_4 = (0, 0, 0, 1)^t$	$\beta_1 = (1, 0, 0, 0)^t$ $\beta_2 = (0, 1, 0, 0)^t$ $\beta_3 = (0, 0, 1, 0)^t$ $\beta_4 = (0, 0, 0, 1)^t$	$\rho_1 = 0.90$ $\rho_2 = 0.50$ $\rho_3 = 1/3$ $\rho_4 = 1/4$

Then, a data set $\left(\mathbf{x}_i^t, \mathbf{y}_i^t\right)^t \in \mathbb{R}^{p+q},\ i = 1, \dots, n$ is generated from a mixture model,

$$(1 - \varepsilon)\, N_{p+q}\left(\mathbf{0}, \Sigma_l\right) + \varepsilon N_{p+q}\left(m\mathbf{1}, \nu^2 \Sigma_l\right), \tag{42}$$

where $\varepsilon \in [0, 0.5]$ stands for the level of contamination, $\mathbf{1} = (1, \dots, 1)^t \in \mathbb{R}^{p+q}$, ν is a "small" positive scalar, and m is a positive scalar which moves along a grid to render asymmetric cases which likely yield the most unfavorable cases for the estimation. More precisely, the level of contamination ε takes the values 0, 0.1 and 0.2, ν is equal to 0.5, and the point mass m moves on the grid $G = \{1, 2, 3, 5, 10, 12, 15, 20\}$. Then, the performance of several estimators for CCA is assessed by using the measures (39)–(41). We describe briefly the candidates included in the list of estimators considered in the analysis. The classical estimator (**Class**) is defined by (11) and (12), replacing the population covariances by their sample counterparts. Three robust multivariate scatter and location estimators were considered as candidates to $\Sigma^{(R)}$ to calculate the eigensystem (13). The M-estimators $\left(\hat{\boldsymbol{\mu}}, \hat{\Sigma}^{(M)}\right)$ are defined as the solutions to the system

$$\begin{aligned} &\sum_{i=1}^{n} W_1(d_i)\left(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}\right) = \mathbf{0} \\ &\frac{1}{n}\sum_{i=1}^{n} W_2(d_i)\left(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}\right)\left(\mathbf{z}_i - \widehat{\boldsymbol{\mu}}\right)^t = \hat{\Sigma}^{(M)}, \end{aligned} \tag{43}$$

where $d_i = d^2\left(\mathbf{z}_i, \widehat{\boldsymbol{\mu}}, \hat{\Sigma}^{(M)}\right)$ are the squared Mahalanobis distances defined in (20) and W_1 and W_2 are weight functions. These weight functions correspond to a Huber's M-estimator, obtained by taking $W_1(d^2) = \min(1, \tau/d)$ with $\tau = \chi^2_{m,0.95}$ and $W_2(d^2) = c\min(1, (\tau/d)^2)$, with c a constant to obtain a consistent estimator of the covariance matrix at normal distributions. The MCD estimator is defined as follows. Given $\mathbf{z}_i \in \mathbb{R}^m,\ i = 1, \dots, n$, random vectors, and $m + 1 \le h < n$, then one takes $d_{(1)} \le d_{(2)} \le \cdots \le d_{(n)}$ the sorted values in increasing order of the squared Mahalanobis distances $d_i = d^2\left(\mathbf{z}_i, \boldsymbol{\mu}, \Sigma\right)$ defined in (20). Hence, a trimming mean is computed by including the $h = [0.75n]$ smallest squared Mahalanobis distances, $\widehat{\sigma}(\boldsymbol{\mu}, \Sigma) = \sum_{i=1}^{h} d_{(i)}$, and the MCD estimator for multivariate location and shape $\left(\widehat{\boldsymbol{\mu}}, \hat{\Sigma}^{(\mathrm{MCD})}\right)$ is computed as

$$\left(\widehat{\boldsymbol{\mu}}, \tilde{\Sigma}^{(\mathrm{MCD})}\right) = \mathop{\arg\min}_{\boldsymbol{\mu}\in\mathbb{R}^m,\, \Sigma\in S_m, \det(\Sigma)=1} \widehat{\sigma}\left(\boldsymbol{\mu}, \Sigma\right). \tag{44}$$

Finally, the estimator for the covariance matrix is given by

$$\hat{\Sigma}^{(\mathrm{MCD})} = \left[\widehat{\sigma}\left(\widehat{\boldsymbol{\mu}}, \tilde{\Sigma}^{(\mathrm{MCD})}\right)\right]^{1/m} \tilde{\Sigma}^{(\mathrm{MCD})}.$$

S-estimators for multivariate location and scatter (Davies [6]) are defined as

$$\left(\widehat{\boldsymbol{\mu}}, \hat{\Sigma}^{(S)}\right) = \arg\min_{\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}\in\mathcal{C}} \det(\Sigma), \tag{45}$$

with

$$\mathcal{C} = \left\{ \boldsymbol{\mu} \in \mathbb{R}^m, \Sigma \in S_m : \frac{1}{n} \sum_{i=1}^{n} \rho\left((\mathbf{z}_i - \boldsymbol{\mu})^t \, \Sigma^{-1} \, (\mathbf{z}_i - \boldsymbol{\mu})^{1/2} \right) = \delta \right\}, \quad \delta \in (0, 1). \tag{46}$$

ρ is taken as the biweight Tukey function $\rho(u) = \left[1 - \left(1 - (u/c)^2\right)^3\right] 1_{[0,c]}(|u|)$ with $c = 1.547$ and $\delta = 0.5$ to obtain an S-estimate with high breakdown point.

M-, MCD or S-estimators can be used in (13) to compute $\Sigma^{(R)}$ (they will be denoted by **M**, **MCD** and **S** respectively in the Tables and Figures). The MCD (or -M) projection pursuit estimators for CCA (denoted as **pp-MCD** and **pp-M**) are defined by using a robust index projection $I_R\left(\mathbf{a}^t\tilde{\mathbf{x}}, \mathbf{b}^t\tilde{\mathbf{y}}\right) = \hat{\sigma}_{12}/(\hat{\sigma}_{11}\hat{\sigma}_{22})$, with the coefficients $\hat{\sigma}_{ij}$ coming from a 2×2 MCD or M-dispersion matrix based on $\left(\mathbf{a}^t\tilde{\mathbf{x}}, \mathbf{b}^t\tilde{\mathbf{y}}\right)$ and standardized observations $\tilde{\mathbf{x}} = \left(\hat{\Sigma}^{(\mathrm{MCD})}_{\mathbf{xx}}\right)^{-1/2} \mathbf{x}$ and $\tilde{\mathbf{y}} = \left(\hat{\Sigma}^{(\mathrm{MCD})}_{\mathbf{yy}}\right)^{-1/2} \mathbf{y}$, with $\hat{\Sigma}^{(\mathrm{MCD})}$ an MCD estimator based on $\mathbf{z} = \left(\mathbf{x}^t, \mathbf{y}^t\right)^t$. The projection pursuit estimators for CCA are defined to be

$$\left(\hat{\boldsymbol{\alpha}}_{1,o}, \hat{\boldsymbol{\beta}}_{1,o}\right) = \arg \max_{\|\mathbf{a}\|=1=\|\mathbf{b}\|} I_R\left(\mathbf{a}^t\tilde{\mathbf{x}}, \mathbf{b}^t\tilde{\mathbf{y}}\right),$$

$$\left(\hat{\boldsymbol{\alpha}}_{l,o}, \hat{\boldsymbol{\beta}}_{l,o}\right) = \arg \max_{\substack{\|\mathbf{a}\|=1=\|\mathbf{b}\|, \\ \mathbf{a} \in \left\langle \hat{\boldsymbol{\alpha}}_{1,o}, \ldots, \hat{\boldsymbol{\alpha}}_{l-1,o} \right\rangle^{\perp}, \mathbf{b} \in \left\langle \hat{\boldsymbol{\beta}}_{1,o}, \ldots, \hat{\boldsymbol{\beta}}_{l-1,o} \right\rangle^{\perp}}} I_R\left(\mathbf{a}^t\tilde{\mathbf{x}}, \mathbf{b}^t\tilde{\mathbf{y}}\right), \quad l > 1,$$

$$\left(\hat{\boldsymbol{\alpha}}_l, \hat{\boldsymbol{\beta}}_l\right) = \left(\left(\hat{\Sigma}^{(\mathrm{MCD})}_{\mathbf{xx}}\right)^{-1/2} \hat{\boldsymbol{\alpha}}_{l,o}, \left(\hat{\Sigma}^{(\mathrm{MCD})}_{\mathbf{yy}}\right)^{-1/2} \hat{\boldsymbol{\beta}}_{l,o} \right), \quad l = 1, \ldots, r.$$

$$\hat{\rho}_l = I_R\left(\hat{\boldsymbol{\alpha}}_l^t\mathbf{x}, \hat{\boldsymbol{\beta}}_l^t\mathbf{y}\right), \quad l = 1, \ldots, r.$$

The Robust Alternating Regression method (**RAR**) is carefully described in Branco et al. [3]. Let us include a very brief description to understand the procedure. The data matrices $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$ are centered by using the columnwise median $\mathbf{m}_{\mathbf{x}} \in \mathbb{R}^p$ and $\mathbf{m}_{\mathbf{y}} \in \mathbb{R}^q$ as location centers, then one takes the residual matrices $X_0 = X - \mathbf{1m}^t_{\mathbf{x}}$ and $Y_0 = Y - \mathbf{1m}^t_{\mathbf{y}}$. Therefore, one must estimate the first principal component $\mathbf{z}^{(0)}_1 \in \mathbb{R}^n$ based on X_0. Next, one must regress $\mathbf{z}^{(0)}_1$ on Y_0 by using weighted L_1-regression to get the regression estimate $\mathbf{b}^{(0)}_1$ and the canonical variates $\mathbf{v}^{(0)}_1 = Y_0\boldsymbol{\beta}^{(0)}_1$, with $\boldsymbol{\beta}^{(0)}_1 = \mathbf{b}^{(0)}_1 / \left\|\mathbf{b}^{(0)}_1\right\|$. Soon after, one regresses $\mathbf{v}^{(0)}_1$ on X_0 to get the regression estimate $\mathbf{a}^{(0)}_1$ and the canonical variates $\mathbf{u}^{(0)}_1 = X_0\boldsymbol{\alpha}^{(0)}_1$ with $\boldsymbol{\alpha}^{(0)}_1 = \mathbf{a}^{(0)}_1 / \left\|\mathbf{a}^{(0)}_1\right\|$, and $\mathbf{u}^{(0)}_1$ on Y_0 to yield $\mathbf{v}^{(1)}_1 = Y_0\boldsymbol{\beta}^{(1)}_1$ with $\boldsymbol{\beta}^{(1)}_1 = \mathbf{b}^{(1)}_1 / \left\|\mathbf{b}^{(1)}_1\right\|$, and so on. One must go back and forth with this procedure until convergence. After running some iterations, call $\mathbf{u}^*_1 = X_0\hat{\boldsymbol{\alpha}}_1$ and $\mathbf{v}^*_1 = Y_0\hat{\boldsymbol{\beta}}_1$ the final estimates for the first canonical variates. To get the upper canonical variates one proceeds recursively: If $l > 1$, and X_{l-2} and Y_{l-2} are the current residual matrices, one constructs the residual matrices $X_{l-1} = X_{l-2} - \mathbf{u}^*_{l-1}\hat{\mathbf{c}}^t$, $Y_{l-1} = Y_{l-2} - \mathbf{v}^*_{l-1}\hat{\mathbf{d}}^t$, with $\hat{\mathbf{c}} \in \mathbb{R}^p$, $\hat{\mathbf{d}} \in \mathbb{R}^q$ obtained by using a robust regression procedure. Then, the former alternating procedure used to obtain the first canonical variates is used to render final estimates for the lth canonical variates $\mathbf{u}^*_l = X_{l-1}\boldsymbol{\alpha}^*_l \in \mathbb{R}^n$ and $\mathbf{v}^*_l = Y_{l-1}\boldsymbol{\beta}^*_l \in \mathbb{R}^n$. Call $\mathbf{u}_1 = \mathbf{u}^*_1$ and $\mathbf{v}_1 = \mathbf{v}^*_1$ the final estimates for the first canonical variates. The final estimates for the lth canonical variates, $l > 1$, are obtained in the following way. Regress the current canonical variates $\mathbf{u}^*_l$ and $\mathbf{v}^*_l$ on the final estimated $l - 1$ canonical variates $\mathbf{u}_1, \ldots, \mathbf{u}_{l-1}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_{l-1}$ respectively by using the robust LTS (Least Trimmed of Squares) regression. Call $U_{l-1} \in \mathbb{R}^{n\times(l-1)}$ and $V_{l-1} \in \mathbb{R}^{n\times(l-1)}$ the matrices whose columns are $\mathbf{u}_1, \ldots, \mathbf{u}_{l-1}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_{l-1}$ respectively. Hence, one gets the regression vectors $\hat{\mathbf{e}} \in \mathbb{R}^{l-1}$ and $\hat{\mathbf{f}} \in \mathbb{R}^{l-1}$ such that the residuals $\mathbf{u}^*_l - U_{l-1}\hat{\mathbf{e}}$ and $\mathbf{v}^*_l - V_{l-1}\hat{\mathbf{f}}$ are regressed on X_0 and Y_0 respectively, rendering robust regression estimates $\hat{\boldsymbol{\alpha}}_l$ and $\hat{\boldsymbol{\beta}}_l$, and the final canonical variates $\hat{\mathbf{u}}_l = X_0\hat{\boldsymbol{\alpha}}_l$ and $\hat{\mathbf{v}}_l = Y_0\hat{\boldsymbol{\beta}}_l$, $l \geq 1$. To perform the simulation study, we used the R-package "rrcov" for the S-estimator and the R-codes available at http://www.statistik.tuwien.ac.at/public/filz/programs.html for the procedures M, MCD, pp-M, pp-MCD and RAR.

To implement the SM-estimator for robust CCA, we have chosen the function

$$\rho(t) = \min(1, 1 - (1 - |t|)^3), \tag{47}$$

considered by Maronna [12] to perform robust estimation in PCA. This function allows for a decreasing sequence of scales as the iterative procedure moves forward as we have seen in Lemma 1. The computing algorithm introduced in Section 4 to calculate canonical vectors and correlations needs to define a bunch of parameters. One of the most important issues to define is the initial robust dispersion matrix to standardize $\mathbf{x}$ and $\mathbf{y}$ respectively at the beginning of the iterative procedure. Maronna [12] proposed S-estimators in the context of PCA. More precisely, he takes a scale σ to evaluate the largeness of the "residuals" $r = \|\mathbf{z} - \boldsymbol{\mu} - P_V(\mathbf{z} - \boldsymbol{\mu})\|^2$, with P_V an orthogonal projection on a subspace $V \subset \mathbb{R}^m$, $\dim(V) = l < m$, and $\boldsymbol{\mu} \in \mathbb{R}^m$. Then, the residuals become $r = r(B, \mathbf{a}) = \|B\mathbf{z} - \mathbf{a}\|^2$ with $B \in \mathbb{R}^{k\times m}$, $k = m - l$ and $BB^t = I_k$. Then, he takes the S-estimators for the principal directions and location as

$$(\hat{B}, \hat{\mathbf{a}}) = \arg \min_{\substack{B:BB^t=I_k \\ \mathbf{a}\in\mathbb{R}^k}} \sigma(B, \mathbf{a}).$$

Table 3
MRPE for the covariance structure Σ_3 under contamination, $k = 1$ and sample size $n = 500$.

ε	m	Class	MCD	M	S	pp-MCD	pp-M	RAR	SM
0.1	1	0.017	0.022	0.017	0.024	1.179	0.020	0.036	0.023
	2	0.063	0.065	0.064	0.079	1.255	0.062	0.094	0.016
	3	0.143	0.016	0.145	0.145	0.666	0.099	0.071	0.015
	5	0.263	0.016	0.265	0.015	1.035	0.474	0.055	0.014
	10	0.353	0.016	0.354	0.015	0.830	0.267	0.035	0.014
	12	0.364	0.016	0.365	0.015	0.733	0.287	0.033	0.014
	15	0.374	0.016	0.374	0.015	0.660	0.241	0.030	0.014
	20	0.381	0.016	0.382	0.015	0.725	0.153	0.028	0.014
0.2	1	0.030	0.044	0.032	0.048	1.126	0.035	0.060	0.046
	2	0.129	0.155	0.132	0.176	1.175	0.132	0.216	0.050
	3	0.227	0.252	0.230	0.268	0.777	0.229	0.207	0.018
	5	0.321	0.023	0.323	0.343	0.912	0.324	0.152	0.018
	10	0.375	0.018	0.375	0.018	0.440	0.385	0.093	0.018
	12	0.381	0.018	0.381	0.018	0.387	0.391	0.080	0.018
	15	0.386	0.018	0.386	0.018	0.260	0.406	0.068	0.018
	20	0.390	0.018	0.390	0.018	0.139	0.399	0.056	0.018

If one considers an M-scale $\sigma(B, \mathbf{a})$ based on $r(B, \mathbf{a})$, that is,

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{\|B\mathbf{z}_i - \mathbf{a}\|^2}{\sigma(B, \mathbf{a})}\right) = \delta, \quad 0 < \delta < 1,$$

then, Maronna [12] shows that at local extrema, the rows $\mathbf{b}_1, \ldots, \mathbf{b}_k$ of B are obtained by applying an iterative scheme to solve the eigensystem

$$\sum_{i=1}^{n} w_i(\mathbf{z}_i - \boldsymbol{\mu})(\mathbf{z}_i - \boldsymbol{\mu})^t\mathbf{b} = \lambda\mathbf{b}, \tag{48}$$

where

$$w_i = \rho'\left(\frac{\|B\mathbf{z}_i - \mathbf{a}\|^2}{\sigma}\right), \qquad \boldsymbol{\mu} = \frac{\sum_{i=1}^{n} w_i\mathbf{z}_i}{\sum_{i=1}^{n} w_i} \quad \text{and} \quad \mathbf{a} = \mathbf{B}\boldsymbol{\mu}.$$

If $\lambda_1, \ldots, \lambda_p, \mathbf{v}_1, \ldots, \mathbf{v}_p$ (respectively $\gamma_1, \ldots, \gamma_q, \mathbf{w}_1, \ldots, \mathbf{w}_q$) solve (48) based on the sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (respectively based on the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$), then we take initial estimators for $\Sigma_{\mathbf{xx}}$ and $\Sigma_{\mathbf{yy}}$ as

$$\hat{\Sigma}_{\mathbf{xx}}^{(\mathrm{SM})} = \sum_{i=1}^{p}\lambda_i\mathbf{v}_i\mathbf{v}_i^t \quad \text{and} \quad \hat{\Sigma}_{\mathbf{yy}}^{(\mathrm{SM})} = \sum_{i=1}^{q}\gamma_i\mathbf{w}_i\mathbf{w}_i^t.$$

We have chosen $Nran$, $kran$, N_1, N_2, δ and tol as 50, 10, 5, 5, 0.5 and 0.01.

SM-1 and **SM-2** will refer to estimated canonical correlations (35) and (36). All the cases considered used $n = 500$ and $n = 50$ as sample sizes, with $n_r = 300$ replications.

We have included only the case of Σ_3, since the conclusions for the other cases are totally similar. Tables 3 and 4 display the MRPE for $k = 1$ in the case of having sample sizes $n = 500$ and $n = 50$. In both cases the tables show a similar behavior, with larger values in the case $n = 50$ as it is expected but the relative behavior amongst the estimates is maintained.

SM shows an outstanding global performance, since it has an efficiency close to Class in the normal case (see Tables 5 and 6), and a remarkable prediction performance in the presence of contamination. When MSE is used to assess the goodness of the estimation, SM also seems to outperform other competitors. MCD also displays a very reasonable performance. The MCD, S and SM show the typical redescending behavior in which the worst performance is produced by contaminations not exceedingly far away from the bulk of the core data set. Rather, monotone estimators like classical and M-, increase their errors as the contamination mean increases. The projection pursuit estimates show a poor performance. RAR seems to provide a weaker performance compared to that of the SM and MCD procedures but better than that of the projection pursuit devices. The simulation study supports the fact that SM-estimation for CCA works more accurately than (13) with multivariate S-estimators.

The computing algorithm for the SM-estimator proceeds with some initial scatter matrices for **x** and **y** respectively to perform the iterative procedure without updating the standardization. We tried with M- or MCD scatter matrices as initial scatter estimators but the results were discouraging and we do not report them. Large and small sample sizes ($n = 500$ and $n = 50$ respectively) were used and the conclusions are quite similar.

Table 4
MRPE for the covariance structure Σ_3 under contamination, $k = 1$ and sample size $n = 50$.

ε	m	Class	MCD	M	S	pp-MCD	pp-M	RAR	SM
0.1	1	0.166	0.398	0.166	0.207	0.956	0.182	0.356	0.191
	2	0.200	0.421	0.203	0.250	0.827	0.206	0.451	0.197
	3	0.278	0.318	0.282	0.271	0.906	0.330	0.391	0.205
	5	0.396	0.303	0.399	0.212	0.833	0.514	0.359	0.205
	10	0.486	0.301	0.488	0.212	0.603	0.428	0.353	0.205
	12	0.497	0.301	0.499	0.212	0.530	0.379	0.352	0.205
	15	0.506	0.301	0.508	0.212	0.499	0.434	0.347	0.205
	20	0.514	0.301	0.516	0.212	0.519	0.533	0.361	0.205
0.2	1	0.195	0.480	0.194	0.255	0.948	0.230	0.447	0.221
	2	0.298	0.603	0.300	0.387	1.269	0.330	0.664	0.228
	3	0.403	0.685	0.404	0.484	1.212	0.424	0.619	0.227
	5	0.501	0.541	0.501	0.562	1.284	0.519	0.523	0.227
	10	0.557	0.244	0.557	0.253	0.656	0.587	0.522	0.227
	12	0.563	0.244	0.563	0.235	0.580	0.592	0.536	0.227
	15	0.569	0.244	0.568	0.235	0.541	0.592	0.529	0.227
	20	0.573	0.244	0.573	0.235	0.516	0.600	0.452	0.227

Table 5
Efficiency for the canonical vectors $\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_4$ associated with **x** and $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_4$ associated with **y** corresponding to Σ_3 in the case of normal samples ($\varepsilon = 0$).

Sample size		Class	MCD	M	S	pp-MCD	pp-M	RAR	SM
500	MRPE	0.014	0.017	0.014	0.016	0.821	0.015	0.023	0.016
	MSE($\hat{\alpha}_1$)	0.040	0.044	0.041	0.043	0.221	0.042	0.054	0.043
	MSE($\hat{\alpha}_2$)	0.184	0.206	0.185	0.201	0.453	0.196	0.283	0.200
	MSE($\hat{\alpha}_3$)	0.397	0.442	0.404	0.434	0.543	0.406	0.540	0.432
	MSE($\hat{\alpha}_4$)	0.369	0.415	0.377	0.403	0.484	0.374	0.522	0.401
	MSE($\hat{\beta}_1$)	0.039	0.044	0.039	0.043	0.224	0.041	0.053	0.042
	MSE($\hat{\beta}_2$)	0.188	0.209	0.190	0.205	0.451	0.200	0.278	0.203
	MSE($\hat{\beta}_3$)	0.399	0.442	0.407	0.433	0.529	0.404	0.543	0.429
	MSE($\hat{\beta}_4$)	0.373	0.411	0.380	0.402	0.470	0.375	0.504	0.398
50	MRPE	0.163	0.369	0.165	0.209	0.735	0.178	0.322	0.184
	MSE($\hat{\alpha}_1$)	0.135	0.205	0.136	0.154	0.247	0.141	0.195	0.144
	MSE($\hat{\alpha}_2$)	0.636	0.780	0.641	0.676	0.626	0.649	0.807	0.664
	MSE($\hat{\alpha}_3$)	0.919	0.964	0.921	0.931	0.797	0.899	0.989	0.918
	MSE($\hat{\alpha}_4$)	0.789	0.883	0.790	0.821	0.759	0.780	0.902	0.797
	MSE($\hat{\beta}_1$)	0.139	0.203	0.139	0.154	0.261	0.144	0.199	0.146
	MSE($\hat{\beta}_2$)	0.629	0.810	0.637	0.661	0.620	0.641	0.803	0.655
	MSE($\hat{\beta}_3$)	0.916	0.973	0.917	0.929	0.763	0.894	1.015	0.926
	MSE($\hat{\beta}_4$)	0.794	0.878	0.800	0.832	0.749	0.789	0.908	0.821

Table 6
Efficiency for the canonical correlations $\widehat{\boldsymbol{\rho}}_1, \ldots, \widehat{\boldsymbol{\rho}}_4$ corresponding to Σ_3 in the case of normal samples ($\varepsilon = 0$).

Sample size		Class	MCD	M	S	pp-MCD	pp-M	RAR	SM-1	SM-2
500	MSE($\hat{\rho}_1$)	0.002	0.003	0.002	0.003	0.167	0.003	0.003	0.016	0.003
	MSE($\hat{\rho}_2$)	0.002	0.003	0.002	0.002	0.044	0.003	0.004	0.005	0.003
	MSE($\hat{\rho}_3$)	0.002	0.002	0.002	0.002	0.022	0.003	0.003	0.003	0.003
	MSE($\hat{\rho}_4$)	0.002	0.002	0.002	0.002	0.008	0.003	0.004	0.004	0.003
50	MSE($\hat{\rho}_1$)	0.026	0.069	0.026	0.035	0.127	0.054	0.057	0.023	0.044
	MSE($\hat{\rho}_2$)	0.034	0.083	0.034	0.045	0.114	0.071	0.067	0.018	0.061
	MSE($\hat{\rho}_3$)	0.016	0.028	0.016	0.019	0.085	0.037	0.043	0.012	0.032
	MSE($\hat{\rho}_4$)	0.019	0.021	0.019	0.020	0.032	0.023	0.034	0.021	0.024

10%-upper trimmed MSE was also recorded but it was not included in the paper since it allows for conclusions which are totally equivalent to our study with MSE. Figs. 1 and 2 depict the behavior of the MSE for the four canonical vectors associated with the vector **x** when the sample sizes are $n = 500$ and $n = 50$ respectively. The plots for the MSE corresponding to the estimated canonical vectors associated with the vector **y** were omitted since the visual inspection does not provide additional discernment and a similar performance to that of Figs. 1 and 2 is observed, showing that SM outperforms the other competitors in most of the cases, while pp-MCD improves its behavior in Fig. 2(c) and (d). We have not included the case with M-estimation since it was barely different from that of the classical correlation.

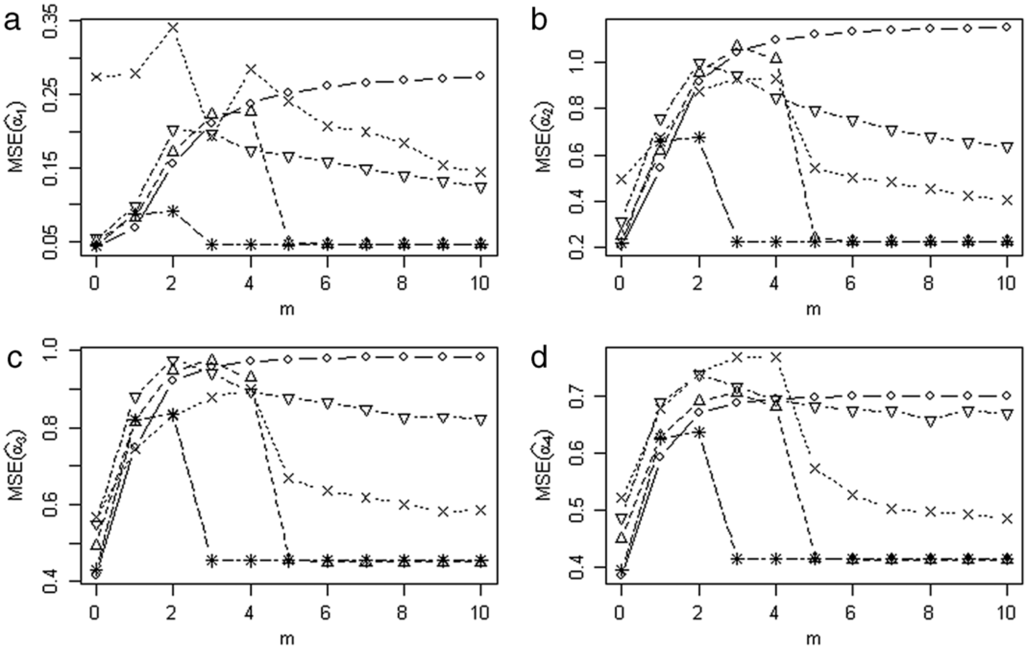

Fig. 1. MSE for the estimated canonical vectors associated with vector **x**, corresponding to the matrix Σ_3 in the case of Classical (○), MCD (△), pp-MCD (×), RAR (▽) and SM (∗) estimators. The underlying distribution is given by formula (42) with the mean m moving from 1 to 10 and the contamination level ε is equal to 0.2. The sample size is $n = 500$. M-, S-, and pp-M estimators were omitted because of their poor behavior.

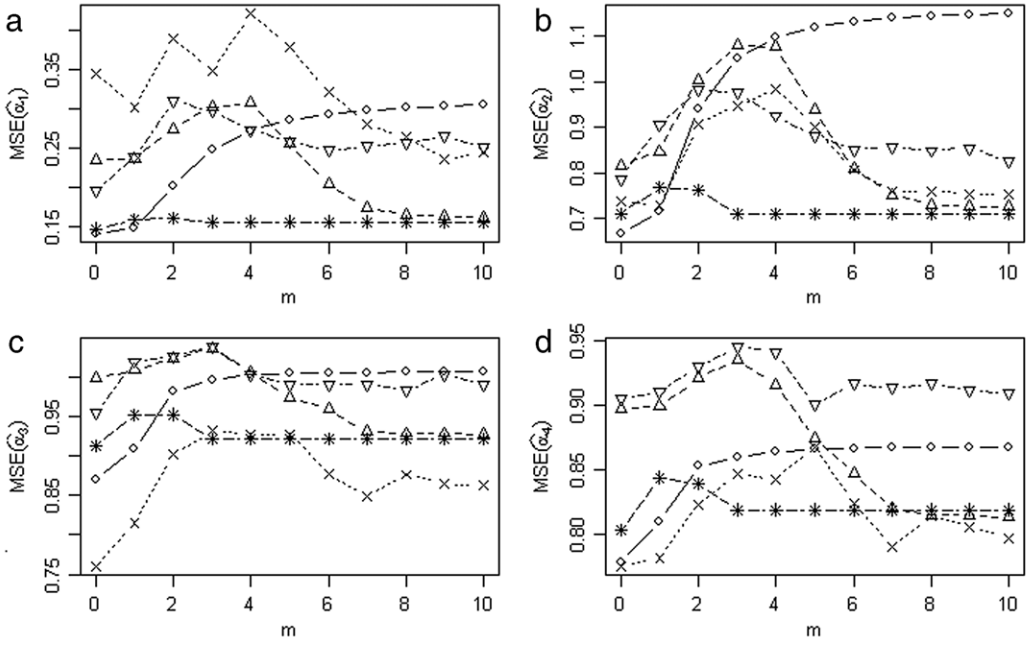

Fig. 2. MSE for the estimated canonical vectors associated with vector **x**, corresponding to the matrix Σ_3 in the case of Classical (○), MCD (△), pp-MCD (×), RAR (▽) and SM (∗) estimators. The underlying distribution is given by formula (42) with the mean m moving from 1 to 10 and the contamination level ε is equal to 0.2. The sample size is $n = 50$. M-, S-, and pp-M estimators were omitted because of their poor behavior.

Figs. 3 and 4 depict the estimated canonical correlations related to the different proposals when the sample sizes are $n = 500$ and $n = 50$ respectively. The measures (35) and (36) for squared canonical correlation are included denoted by SM-1 and SM-2 respectively. SM-1 has a remarkable performance while SM-2 worsens its behavior as the correlation order increases. The redescending behavior displayed for MCD and SM is also noticeable in the MSE analysis while the MSE for the monotone estimators worsens as the contamination mean increases. The profile for each estimator allows for a clearer comparison amongst them showing the outstanding performance of SM-estimators.

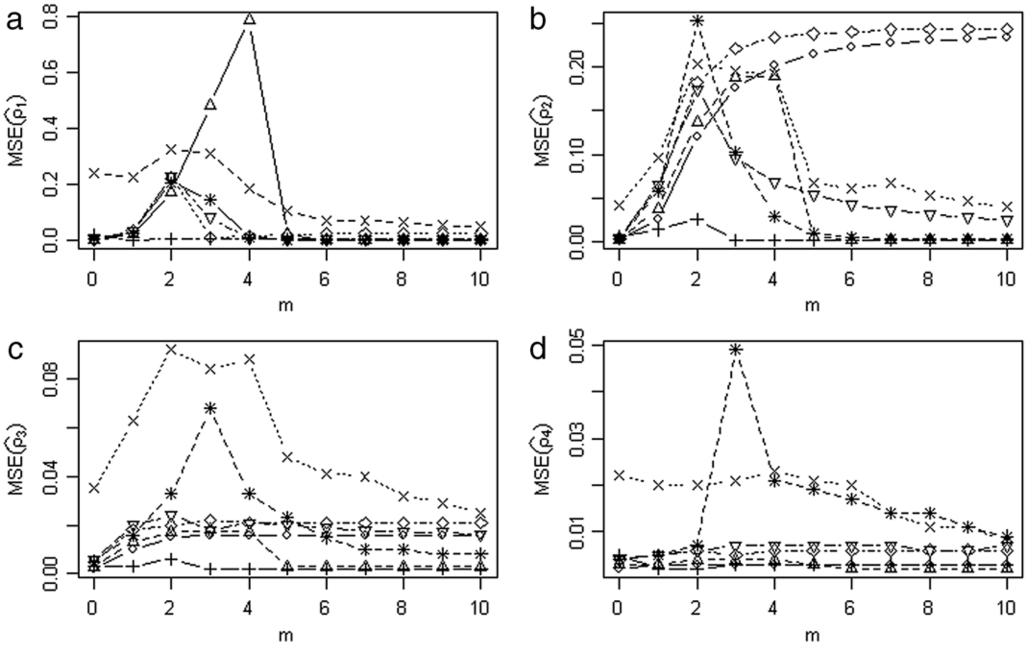

Fig. 3. MSE for the estimated canonical correlations corresponding to the matrix Σ_3 in the case of Classical (○), MCD (△), pp-MCD (×), pp-M (◊), RAR (▽), SM-1 (+), SM-2 (∗) estimators. The underlying distribution is given by formula (42), with the mean m moving from 1 to 10 and the contamination level ε is equal to 0.2. The sample size is $n = 500$. M- and S- estimators were omitted because of their poor behavior. The classical estimator was not included for the first canonical correlation since its MSE greatly exceeds to that of the other estimates.

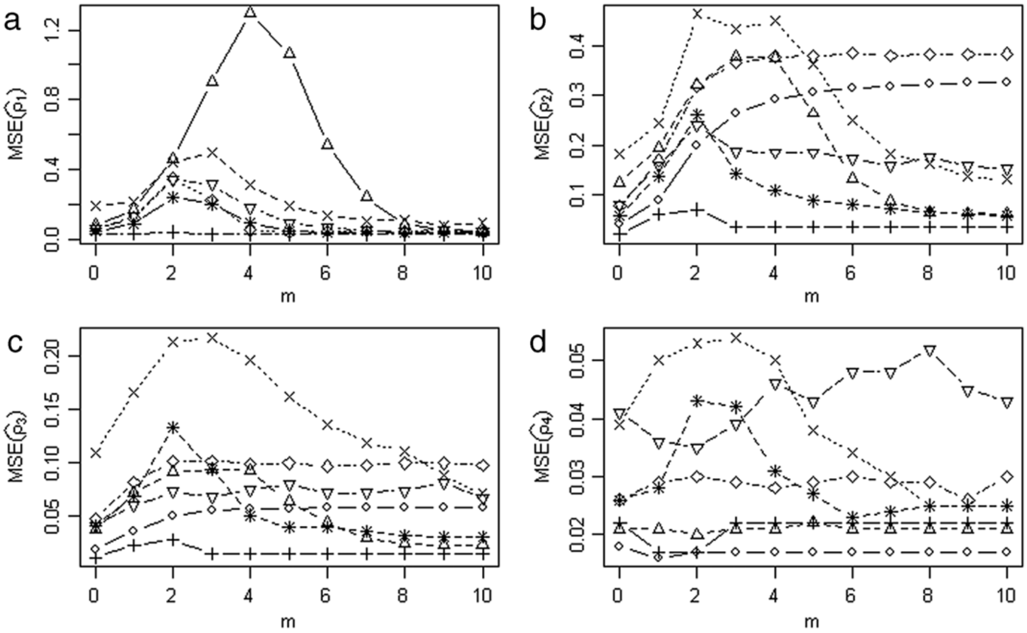

Fig. 4. MSE for the estimated canonical correlations corresponding to the matrix Σ_3 in the case of Classical (○), MCD (△), pp-MCD (×), pp-M (◊), RAR (▽), SM-1 (+), SM-2 (∗) estimators. The underlying distribution is given by formula (42), with the mean m moving from 1 to 10 and the contamination level ε is equal to 0.2. The sample size is $n = 50$. M- and S-estimators were omitted because of their poor behavior. The classical estimator was not included for the first canonical correlation since its MSE greatly exceeds to that of the other estimates.

6. Concluding remarks

If we take into account the predictive aspect of CCA and we consider a robust M-scale to evaluate the squared difference between two lower dimensional linear combinations of the original vectors **x** and **y**, we come up with SM-estimators for canonical vectors and correlations. The new proposal turns out to be Fisher-consistent. We also propose a computing algorithm following a similar approach dealt with by Maronna [12]. By taking a sufficiently large number of different starting points as it is described in Step 3 of the algorithm, one generates a number of candidates which are close enough to the global minimum. The simulation study seems to support that the global optimum is well approximated. As to computation time, the SM-procedure looks rather slow compared to the other competitors, in spite of its excellent statistical performance. However, any issue on running time will become easily out of fashion due to the persistent improvement in processors technology. The performance of the new proposal and some other robust estimators for CCA was analyzed through a simulation study, by using a predictive measure and the mean squared error, to evaluate the efficiency under the normal multivariate distribution and the robustness under contamination. In both cases, the SM-estimator for CCA behaves remarkably well compared with other candidates. Therefore, we can conclude that the S-estimation, tailored in the PCA and CCA contexts, seems to be appropriate as robust methodology.

Acknowledgments

We would like to thank two anonymous referees and the Associate Editor for their comments and suggestions that have resulted in a much improved paper. First author's research was partially supported by Grants PICT 0821 and 0397 from ANPCYT, Grant 05/B424 Secyt, UNC, and Grant 20020100100276 from Secyt, UBA, Argentina. Second author's research was partially supported by Grants PICT 0397 from ANPCYT and Grant 05/B424 Secyt, UNC, Argentina.

Appendix

A technical lemma before the main result about consistency is needed.

Proposition A.1. *If $C \in \mathbb{R}^{r\times m},\ r \leq m$, is a matrix with orthonormal rows, then C^tC is an orthogonal projection onto the rows of C.*

Proof. If $\mathbf{c}^t$ is a row of C, then $C^tC\mathbf{c} = \mathbf{c}$. If $\mathbf{c}$ is orthogonal with respect to the rows of C, then $C^tC\mathbf{c} = 0_{m\times m}$. Therefore the proposition holds.

Proof of Theorem 1. Let us consider

$$\|A\widetilde{\mathbf{x}} - B\widetilde{\mathbf{y}} - \mathbf{a}\|^2 = 2\left(\mathbf{z} - \mathbf{a}_0\right)^t \frac{C^t}{\sqrt{2}}\frac{C}{\sqrt{2}}\left(\mathbf{z} - \mathbf{a}_0\right),$$

with

$$C = \begin{pmatrix} A_{r\times p} & -B_{r\times q}\end{pmatrix}, \qquad C^tC = \begin{pmatrix} A^tA & -A^tB \\ -B^tA & B^tB \end{pmatrix}, \qquad C\mathbf{a}_0 = \mathbf{a}.$$

Recall that $\Sigma_0 = c\Sigma,\ \Sigma$ given in (4) for some constant $c > 0$. Then $(\widetilde{\mathbf{x}}^t, \widetilde{\mathbf{y}}^t)^t$ is elliptically contoured with density $\widetilde{f}(\mathbf{z}) = \det(M_0^{-1/2}) f_0\left(\mathbf{z}^t M_0^{-1}\mathbf{z}\right)$, with $M_0 = cM,\ M$ given by (17). Then,

$$\begin{aligned}
\delta &= E\rho\left(\frac{\|A\widetilde{\mathbf{x}} - B\widetilde{\mathbf{y}} - \mathbf{a}\|^2}{\sigma}\right) \\
&= E\rho\left(\frac{\left(\mathbf{z} - \mathbf{a}_0\right)^t C^tC\left(\mathbf{z} - \mathbf{a}_0\right)}{\sigma}\right) \\
&= \det(M_0^{-1/2})\int \rho\left(\frac{\left(\mathbf{z} - \mathbf{a}_0\right)^t C^tC\left(\mathbf{z} - \mathbf{a}_0\right)}{\sigma}\right) f_0\left(\mathbf{z}^t M_0^{-1}\mathbf{z}\right) d\mathbf{z}.
\end{aligned}$$

If we make a change of variables $\mathbf{u} = M_0^{-1/2}\mathbf{z}$ and call $\mathbf{b}_0 = M_0^{-1/2}\mathbf{a}_0$ we get that

$$\delta = \int \rho\left(\frac{\left(\mathbf{u} - \mathbf{b}_0\right)^t M_0^{1/2}\left(C^tC\right)M_0^{1/2}\left(\mathbf{u} - \mathbf{b}_0\right)}{\sigma}\right) f_0\left(\mathbf{u}^t\mathbf{u}\right) d\mathbf{u}.$$

Since f_0 is decreasing, if we proceed by successive steps, starting with an indicator function as a ρ function, afterwards, taking simple functions and the Monotone Convergence Theorem in the end, we can finally assure that,

$$\int \rho\left(\frac{\left(\mathbf{u} - \mathbf{b}_0\right)^t M_0^{1/2}\left(C^tC\right)M_0^{1/2}\left(\mathbf{u} - \mathbf{b}_0\right)}{\sigma}\right) f_0\left(\mathbf{u}^t\mathbf{u}\right) d\mathbf{u} \geq \int \rho\left(\frac{\mathbf{u}^t M_0^{1/2}\left(C^tC\right)M_0^{1/2}\mathbf{u}}{\sigma}\right) f_0\left(\mathbf{u}^t\mathbf{u}\right) d\mathbf{u}.$$

If $M_0 = P\Gamma P^t$ with Γ a diagonal matrix with the elements in decreasing order and P an orthogonal matrix, then

$$\begin{aligned}\int \rho\left(\frac{\mathbf{u}^t M_0^{1/2}\left(C^tC\right)M_0^{1/2}\mathbf{u}}{\sigma}\right) f_0\left(\mathbf{u}^t\mathbf{u}\right) d\mathbf{u} &= \int \rho\left(\frac{\mathbf{u}^t P\Gamma^{1/2}P^t\left(C^tC\right)P\Gamma^{1/2}\left(P^t\mathbf{u}\right)}{\sigma}\right) f_0\left(\mathbf{u}^t\mathbf{u}\right) d\mathbf{u} \\ &= \int \rho\left(\frac{(P^t\mathbf{u})^t\Gamma^{1/2}(CP)^t(CP)\Gamma^{1/2}\left(P^t\mathbf{u}\right)}{\sigma}\right) f_0\left(\mathbf{u}^t\mathbf{u}\right) d\mathbf{u} \\ &= \int \rho\left(\frac{\mathbf{z}^t\Gamma^{1/2}(CP)^t(CP)\Gamma^{1/2}\mathbf{z}}{\sigma}\right) f_0\left(\mathbf{z}^t\mathbf{z}\right) d\mathbf{z}. \end{aligned} \tag{49}$$

CP is another matrix with orthonormal rows since $\left(\mathbf{c}_jP\right)\left(\mathbf{c}_iP\right)^t = \delta_{ij}$, $1 \le i, j \le r$. Therefore, $(CP)^t(CP)$ is a projection matrix by Proposition A.1. Then, $P_V = P^tC^tCP$, with V an r dimensional subspace in $\mathbb{R}^m$. Hence

$$\int \rho\left(\frac{\mathbf{z}^t\Gamma^{1/2}(CP)^t(CP)\Gamma^{1/2}\mathbf{z}}{\sigma}\right) f_0\left(\mathbf{z}^t\mathbf{z}\right) d\mathbf{z} = \int \rho\left(\frac{\mathbf{z}^t\Gamma^{1/2}P_V\Gamma^{1/2}\mathbf{z}}{\sigma}\right) f_0\left(\mathbf{z}^t\mathbf{z}\right) d\mathbf{z}.$$

Let us call $R = \Gamma^{1/2}P_V\Gamma^{1/2}$. R is a symmetric nonnegative definite matrix and its spectral decomposition is given by $R = V\Lambda V^t$, $VV^t = V^tV = I_m$, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$. Since the $\dim\operatorname{Ker} R = m - r$, we can suppose without loss of generality that $\lambda_{r+1} = \cdots = \lambda_m = 0$. Then, taking $\mathbf{w} = V^t\mathbf{z}$ we get,

$$\int \rho\left(\frac{\mathbf{z}^t\Gamma^{1/2}P_V\Gamma^{1/2}\mathbf{z}}{\sigma}\right) f_0\left(\mathbf{z}^t\mathbf{z}\right) d\mathbf{z} = \int \rho\left(\frac{\mathbf{w}^t\Lambda\mathbf{w}}{\sigma}\right) f_0\left(\mathbf{w}^t\mathbf{w}\right) d\mathbf{w},$$

since $|\det(V)| = 1$. Put $D = \Gamma^{1/2}$ and take $\mathbf{v} \ne \mathbf{0}$ an eigenvector associated with the eigenvalue $\lambda \ne 0$, then

$$\begin{aligned}&DP_VD\mathbf{v} = \lambda\mathbf{v}, \\ &P_VD\mathbf{v} = \lambda D^{-1}\mathbf{v}.\end{aligned}$$

This means that $D^{-1}\mathbf{v} \in V$. Then $D^{-1}\mathbf{v}$ is orthogonal to $D\mathbf{v} - \lambda D^{-1}\mathbf{v}$, that is,

$$\begin{aligned}&0 = \mathbf{v}^tD^{-1}(D\mathbf{v} - \lambda D^{-1}\mathbf{v}) = \mathbf{v}^t\left(I - \lambda D^{-2}\right)\mathbf{v} \\ &\lambda = \frac{1}{\mathbf{v}^tD^{-2}\mathbf{v}} = \frac{1}{\sum\limits_{i=1}^m d_i^{-2}v_i^2} \\ &\frac{1}{d_m^{-2}\sum\limits_{i=1}^m v_i^2} = d_m^2 = \gamma_m \le \lambda \le \frac{1}{d_1^{-2}\sum\limits_{i=1}^m v_i^2} = d_1^2 = \gamma_1.\end{aligned}$$

If $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ denotes the set of eigenvectors related to the nonnull eigenvalues of R, then $\Lambda = \operatorname{diag}((\mathbf{v}_1^tD^{-2}\mathbf{v}_1)^{-1}, \ldots, \left(\mathbf{v}_r^tD^{-2}\mathbf{v}_r\right)^{-1}, 0, \ldots, 0)$ with $V = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \in \mathbb{R}^{(p+q)\times r}$, $V^tV = I_r$. Let us consider the function

$$f_2(V) = \mathbf{z}^t\Lambda\mathbf{z} = \sum_{k=1}^r \left(\mathbf{e}_k^tV^tD^{-2}V\mathbf{e}_k\right)^{-1} z_k^2,$$

with $\{\mathbf{e}_1, \ldots, \mathbf{e}_r\}$ the canonical basis in $\mathbb{R}^r$. Observe that

$$\frac{\partial\left(\mathbf{e}_k^tV^tD^{-2}V\mathbf{e}_k\right)}{\partial V} = 2D^{-2}V\mathbf{e}_k\mathbf{e}_k^t.$$

Since we are looking for restricted minimum, we consider the Lagrangian function for $f_2(V)$ given by

$$\begin{aligned}h_2(V) &= \mathbf{z}^t\Lambda\mathbf{z} + \sum_{1\le j,\ k\le r} \lambda_{kj}\left(\mathbf{e}_j^tV^tV\mathbf{e}_k - \delta_{kj}\right) \\ &= \sum_{k=1}^r \left(\mathbf{e}_k^tV^tD^{-2}V\mathbf{e}_k\right)^{-1} z_k^2 + \sum_{1\le j,\ k\le r} \lambda_{kj}\left(\mathbf{e}_j^tV^tV\mathbf{e}_k - \delta_{kj}\right). \\ \frac{\partial h_2}{\partial V} &= -2\sum_{k=1}^r \left(\mathbf{e}_k^tV^tD^{-2}V\mathbf{e}_k\right)^{-2} D^{-2}V\mathbf{e}_k\mathbf{e}_k^tz_k^2 + 2V\sum_{1\le j,\ k\le r}^{r} \lambda_{kj}\mathbf{e}_k\mathbf{e}_j^t = 0_{m\times r}.\end{aligned}$$

If we premultiply $\frac{\partial h_2}{\partial V}$ by $(V\mathbf{e}_l)^t$ and postmultiply by $\mathbf{e}_l$, $l \in \{1,\ldots,r\}$, we get

$$0 = -\left(\mathbf{e}_l^t V^t D^{-2} V\mathbf{e}_l\right)^{-2} \mathbf{e}_l^t V^t D^{-2} V\mathbf{e}_l z_l^2 + \lambda_{ll},$$
$$\lambda_{ll} = \left(\mathbf{e}_l^t V^t D^{-2} V\mathbf{e}_l\right)^{-1} z_l^2.$$

If we premultiply $\frac{\partial h_2}{\partial V}$ by $(V\mathbf{e}_k)^t$ and postmultiply by $\mathbf{e}_j$, $1 \le k \ne j \le r$, we obtain

$$\lambda_{kj} = \left(\mathbf{e}_j^t V^t D^{-2} V\mathbf{e}_j\right)^{-2} \mathbf{e}_k^t V^t D^{-2} V\mathbf{e}_j z_j^2, \quad k \ne j.$$

Therefore, the partial derivative of h_2 with respect to V is

$$\frac{\partial h_2}{\partial V} = -\sum_{k=1}^{r} \frac{D^{-2}\mathbf{v}_k}{\left(\mathbf{v}_k^t D^{-2}\mathbf{v}_k\right)^2}\mathbf{e}_k^t z_k^2 + \sum_{l=1}^{r} \frac{\mathbf{v}_l}{\mathbf{v}_l^t D^{-2}\mathbf{v}_l}\mathbf{e}_l^t z_l^2 + \sum_{1\le j\ne k\le r} \frac{\mathbf{v}_k^t D^{-2}\mathbf{v}_j}{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)^2}\mathbf{v}_k z_j^2 \mathbf{e}_j^t$$
$$0_{m\times r} = \sum_{j=1}^{r}\left[\frac{D^{-2}\mathbf{v}_j z_j^2}{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)^2} - \frac{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)\mathbf{v}_j z_j^2}{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)^2} - \sum_{k\ne j}\frac{\mathbf{v}_k^t D^{-2}\mathbf{v}_j z_j^2}{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)^2}\mathbf{v}_k\right]\mathbf{e}_j^t$$
$$\mathbf{0}_m = \frac{D^{-2}\mathbf{v}_j z_j^2}{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)^2} - \frac{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)\mathbf{v}_j z_j^2}{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)^2} - \sum_{k\ne j}\frac{\mathbf{v}_k^t D^{-2}\mathbf{v}_j z_j^2}{\left(\mathbf{v}_j^t D^{-2}\mathbf{v}_j\right)^2}\mathbf{v}_k, \quad \text{for all } j = 1,\ldots,r$$
$$D^{-2}\mathbf{v}_j = \sum_{k=1}^{r}\left(\mathbf{v}_k^t D^{-2}\mathbf{v}_j\right)\mathbf{v}_k.$$

Then, we can conclude that the subspace V generated by $\{\mathbf{v}_1,\ldots,\mathbf{v}_r\}$ is D^{-2}-invariant and therefore, by using a well known result from linear algebra (see Hoffman and Kunze [9, p. 263]), V must be a direct sum of the subspaces $V \cap V_i$, where V_i, $i = 1,\ldots,m$ stands for the ith eigenspace of D^{-2}, that is, $V_i = \langle \mathbf{f}_i \rangle$, $i = 1,\ldots,m$, with $\{\mathbf{f}_1,\ldots,\mathbf{f}_m\}$ the canonical basis in $\mathbb{R}^m$. Then to minimize $f_2(V)$ we should take $V_0 = \langle \mathbf{f}_{m-r+1},\ldots,\mathbf{f}_m\rangle$, $\lambda_j = \gamma_{m-r+j}, j = 1,\ldots,r$, and consequently $f_2(V_0) = \sum_{i=1}^r \gamma_{m-r+i} z_i^2$. Therefore,

$$\int \rho\left(\frac{\mathbf{z}^t \Gamma^{1/2} P_V \Gamma^{1/2}\mathbf{z}}{\sigma}\right) f_0\left(\mathbf{z}^t\mathbf{z}\right) d\mathbf{z} = \int \rho\left(\frac{\mathbf{w}^t \Lambda \mathbf{w}}{\sigma}\right) f_0\left(\mathbf{w}^t\mathbf{w}\right) d\mathbf{w}$$
$$> \int \rho\left(\frac{\mathbf{w}^t \Lambda_0 \mathbf{w}}{\sigma}\right) f_0\left(\mathbf{w}^t\mathbf{w}\right) d\mathbf{w},$$

with $\Lambda_0 = \mathrm{diag}(\gamma_{m-r+1},\ldots,\gamma_m,0,\ldots,0)$ and $P_{V_0} = \sum_{i=1}^r \mathbf{f}_{m-r+i}\mathbf{f}_{m-r+i}^t$. This entails that if $\sigma(P_V)$ and $\sigma(P_{V_0})$ solve respectively the equations

$$\delta = \int \rho\left(\frac{\mathbf{z}^t \Gamma^{1/2} P_V \Gamma^{1/2}\mathbf{z}}{\sigma(P_V)}\right) f_0\left(\mathbf{z}^t\mathbf{z}\right) d\mathbf{z}$$
$$\delta = \int \rho\left(\frac{\mathbf{z}^t \Gamma^{1/2} P_{V_0} \Gamma^{1/2}\mathbf{z}}{\sigma(P_{V_0})}\right) f_0\left(\mathbf{z}^t\mathbf{z}\right) d\mathbf{z},$$

then $\sigma(P_{V_0}) < \sigma(P_V)$. Recall that $M_0 = P\Gamma P^t$. From (49), $P_{V_0} = (CP)^t CP = P^t C^t CP = \sum_{i=1}^r \mathbf{f}_{m-r+i}\mathbf{f}_{m-r+i}^t$. Then,

$$C^t C = \sum_{i=1}^{r} (P\mathbf{f}_{m-r+i})(P\mathbf{f}_{m-r+i})^t,$$

and the Fisher consistency follows.

Proof of Corollary 1. If we take $(\widetilde{\mathbf{x}}^t, \widetilde{\mathbf{y}}^t)^t$ as in Theorem 1, then we obtain that the location parameter and the s smallest eigenvalues and eigenvectors of the matrix M in (17) minimize the robust scale σ of the residuals, with $s = \min(p,q)$. If we look for the critical points of $h(C,\mathbf{a},\Lambda) = \sigma(C,\mathbf{a}) + \mathrm{tr}(CC^t\Lambda^t) - \mathrm{tr}(\Lambda^t)$, with $C \in \mathbb{R}^{s\times(p+q)}$, $\Lambda \in \mathbb{R}^{s\times s}$, we get the eigensystem

$$\tilde{M}C^t = C^t\Delta,$$

with $\Delta \in \mathbb{R}^{s\times s}$ a diagonal positive definite matrix and $\tilde{M} = E_{\tilde{F}}\psi(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^t$, if $\mathbf{z} \sim \tilde{F}$ denotes the distribution of $(\widetilde{\mathbf{x}}^t, \widetilde{\mathbf{y}}^t)^t$. Therefore $\tilde{M}$ and M_0 share the same eigenvectors and eigenvalues, which would say that they must coincide. Therefore, if F stands for the elliptical distribution, then

$$E_F\psi(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^t = \begin{pmatrix} \Sigma_{\mathbf{xx}}^{1/2} & 0_{p\times q} \\ 0_{q\times p} & \Sigma_{\mathbf{yy}}^{1/2} \end{pmatrix} M_0 \begin{pmatrix} \Sigma_{\mathbf{xx}}^{1/2} & 0_{p\times q} \\ 0_{q\times p} & \Sigma_{\mathbf{yy}}^{1/2} \end{pmatrix} = c\Sigma.$$

Proof of Lemma 1. Let $D_o^0 = \begin{pmatrix} A_o^0 & -B_o^0 \end{pmatrix}$ and $\mathbf{a}_o^0 = A_o^0 \tilde{\boldsymbol{\mu}}_{\mathbf{x}} - B_o^0 \tilde{\boldsymbol{\mu}}_{\mathbf{y}}$ be the current values, with $\tilde{\boldsymbol{\mu}}_{\mathbf{x}} = \frac{\sum \psi_i \tilde{\mathbf{x}}_i}{\sum \psi_i}$, $\tilde{\boldsymbol{\mu}}_{\mathbf{y}} = \frac{\sum \psi_i \tilde{\mathbf{y}}_i}{\sum \psi_i}$, and let $D_o^1 = \begin{pmatrix} A_o^1 & -B_o^1 \end{pmatrix}$and $\mathbf{a}_o^1$ be the values corresponding to the next iteration. Recall that $A_j = A_o^j \left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2}$ and $B_j = B_o^j \left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2}$, with $A_o^j \left(A_j^o\right)^t = I_r = B_o^j \left(B_o\right)^t$, $j = 0, 1$. Then $D_j = \left(A_o^j \left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2} \quad -B_o^j \left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2}\right) = D_o^j \hat{S}, j = 0, 1,$ with $D_j = \begin{pmatrix} A_j & -B_j \end{pmatrix}$ and $\hat{S} = \begin{pmatrix} \left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2} & 0_{p\times q} \\ 0_{q\times p} & \left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2} \end{pmatrix}$. For $j = 0, 1$ put $r_{ji} = \left\| A_j \mathbf{x}_i - B_j \mathbf{y}_i - \mathbf{a}_j \right\|^2$, $i = 1, \ldots, n$ and define σ_j, as the solution to

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{r_{ji}}{\sigma_j}\right) = \delta.$$

$\sigma_1 \le \sigma_0$ is equivalent to the inequality

$$\sum_{i=1}^{n} \rho\left(\frac{r_{1i}}{\sigma_0}\right) \le \sum_{i=1}^{n} \rho\left(\frac{r_{0i}}{\sigma_0}\right). \tag{50}$$

Then, let us prove (50). Given $a, b \in (0, \infty)$, the concavity of ρ implies that $\rho(b) - \rho(a) \le \psi(a)(b - a)$, with $\psi = \rho'$. Then,

$$\sum_{i=1}^{n} \rho\left(\frac{r_{1i}}{\sigma_0}\right) - \sum_{i=1}^{n} \rho\left(\frac{r_{0i}}{\sigma_0}\right) \le \sum_{i=1}^{n} \psi_i \left(\frac{r_{1i}}{\sigma_0} - \frac{r_{0i}}{\sigma_0}\right). \tag{51}$$

Let us analyze that $\sum_{i=1}^n \psi_i r_{1i} \le \sum_{i=1}^n \psi_i r_{0i}$. If M_0 is the current matrix given in (34), then it can be easily derived that,

$$\sum_{i=1}^{n} \psi_i r_{ji} = \operatorname{tr}\left(D_o^j M_0 \left(D_o^j\right)^t\right), \quad j = 0, 1.$$

If $j = 1$ then $\sum_{i=1}^n \psi_i r_{ji} = \operatorname{tr}\left(D_o^1 M_0 \left(D_o^1\right)^t\right) = \sum_{k=1}^r \lambda_k^{(1)}$, with $\lambda_1^{(1)} \le \cdots \le \lambda_r^{(1)}$ the first r smallest eigenvalues of (34). It is well known from the linear algebra that $\operatorname{tr}\left(D_o^1 M_0 \left(D_o^1\right)^t\right) \le \operatorname{tr}\left(D_o^0 M_0 \left(D_o^0\right)^t\right)$. Then $\sum_{i=1}^n \psi_i r_{1i} - \sum_{i=1}^n \psi_i r_{0i} \le 0$ and (51) entail (50).

Proof of Lemma 2. According to Lemma 1 the sequence $\sigma^{(k)}$ is decreasing and nonnegative, therefore $\lim_{k\to\infty} \sigma^{(k)} = \sigma_0$. Consider $\left(\tilde{D}, \tilde{\mathbf{a}}\right)$ any accumulation point of $\left\{\left(D^{(k)}, \mathbf{a}^{(k)}\right)\right\}_{k=1}^{\infty}$. Let us observe that $\sigma(\tilde{D}, \tilde{\mathbf{a}}) = \sigma_0$ and that if $(\tilde{D}_1, \tilde{\mathbf{a}}_1)$ is the result of applying the iteration step to $\left(\tilde{D}, \tilde{\mathbf{a}}\right)$ then $\sigma(\tilde{D}_1, \tilde{\mathbf{a}}_1) = \sigma_0$. To see this, take any subsequence $\left\{\left(D^{(k_j)}, \mathbf{a}^{(k_j)}\right)\right\}_{j=1}^{\infty}$ converging to $\left(\tilde{D}, \tilde{\mathbf{a}}\right)$. Call $T : \mathbb{R}^{r\times(p+q)} \times \mathbb{R}^r \to \mathbb{R}^{r\times(p+q)} \times \mathbb{R}^r$ the transformation such that $T(D, \mathbf{a})$ is the result of applying Step 2 to $(D, \mathbf{a})$. By assumption, this transformation is continuous and $(D, \mathbf{a}) \to \boldsymbol{\sigma}(D, \mathbf{a})$ is also a continuous function. Since $\left(D^{(k_j+1)}, \mathbf{a}^{(k_j+1)}\right) = T\left(D^{(k_j)}, \mathbf{a}^{(k_j)}\right)$ then we have

$$\begin{aligned}
&\lim_{j\to\infty} \left(D^{(k_j+1)}, \mathbf{a}^{(k_j+1)}\right) = T\left(\tilde{D}, \tilde{\mathbf{a}}\right) = (\tilde{D}_1, \tilde{\mathbf{a}}_1), \\
&\lim_{j\to\infty} \sigma\left(D^{(k_j)}, \mathbf{a}^{(k_j)}\right) = \sigma(\tilde{D}, \tilde{\mathbf{a}}), \quad \text{and} \\
&\sigma(\tilde{D}_1, \tilde{\mathbf{a}}_1) = \lim_{j\to\infty} \sigma\left(D^{(k_j+1)}, \mathbf{a}^{(k_j+1)}\right) = \sigma_0 = \lim_{j\to\infty} \sigma\left(D^{(k_j)}, \mathbf{a}^{(k_j)}\right) = \sigma(\tilde{D}, \tilde{\mathbf{a}}).
\end{aligned}$$

Thus, if $\tilde{\mathbf{z}} = \begin{pmatrix} \left(\hat{\Sigma}_{\mathbf{xx}}^{(R)}\right)^{-1/2} \mathbf{x} \\ \left(\hat{\Sigma}_{\mathbf{yy}}^{(R)}\right)^{-1/2} \mathbf{y} \end{pmatrix}$, we have that

$$\sum_{i=1}^{n} \rho\left(\frac{\left\|\tilde{D}\tilde{\mathbf{z}}_i - \tilde{\mathbf{a}}\right\|^2}{\sigma_0}\right) = \delta = \sum_{i=1}^{n} \rho\left(\frac{\left\|\tilde{D}_1\tilde{\mathbf{z}}_i - \tilde{\mathbf{a}}_1\right\|^2}{\sigma_0}\right).$$

On the other hand, if $M_0(D^{(k_j)}, \mathbf{a}^{(k_j)})$ stands for the current matrix at the k_jth step then $\operatorname{tr}(D^{(k_j)} M_0(D^{(k_j)}, \mathbf{a}^{(k_j)}) \left(D^{(k_j)}\right)^t) - \operatorname{tr}(D M_0(D^{(k_j)}, \mathbf{a}^{(k_j)}) D^t) \le 0$ for any $D = (D_1, D_2) \in \mathbb{R}^{r\times(p+q)}$, $D_1 D_1^t = I_r = D_2 D_2^t$. Then, by taking a subsequence if necessary and the continuity of the trace function, it also holds that $\operatorname{tr}(\tilde{D}\tilde{M}(\tilde{D},\tilde{\mathbf{a}})\tilde{D}^t) - \operatorname{tr}(D\tilde{M}(\tilde{D},\tilde{\mathbf{a}})D^t) \le 0$ if $\tilde{M}(\tilde{D},\tilde{\mathbf{a}}) = \lim_{j\to\infty} M_0(D^{(k_j)},\mathbf{a}^{(k_j)})$. Then, if $\psi_i = \rho'\left(\frac{\|\tilde{D}\tilde{\mathbf{z}}_i-\tilde{\mathbf{a}}\|^2}{\sigma_0}\right)$, we get that

$$\sum_{i=1}^{n}\rho\left(\frac{\left\|\tilde{D}\tilde{\mathbf{z}}_i-\tilde{\mathbf{a}}\right\|^2}{\sigma_0}\right)-\sum_{i=1}^{n}\rho\left(\frac{\left\|D\tilde{\mathbf{z}}_i-\mathbf{a}\right\|^2}{\sigma_0}\right)\le\sum_{i=1}^{n}\psi_i\left(\frac{\left\|\tilde{D}\tilde{\mathbf{z}}_i-\tilde{\mathbf{a}}\right\|^2}{\sigma_0}-\frac{\left\|D\tilde{\mathbf{z}}_i-\mathbf{a}\right\|^2}{\sigma_0}\right)$$
$$=\operatorname{tr}(\tilde{D}\tilde{M}(\tilde{D},\tilde{\mathbf{a}})\tilde{D}^t)-\operatorname{tr}(D\tilde{M}(\tilde{D},\tilde{\mathbf{a}})D^t)\le 0.$$

Then, $\left(\tilde{D},\tilde{\mathbf{a}}\right)$ is a local minimum of the function $h(D,\mathbf{a})=\sum_{i=1}^{n}\rho\left(\frac{\|D\tilde{\mathbf{z}}_i-\mathbf{a}\|^2}{\sigma_0}\right)$. Let us call $N(\tilde{D},\tilde{\mathbf{a}})$ a neighborhood of $\left(\tilde{D},\tilde{\mathbf{a}}\right)$ in which $h(\tilde{D},\tilde{\mathbf{a}})\le h(D,\mathbf{a})$ for any $(D,\mathbf{a})\in N(\tilde{D},\tilde{\mathbf{a}})$. This entails that $\sigma_0=\sigma(\tilde{D},\tilde{\mathbf{a}})\le\sigma(D,\mathbf{a})$ for any $(D,\mathbf{a})\in N(\tilde{D},\tilde{\mathbf{a}})$ since $\delta=h(\tilde{D},\tilde{\mathbf{a}})\le h(D,\mathbf{a})$ for any $(D,\mathbf{a})\in N(\tilde{D},\tilde{\mathbf{a}})$.

ten Berge [18] found the relationship between CCA and PCA which was used to define the SM-estimators for canonical correlations. Next, Lemma 3 rephrases the result to make it clearer.

Lemma 3. *Let $\left(\mathbf{x}^t,\mathbf{y}^t\right)^t\in\mathbb{R}^{p+q}$ be a random vector with finite second moments. Let E be the set of eigenvalues of M in (17).*

(a) *Let $\lambda\in E$, with $\lambda\ne 1$, and $\left(\mathbf{v}^t,\mathbf{w}^t\right)^t\in\mathbb{R}^{p+q}$, with $\mathbf{v}\in\mathbb{R}^p$ and $\mathbf{w}\in\mathbb{R}^q$ such that $M\left(\mathbf{v}^t,\mathbf{w}^t\right)^t=\lambda\left(\mathbf{v}^t,\mathbf{w}^t\right)^t$. Then* (i) *$\Sigma_{\mathbf{xx}}^{-1/2}\mathbf{v}$ and $\Sigma_{\mathbf{yy}}^{-1/2}\mathbf{w}$ are canonical vectors verifying (11) and (12) associated to the squared canonical correlation $(\lambda-1)^2$. Furthermore, $\lambda\in(0,2)$.* (ii) *If $\lambda_j\in E, j=1,2,\ \lambda_1\ne\lambda_2, 2-\lambda_1-\lambda_2\ne 0,\ \left(\mathbf{v}_j^t,\mathbf{w}_j^t\right)^t\in\mathbb{R}^{p+q}$, with $\mathbf{v}_j\in\mathbb{R}^p$ and $\mathbf{w}_j\in\mathbb{R}^q$ such that they verify that $M\left(\mathbf{v}_j^t,\mathbf{w}_j^t\right)^t=\lambda_j\left(\mathbf{v}_j^t,\mathbf{w}_j^t\right)^t$, then $\mathbf{v}_1^t\mathbf{v}_2=0=\mathbf{w}_1^t\mathbf{w}_2$.*

(b) *Let $1\le j\le r$ and $\mathbf{v}_j$ be an eigenvector of (11) with eigenvalue γ_j. Then,* (i) *$\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\mathbf{v}_j$ is an eigenvector of (12) associated with the same eigenvalue γ_j, that is,*

$$\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}\left(\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\mathbf{v}_j\right)=\gamma_j\left(\Sigma_{\mathbf{yy}}^{-1}\Sigma_{\mathbf{yx}}\mathbf{v}_j\right).$$

(ii) *If $\gamma_j>0$, then $\left(\mathbf{v}_j^t\Sigma_{\mathbf{xx}}^{1/2},\gamma_j^{-1/2}\mathbf{v}_j^t\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2}\right)^t$ and $\left(\mathbf{v}_j^t\Sigma_{\mathbf{xx}}^{1/2},-\gamma_j^{-1/2}\mathbf{v}_j^t\Sigma_{\mathbf{xy}}\Sigma_{\mathbf{yy}}^{-1/2}\right)^t$ are eigenvectors for M, that is,*

$$M\begin{pmatrix}\Sigma_{\mathbf{xx}}^{1/2}\mathbf{v}_j\\ \gamma_j^{-1/2}\Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}}\mathbf{v}_j\end{pmatrix}=\left(1+\gamma_j^{1/2}\right)\begin{pmatrix}\Sigma_{\mathbf{xx}}^{1/2}\mathbf{v}_j\\ \gamma_j^{-1/2}\Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}}\mathbf{v}_j\end{pmatrix}$$
$$M\begin{pmatrix}\Sigma_{\mathbf{xx}}^{1/2}\mathbf{v}_j\\ -\gamma_j^{-1/2}\Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}}\mathbf{v}_j\end{pmatrix}=\left(1-\gamma_j^{1/2}\right)\begin{pmatrix}\Sigma_{\mathbf{xx}}^{1/2}\mathbf{v}_j\\ -\gamma_j^{-1/2}\Sigma_{\mathbf{yy}}^{-1/2}\Sigma_{\mathbf{yx}}\mathbf{v}_j\end{pmatrix}.$$

Furthermore, the eigenvalues of M are symmetrically located around 1.

(c) *$1\in E$ if and only if either $p\ne q$ or* $\operatorname{rank}(\Sigma_{\mathbf{xy}})<\min\{p,q\}$.

References

[1] J.R. Berrendero, Contribuciones a la Teoría de la Robustez respecto al Sesgo (Ph.D. thesis), Universidad Carlos III de Madrid (in Spanish).

[2] G. Boente, L. Orellana, A robust approach to common principal components, in: Trends in Mathematics, Birkhäuser Verlag, 2001, pp. 117–145.

[3] J.A. Branco, C. Croux, P. Filzmoser, M.R. Oliveira, Robust canonical correlations: a comparative study, Comput. Statist. 20 (2) (2005) 203–229.

[4] C. Croux, C. Dehon, Estimators of the multiple correlation coeficient: Local robustness and confidence intervals, Statist. Papers 44 (2003) 315–334.

[5] S. Das, P.K. Sen, Canonical correlations, in: P. Armitage, T. Colton (Eds.), Encyclopedia of Biostatistics, John Wiley & Sons, Ltd., New York, 1998, pp. 468–482.

[6] P.L. Davies, Asymptotic behavior of S-estimators of multivariate location estimators and dispersion matrices, Ann. Statist. 15 (1987) 1269–1292.

[7] P. Filzmoser, C. Dehon, C. Croux, Outlier resistant estimators for canonical correlation analysis, in: J.G. Betlehem, P.G.M. van der Heijden (Eds.), COMPSTAT: Proceedings in Computational Statistics, Physica-Verlag, Heldelber, 2000, pp. 301–306.

[8] R. Furrer, M.G. Genton, Aggregation-cokriging for highly multivariate spatial data, Biometrika 98 (2011) 615–631.

[9] K. Hoffman, R. Kunze, Linear Algebra, second ed., PHI Learning, 2009.

[10] H. Hotelling, Relations between two sets of variables, Biometrika 28 (1936) 321–377.

[11] G. Karnel, Robust canonical correlation and correspondence analysis, in: The Frontiers of Statistical Scientific and Industrial Applications, (Vol. II of the Proceedings of ICOSCO-I, The First International Conference on Statistical Computing), American Sciences Press, Strasbourg, 1991, pp. 335–354.

[12] R.A. Maronna, Principal components and orthogonal regression based on robust scales, Technometrics 47 (3) (2005) 264–273.

[13] C.Radhakrishna Rao, H. Toutenberg, Linear Models: Least Squares and Alternatives, second ed., Springer, 1999.

[14] M. Romanazzi, Influence in canonical correlation analysis, Psychometrika 57 (1992) 237–259.

[15] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossmann, G. Pflug, I. Vincze, W. Wertz (Eds.), Mathematical Statistics and Applications, Vol. B, Reidel, Dordrecht, 1985, pp. 283–297.

[16] G.A.F. Seber, Multivariate Observations, second ed., John Wiley & Sons, Inc., 1984.

[17] S. Taskinen, C. Croux, A. Kankainen, E. Ollila, H. Oja, Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices, J. Multivariate Anal. 97 (2006) 359–384.

[18] J.M.F. ten Berge, On the equivalence of two oblique congruence rotation methods and orthogonal approximations, Psychometrika 44 (1979) 359–364.

[19] H. Wold, Nonlinear estimation by iterative least squares procedures, in: F.N. Davis (Ed.), A Festschrift for J. Neyman, Wiley, New York, 1966, pp. 411–444.

[20] V.J. Yohai, M. García Ben, Canonical variables as optimal predictors, Ann. Statist. 8 (4) (1980) 865–869.

[21] R.H. Zamar, Bias robust estimation in orthogonal regression, Ann. Statist. 20 (4) (1992) 1875–1888.