Robust Canonical Correlations: A Comparative Study

J.A. Branco[1], C. Croux[2], P. Filzmoser[3], and M.R. Oliveira[1]

[1] Department of Mathematics and Center for Mathematics and its Applications, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[2] Department of Applied Economics, K.U.Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[3] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria

Summary

Several approaches for robust canonical correlation analysis will be presented and discussed. A first method is based on the definition of canonical correlation analysis as looking for linear combinations of two sets of variables having maximal (robust) correlation. A second method is based on alternating robust regressions. These methods are discussed in detail and compared with the more traditional approach to robust canonical correlation via covariance matrix estimates. A simulation study compares the performance of the different estimators under several kinds of sampling schemes. Robustness is studied as well by breakdown plots.

Keywords: Alternating Regressions, Canonical Correlation, Correlation Measures, Projection Pursuit, Robustness, Robust Covariance Estimation, Robust Regression

1 Introduction

Canonical correlation analysis (CCA) is a multivariate statistical method which was introduced by Hotelling (1936). The aim of CCA is to identify and quantify the relations between a p-dimensional random variable $\boldsymbol{x}$ and a q-dimensional random variable $\boldsymbol{y}$. (Throughout the paper vectors will be denoted in bold.) Without loss of generality we assume $p \leq q$. We use the following notations for the expectations and covariances,

$$E(\boldsymbol{x}) = \boldsymbol{\mu}_x \qquad \text{and} \qquad E(\boldsymbol{y}) = \boldsymbol{\mu}_y,$$

$$\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{xx} \qquad \text{and} \qquad \mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{\Sigma}_{yy},$$

$$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_{yx}^t,$$

and denote the joint covariance matrix of $(\boldsymbol{x}^t, \boldsymbol{y}^t)^t$ by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}. \tag{1.1}$$

We assume that $\boldsymbol{\Sigma}$ has full rank $p + q$.

In CCA we want to study the associations between $\boldsymbol{x}$ and $\boldsymbol{y}$ as measured by the correlations between linear combinations of both sets of variables. Herefore one looks for

$$(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1) = \underset{\boldsymbol{a}, \boldsymbol{b}}{\mathrm{argmax}}\, \mathrm{Corr}(\boldsymbol{a}^t \boldsymbol{x}, \boldsymbol{b}^t \boldsymbol{y}), \tag{1.2}$$

where $\boldsymbol{\alpha}_1 \in I\!R^p$ and $\boldsymbol{\beta}_1 \in I\!R^q$ is the resulting first pair of canonical vectors or canonical coefficients. The linear combinations

$$u_1 = \boldsymbol{\alpha}_1^t \boldsymbol{x} \qquad \text{and} \qquad v_1 = \boldsymbol{\beta}_1^t \boldsymbol{y}$$

are called the first pair of canonical variates. Note that the vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ are only determined up to a multiple by (1.2). To identify them (up to sign) one usually requires that $\mathrm{Var}(u_1) = \mathrm{Var}(v_1) = 1$.

If $rank\,(\boldsymbol{\Sigma}_{xy}) > 1$, the first canonical vectors are not describing the complete dependency structure between $\boldsymbol{x}$ and $\boldsymbol{y}$. Higher order canonical vectors are then recursively defined for $l = 2, \ldots, p$ as

$$(\boldsymbol{\alpha}_l, \boldsymbol{\beta}_l) = \underset{\boldsymbol{a}, \boldsymbol{b}}{\mathrm{argmax}}\, \mathrm{Corr}(\boldsymbol{a}^t \boldsymbol{x}, \boldsymbol{b}^t \boldsymbol{y}), \tag{1.3}$$

yielding the pair of canonical variates of order l

$$u_l = \boldsymbol{\alpha}_l^t \boldsymbol{x} \qquad \text{and} \qquad v_l = \boldsymbol{\beta}_l^t \boldsymbol{y},$$

that need to verify the restrictions

$$\text{Cov}(u_l, u_j) = \boldsymbol{\alpha}_l^t \boldsymbol{\Sigma}_{xx} \boldsymbol{\alpha}_j = \text{Cov}(v_l, v_j) = \boldsymbol{\beta}_l^t \boldsymbol{\Sigma}_{yy} \boldsymbol{\beta}_j = \delta_{lj}, \tag{1.4}$$

where $\delta_{lj} = 1$ if $l = j$ and 0 otherwise ($1 \leq j < l$). The correlation ρ_l between the canonical variates of the l-th pair,

$$\rho_l = \text{Corr}(u_l, v_l),$$

is called the l-th canonical correlation ($l = 1, \ldots, p$).

The above CCA optimization problems (1.2) and (1.3) have a fairly simple solution (see e.g. Johnson and Wichern, 1998, Chapter 10). The canonical coefficients $\boldsymbol{\alpha}_l$ and $\boldsymbol{\beta}_l$ are the eigenvectors corresponding to the eigenvalues $\rho_1^2 \geq \ldots \geq \rho_p^2 > 0$ of the matrices

$$\boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} \qquad \text{and} \qquad \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}. \tag{1.5}$$

Both of the above matrices have the same eigenvalues ρ_l^2 which are the squared canonical correlations. CCA is a standard tool in multivariate analysis and is covered by most textbooks on multivariate statistics. In a review by Das and Sen (1998) CCA is shown to be a useful method of data reduction which helps to understand complex relationships among sets of variables from a wide range of applied fields.

The canonical coefficients, variates, and correlations are estimated by taking the sample covariances in (1.1) and computing the eigenvectors and -values of the matrices in (1.5). However, since the classical estimator of a covariance matrix is very vulnerable with respect to outlying observations, also the eigenvectors and -values based on the sample covariance matrices will be very sensitive, as was shown in the context of CCA by Romanazzi (1992). An obvious approach to robust CCA is to estimate the population covariance matrix in (1.1) robustly. Karnel (1991) took an M-estimator as robust estimator of $\boldsymbol{\Sigma}$ which, however, has poor robustness properties in higher dimensions. Croux and Dehon (2002) used the minimum covariance determinant (MCD) estimator which has a high breakdown point (Rousseeuw, 1985). Distributional properties for CCA based on robust estimates of the covariance matrix have recently been obtained by Taskinen et al (2003).

Using a robust estimator of $\boldsymbol{\Sigma}$ and solving the eigenvalue problem (1.5) boils down to robustify the solutions of the problems (1.2) and (1.3), but the canonical vectors obtained in this way have no natural interpretation anymore. Therefore, a procedure will be studied which robustifies the initial

definition of CCA, by replacing the correlation measure in (1.2) and (1.3) by robust measures of correlation. This approach, outlined in Section 2, is called the projection pursuit (PP) approach. The idea of Oliveira and Branco (2000) to extract the canonical variates sequentially is followed.

In an old paper of Wold (1966) a completely different approach, avoiding the use of covariance matrices, is suggested. Wold proposes an iterative alternating regression scheme, allowing to obtain the canonical variates when the covariance matrices have rank deficiencies. The latter occurs frequently when the dimension p is large in comparison to the number of observations, as is typical in chemometrics and spectroscopy. Performing robust alternating regressions (RAR) is then another approach to robust CCA. This idea has already been explored in Filzmoser et al. (2000), but only for obtaining the first canonical variates. In Section 3 a full and complete treatment of the RAR approach in the setting of CCA is presented.

The contribution of this paper is that it gives a complete description of algorithms to perform projection pursuit and robust alternating regression based CCA. These algorithms have been implemented in `Splus`, a software package with a large variety of robust methods. The programs are available at
http://www.statistik.tuwien.ac.at/public/filz/programs.html.
Efficiency and robustness comparisons between the more traditional approach using robust covariance matrices and the PP- and RAR-based methods are made in Section 4 by means of a simulation study and breakdown plots. Furthermore, the robust correlation measures are used for tests of independence. Section 5 provides conclusions.

2 Robust CCA based on Projection Pursuit

The principle of projection pursuit (PP) is to find structures in data by looking at lower dimensional projections (Huber, 1985). Canonical analysis can be seen as PP-technique since it searches for two directions $\boldsymbol{a}$ and $\boldsymbol{b}$ maximizing the correlation of the variables $\boldsymbol{x}$ and $\boldsymbol{y}$ projected on these directions: $\mathrm{Corr}(\boldsymbol{a}^t\boldsymbol{x}, \boldsymbol{b}^t\boldsymbol{y})$. The latter quantity is called the projection pursuit index (PI). Taking an ordinary Pearson correlation coefficient yields the classical non robust approach. The idea is therefore to work with a different PI, i.e. a robust estimator of the correlation. Note that it is sufficient to have a measure of bivariate correlation between two univariate variables u and v. Three different robust projection indices will be considered:

- *Correlation derived from a bivariate M estimator* (**PP-M**): Given a 2-dimensional random variable $\boldsymbol{z} = (u, v)^t$, the M estimator of Maronna (1976) with M-location $\boldsymbol{\mu}(\boldsymbol{z})$ and M-scatter matrix $\boldsymbol{C}(\boldsymbol{z})$ is defined

implicitly as solutions of the equations

$$\begin{aligned} \boldsymbol{\mu} &= E\Big[w_1\left((\boldsymbol{z}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)\boldsymbol{z}]/E[w_1\left((\boldsymbol{z}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)\Big] \\ \boldsymbol{C} &= E\Big[w_2\left((\boldsymbol{z}-\boldsymbol{\mu})^t\boldsymbol{C}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right)(\boldsymbol{z}-\boldsymbol{\mu})(\boldsymbol{z}-\boldsymbol{\mu})^t\Big] \end{aligned}$$

where $\boldsymbol{\mu}$ is a bivariate vector and $\boldsymbol{C}$ is a symmetric positive definite two-by-two matrix. Furthermore w_1 and w_2 are specified weight functions. We focus on Huber's M estimator, obtained by taking $w_1(d^2) = \max(1, \tau/d)$ and $w_2(d^2) = c\max(1, (\tau/d)^2)$ with $\tau = \chi^2_{2,0.9}$ the 10% upper quantile of a chi-squared distribution with 2 degrees of freedom and c selected to obtain a consistent estimator of the covariance matrix at normal distributions (Huber, 1981). The correlation measure is then directly computed from $\boldsymbol{C}(\boldsymbol{z})$.

- *Correlation derived from a bivariate MCD estimator* (**PP-MCD**): Instead of a bivariate M estimator, which looses robustness under huge amounts of contamination, a high breakdown multivariate scatter matrix will be taken. A popular estimator is the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1985, and Rousseeuw and Van Driessen, 1999, for a fast algorithm). The minimum covariance determinant (MCD) estimator is determined by that subset of observations of size h which minimizes the determinant of the sample covariance matrix, computed from only these h points. The location estimator is the average of these h points, whereas the scatter estimator is proportional to their covariance matrix. As a compromise between robustness and efficiency, selected $h = \lfloor 0.75n \rfloor$ will be taken (n is the sample size).

- *Spearman's rank correlation* (**PP-SPM**): This well-known measure of association $\rho_S(u, v)$ is defined as the correlation between the ranks of two random variables u and v. It has a non-parametric nature and does not rely on any symmetry conditions. When sampling from a bivariate normal distribution, ρ_S is not estimating the same quantity as Pearson's correlation. To compare the estimation of the canonical correlations by PP-SPM with other methods, we will apply the transformation $2\sin(\pi\rho_S/6)$ to get consistent estimation under normality.

The above projection indices will be used for finding the first pair of canonical vectors as in (1.2), but also for finding canonical vectors of higher order (1.3). To fulfill the restrictions (1.4) for higher order canonical vectors, the following procedure is proposed. We start by estimating $\boldsymbol{\Sigma}$ using a highly robust estimator, like the reweighted MCD estimator (Rousseeuw and Van Driessen, 1999). Then $\boldsymbol{\Sigma}$ is decomposed as in (1.1), and a spectral decomposition of $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ is performed leading to $\boldsymbol{\Sigma}_{xx} = \boldsymbol{U}\boldsymbol{K}\boldsymbol{U}^t$ and $\boldsymbol{\Sigma}_{yy} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^t$, where

$\boldsymbol{K}, \boldsymbol{L}$ are diagonal and $\boldsymbol{U}, \boldsymbol{V}$ orthogonal matrices. The original variables are now transformed into

$$(\mathbf{x}^*, \mathbf{y}^*) = \left(\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{U}^t\mathbf{x}, \boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{V}^t\mathbf{y}\right).$$

Because of the equivariance properties of canonical analysis, the new variables $(\mathbf{x}^*, \mathbf{y}^*)$ have the same canonical correlations as the original variables $(\mathbf{x}, \mathbf{y})$, and the canonical coefficients verify the relations:

$$\boldsymbol{\alpha}_l = \boldsymbol{U}\boldsymbol{K}^{-1/2}\boldsymbol{\alpha}_l^* \text{ and } \boldsymbol{\beta}_l = \boldsymbol{V}\boldsymbol{L}^{-1/2}\boldsymbol{\beta}_l^*,$$

for $l = 1, \ldots, p$.

For finding the first canonical vectors $\boldsymbol{\alpha}_1^*$ and $\boldsymbol{\beta}_1^*$, the projection index $PI(\boldsymbol{a}^t\boldsymbol{x}^*, \boldsymbol{b}^t\boldsymbol{y}^*)$ needs to be maximized under the restrictions $\mathrm{Var}\,(\boldsymbol{a}^t\mathbf{x}^*) = \boldsymbol{a}^t\boldsymbol{a} = 1$ and $\mathrm{Var}\left(\boldsymbol{b}^t\mathbf{y}^*\right) = \boldsymbol{b}^t\boldsymbol{b} = 1$. To get rid of the side constraint in the function to be maximized, it is convenient to write the arguments $\boldsymbol{a}$ and $\boldsymbol{b}$ in polar coordinates. Let $(\theta_1, \ldots, \theta_{p-1})^t$ be the polar coordinates of a vector with norm equal to one. The projection index is then maximized over the set of $(p-1)$ angles corresponding to a unit norm vector $\boldsymbol{a}$ and the set of $(q-1)$ angles corresponding to a unit norm vector $\boldsymbol{b}$. Herefore a standard maximization routine has been used, where the maximization is over a hyperrectangle. Once optimal angles $(\theta_1, \ldots, \theta_{p-1})^t$ are obtained, one can go back to the previous coordinates using the recursive relation:

$$k = 2, \qquad \boldsymbol{\alpha}_{(2)} = \begin{bmatrix} \cos\theta_1 \\ \sin\theta_1 \end{bmatrix}, \theta_1 \in [0, \pi[,$$

$$2 < k \leq p, \quad \boldsymbol{\alpha}_{(k)} = \begin{bmatrix} \boldsymbol{\alpha}_{(k-1)}\sin\theta_{k-1} \\ \cos\theta_{k-1} \end{bmatrix}, \begin{array}{l} \theta_1 \in [0, 2\pi[, \\ \theta_j \in [0, \pi], j = 2, \ldots, k-2, \\ \theta_{k-1} \in [0, \frac{\pi}{2}]. \end{array}$$

resulting in $\boldsymbol{\alpha}_1^* = \boldsymbol{\alpha}_{(p)}$ and similarly for $\boldsymbol{\beta}_1^*$.

Suppose now that the first $(l-1)$ pairs of canonical coefficients, $(\boldsymbol{\alpha}_j^*, \boldsymbol{\beta}_j^*)$, $j = 1, \ldots, l-1$, are already found and we want to estimate the l-th pair $(2 \leq l \leq p)$. Note that the restrictions $\boldsymbol{\alpha}_l^t\boldsymbol{\Sigma}_{xx}\boldsymbol{\alpha}_j = 0$ and $\boldsymbol{\beta}_l^t\boldsymbol{\Sigma}_{yy}\boldsymbol{\beta}_j = 0$ translate into orthogonality constraints $\boldsymbol{\alpha}_l^{*t}\boldsymbol{\alpha}_j^* = 0$ and $\boldsymbol{\beta}_l^{*t}\boldsymbol{\beta}_j^* = 0$ for $l \neq j$. First construct two orthogonal matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as $\boldsymbol{A} = [\boldsymbol{\alpha}_1^* \ldots \boldsymbol{\alpha}_{l-1}^* | \boldsymbol{A}^r]$ and $\boldsymbol{B} = [\boldsymbol{\beta}_1^* \ldots \boldsymbol{\beta}_{l-1}^* | \boldsymbol{B}^r]$ where $\boldsymbol{A}^r$ and $\boldsymbol{B}^r$ are orthonormal bases of the subspaces orthogonal to $\boldsymbol{\alpha}_1^*, \ldots, \boldsymbol{\alpha}_{l-1}^*$ and $\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_{l-1}^*$, respectively. Next we project the variables on the subspaces orthogonal to the canonical vectors already retrieved:

$$(\mathbf{x}^{**}, \mathbf{y}^{**}) = \left((\boldsymbol{A}^r)^t\,\mathbf{x}^*, (\boldsymbol{B}^r)^t\,\mathbf{y}^*\right).$$

Now we need to find $\boldsymbol{\alpha}^{**}$ and $\boldsymbol{\beta}^{**}$ maximizing $PI(\boldsymbol{a}^t\boldsymbol{x}^{**}, \boldsymbol{b}^t\boldsymbol{y}^{**})$ under the constraint that the norm of $\boldsymbol{a}^{**}$ and $\boldsymbol{b}^{**}$ equals one. This can be done as before, by passing to polar coordinates. By back-transformation $\boldsymbol{\alpha}_l^* = \boldsymbol{A}^r\boldsymbol{\alpha}^{**}$

and $\boldsymbol{\beta}_l^* = \boldsymbol{B}^r\boldsymbol{\beta}^{**}$ are obtained. Projecting onto the subspaces has two advantages: (i) the maximization problem is now in a lower dimensional space since $\boldsymbol{A}^r$ and $\boldsymbol{B}^r$ are matrices with dimensions $p \times (p-l+1)$ and $q \times (q-l+1)$, respectively; (ii) it is immediate to see that $\boldsymbol{\alpha}_l^*$ and $\boldsymbol{\beta}_l^*$ will be orthogonal to all previously found $\boldsymbol{\alpha}_j^*$ and $\boldsymbol{\beta}_j^*$, such that the side condition for higher order canonical coefficients will automatically be fulfilled.

After having obtained the canonical vectors for $(\boldsymbol{x}^*, \boldsymbol{y}^*)$, the solution for the original variables is obtained by applying the transformation:

$$(\boldsymbol{\alpha}_l, \boldsymbol{\beta}_l) = \left(\boldsymbol{U}\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{\alpha}_l^*, \boldsymbol{V}\boldsymbol{L}^{-\frac{1}{2}}\boldsymbol{\beta}_l^*\right), \quad l = 2, \cdots, p. \tag{2.1}$$

Finally, the canonical correlations are estimated by computing $\rho_l = PI(u_l, v_l)$ where $(u_l, v_l) = \left(\boldsymbol{\alpha}_l^t\mathbf{x}, \boldsymbol{\beta}_l^t\mathbf{y}\right)$, for $l = 1, \cdots, p$.

3 Robust CCA based on Alternating Regressions

Wold (1966) proposed an alternating regression technique for obtaining the solution to CCA. Also Lyttkens (1972) and Tenenhaus (1998, p. 204) mentioned this approach. The general idea behind the procedure is as follows. Suppose we have an initial value for a canonical vector $\boldsymbol{\beta}$. Then the maximization problem (1.2) reduces to

$$\boldsymbol{\alpha} = \underset{\boldsymbol{a}}{\operatorname{argmax}} \operatorname{Corr}(\boldsymbol{a}^t\boldsymbol{x}, \boldsymbol{\beta}^t\boldsymbol{y}). \tag{3.1}$$

But then it follows from standard results on multiple regression, that $\boldsymbol{\alpha}$ is proportional to the regression coefficients $\boldsymbol{a}$ in the model

$$\boldsymbol{\beta}^t\boldsymbol{y} = \boldsymbol{a}^t\boldsymbol{x} + \gamma_1 + \varepsilon_1. \tag{3.2}$$

On the other hand, for a fixed $\boldsymbol{\alpha}$, the optimal $\boldsymbol{\beta}$ is obtained by the maximization of

$$\boldsymbol{\beta} = \underset{\boldsymbol{b}}{\operatorname{argmax}} \operatorname{Corr}(\boldsymbol{\alpha}^t\boldsymbol{x}, \boldsymbol{b}^t\boldsymbol{y}), \tag{3.3}$$

and the solution $\boldsymbol{\beta}$ is proportional to the regression coefficients $\boldsymbol{b}$ in the regression equation

$$\boldsymbol{\alpha}^t\boldsymbol{x} = \boldsymbol{b}^t\boldsymbol{y} + \gamma_2 + \varepsilon_2. \tag{3.4}$$

This leads to an alternating regression scheme which will be described now in detail. First the classical case, where least squares (LS) regressions are used, is presented. This least squares alternating regression scheme has already been outlined in the literature, but we feel that it is useful to recall the different steps for two reasons: (i) for obtaining higher order canonical

variates it was suggested by Wold (1966) to perform subsequent alternating regressions in the residual space. Since the residual matrices have reduced rank, this needs to be done with care. Many details are missing in Wold (1966) and we feel it useful to give a complete description; (ii) to make the analogy with robust alternating regressions more clear.

We will present the method at the sample level. Assume that n observations are sampled from the same distribution as $\boldsymbol{x}$ and $\boldsymbol{y}$, and arrange them as rows in the data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. The vectors containing the scores on the canonical variates are then given by $\boldsymbol{u}_l = \boldsymbol{X}\boldsymbol{\alpha}_l$ and $\boldsymbol{v}_l = \boldsymbol{X}\boldsymbol{\beta}_l$ and will be called canonical variates as well. Note the difference between $\boldsymbol{u_l}$ and u_l, the latter one being a random variable. The method will also work if the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ have not full rank. In this case the maximum number of components to be extracted is $k = \min\{\text{rank}(\boldsymbol{X}), \text{rank}(\boldsymbol{Y})\}$.

3.1 Least Squares Alternating Regressions

We start with the mean centered data matrices

$$\boldsymbol{X}_0 = \boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^t \qquad \boldsymbol{Y}_0 = \boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^t$$

where $\mathbf{1}$ is a vector of length n with elements 1, and $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ are the mean vectors of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Using mean centered matrices avoids adding the intercept terms γ_1 and γ_2 in (3.2) and (3.4).

The canonical vectors $\boldsymbol{\alpha}_l$ and $\boldsymbol{\beta}_l$ and variates $\boldsymbol{u}_l$ and $\boldsymbol{v}_l$ will be retrieved sequentially, for $l = 1, \ldots, k$. In the case $l = 1$ we directly use the centered data matrices $\boldsymbol{X}_0$ and $\boldsymbol{Y}_0$. For $l > 1$ the alternating regression scheme will be applied using deflated matrices $\boldsymbol{X}_{l-1}$ and $\boldsymbol{Y}_{l-1}$. These are obtained as residuals of regressing $\boldsymbol{X}_{l-2}$ and $\boldsymbol{Y}_{l-2}$ on the previously found canonical variates $\boldsymbol{u}_{l-1}$ and $\boldsymbol{v}_{l-1}$, respectively. So, we estimate regression models

$$\boldsymbol{X}_{l-2} = \boldsymbol{u}_{l-1}\boldsymbol{c}^t + \boldsymbol{\varepsilon}_1, \tag{3.5}$$

and

$$\boldsymbol{Y}_{l-2} = \boldsymbol{v}_{l-1}\boldsymbol{d}^t + \boldsymbol{\varepsilon}_2, \tag{3.6}$$

and set $\boldsymbol{X}_{l-1} = \hat{\boldsymbol{\varepsilon}}_1$ and $\boldsymbol{Y}_{l-1} = \hat{\boldsymbol{\varepsilon}}_2$.

In our case of least squares regression, explicit expressions for the deflated data matrices are given by

$$\boldsymbol{X}_{l-1} = \left(\boldsymbol{I}_n - \frac{\boldsymbol{u}_{l-1}\boldsymbol{u}_{l-1}^t}{\boldsymbol{u}_{l-1}^t\boldsymbol{u}_{l-1}}\right)\boldsymbol{X}_{l-2} \quad \text{and } \boldsymbol{Y}_{l-1} = \left(\boldsymbol{I}_n - \frac{\boldsymbol{v}_{l-1}\boldsymbol{v}_{l-1}^t}{\boldsymbol{v}_{l-1}^t\boldsymbol{v}_{l-1}}\right)\boldsymbol{Y}_{l-2},$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Due to standard results from multiple linear regression, the columns in $\boldsymbol{X}_{l-1}$ are orthogonal to $\boldsymbol{u}_{l-1}$. Since the l-th

canonical variate $\boldsymbol{u}_l$ will be a linear combination of the columns of $\boldsymbol{X}_{l-1}$, it will also be orthogonal, hence uncorrelated, to the previously found canonical variates.

Note that the matrices $\boldsymbol{X}_{l-1}$ and $\boldsymbol{Y}_{l-1}$ have reduced rank, the rank being reduced by $(l-1)$. Later on in the algorithm, problems will occur when the inverse of $\boldsymbol{X}_{l-1}^t\boldsymbol{X}_{l-1}$ is needed. Therefore, generalized inverses need to be taken. Let

$$\boldsymbol{X}_{l-1} = \boldsymbol{U}_X\boldsymbol{D}_X\boldsymbol{V}_X^t = \boldsymbol{U}_X^*\boldsymbol{D}_X^*{\boldsymbol{V}_X^*}^t \tag{3.7}$$

be the singular value decomposition. Here $\boldsymbol{D}_X$ is a diagonal matrix including all p singular values, the diagonal matrix $\boldsymbol{D}_X^*$ contains only the strictly positive singular values and $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ are reduced by the rows and columns corresponding to zero singular values. The generalized inverse is then defined by

$$(\boldsymbol{X}_{l-1}^t\boldsymbol{X}_{l-1})^- = \boldsymbol{V}_X^*{\boldsymbol{D}_X^*}^{-2}{\boldsymbol{V}_X^*}^t. \tag{3.8}$$

The generalized inverse of $\boldsymbol{Y}_{l-1}^t\boldsymbol{Y}_{l-1}$ is defined similarly.

Starting value: To start up the alternating regression scheme on the deflated matrices $\boldsymbol{X}_{l-1}$ and $\boldsymbol{Y}_{l-1}$ a starting value $\boldsymbol{\beta}_l^{(0)}$, by analogy to (3.1), is required. Wold (1966) suggests to take arbitrary numbers. According to Tenenhaus (1998, p. 204) we could take the first column of $\boldsymbol{X}_{l-1}$ and regress this vector on $\boldsymbol{Y}_{l-1}$. Selecting the starting value is not so crucial for the classical method, but for a robust method it is much more important to have a good starting value. We propose to regress the first principal component of $\boldsymbol{X}_{l-1}$, denoted by $\boldsymbol{z}_1^{l-1}$, on $\boldsymbol{Y}_{l-1}$. The associated regression model is then

$$\boldsymbol{z}_1^{l-1} = \boldsymbol{Y}_{l-1}\boldsymbol{b}_l^{(0)} + \varepsilon_3 \tag{3.9}$$

with estimated LS regression coefficients

$$\hat{\boldsymbol{b}}_l^{(0)} = (\boldsymbol{Y}_{l-1}^t\boldsymbol{Y}_{l-1})^-\boldsymbol{Y}_{l-1}^t\boldsymbol{z}_1^{l-1}.$$

The starting value is now defined by

$$\boldsymbol{\beta}_l^{(0)} = \frac{\hat{\boldsymbol{b}}_l^{(0)}}{\|\hat{\boldsymbol{b}}_l^{(0)}\|},$$

and the corresponding canonical variate is

$$\boldsymbol{v}_l^{(0)} = \boldsymbol{Y}_{l-1}\boldsymbol{\beta}_l^{(0)}.$$

Further steps in the alternating regression scheme: In step number s (with $s > 1$) we first regress the approximation we found in step $s-1$ for the

canonical variate of the y-part, $\boldsymbol{v}_l^{(s-1)}$, on the deflated matrix for the x-part. According to (3.2),

$$\boldsymbol{v}_l^{(s-1)} = \boldsymbol{X}_{l-1}\boldsymbol{a}_l^{(s)} + \varepsilon_4, \tag{3.10}$$

leading to an LS-estimate

$$\hat{\boldsymbol{a}}_l^{(s)} = (\boldsymbol{X}_{l-1}^t\boldsymbol{X}_{l-1})^-\boldsymbol{X}_{l-1}^t\boldsymbol{v}_l^{(s-1)},$$

being standardized to

$$\boldsymbol{\alpha}_l^{(s)} = \frac{\hat{\boldsymbol{a}}_l^{(s)}}{\|\hat{\boldsymbol{a}}_l^{(s)}\|}.$$

The canonical variate is then defined as

$$\boldsymbol{u}_l^{(s)} = \boldsymbol{X}_{l-1}\boldsymbol{\alpha}_l^{(s)}.$$

For the second part of step s of the alternating regression scheme we follow (3.4) which results in a regression

$$\boldsymbol{u}_l^{(s)} = \boldsymbol{Y}_{l-1}\boldsymbol{b}_l^{(s)} + \varepsilon_5. \tag{3.11}$$

with LS estimate

$$\hat{\boldsymbol{b}}_l^{(s)} = (\boldsymbol{Y}_{l-1}^t\boldsymbol{Y}_{l-1})^-\boldsymbol{Y}_{l-1}^t\boldsymbol{u}_l^{(s)},$$

yielding an updated coefficient

$$\boldsymbol{\beta}_l^{(s)} = \frac{\hat{\boldsymbol{b}}_l^{(s)}}{\|\hat{\boldsymbol{b}}_l^{(s)}\|}$$

and updated canonical variates

$$\boldsymbol{v}_l^{(s)} = \boldsymbol{Y}_{l-1}\boldsymbol{\beta}_l^{(s)}.$$

Then a next step $s+1$ is carried out, and the alternating regression scheme is carried out further up to convergence. The canonical vectors and variates obtained at this stage are then denoted by $\boldsymbol{\alpha}_l^*$, $\boldsymbol{\beta}_l^*$, $\boldsymbol{u}_l^*$, and $\boldsymbol{v}_l^*$. The estimated l-th canonical correlation coefficient is the bivariate sample correlation coefficient $|\mathrm{Corr}(\boldsymbol{u}_l^*, \boldsymbol{v}_l^*)|$.

Note that the canonical vectors are normed to length 1. By rescaling one can easily meet the unit variance restriction (1.4). But this will not change the canonical correlation. For comparing the performance of estimates of canonical vectors from different methods measures will be used that do not depend on the choice of the normalization constraint.

Final solution: In this last step one wishes to express the canonical coefficients and variates in terms of the data matrices $\boldsymbol{X_0}$ and $\boldsymbol{Y_0}$. For $l = 1$ the resulting

canonical vectors and variates are identical with the final solution, i.e. $\boldsymbol{u}_1 = \boldsymbol{u}_1^*$, $\boldsymbol{v}_1 = \boldsymbol{v}_1^*$, $\hat{\boldsymbol{\alpha}}_1 = \boldsymbol{\alpha}_1^*$, and $\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1^*$. For $l > 1$, the final canonical variates have to fulfill the orthogonality restriction to previously found variates, given by $\boldsymbol{u}_j = \boldsymbol{X}_0\boldsymbol{\alpha}_j$ $(j = 1, \ldots, l-1)$. Construct the matrix $\boldsymbol{U}_{l-1} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{l-1}]$. and consider the regression

$$\boldsymbol{u}_l^* = \boldsymbol{U}_{l-1}\boldsymbol{e} + \varepsilon_6. \tag{3.12}$$

The residuals are then given by

$$\hat{\varepsilon}_6 = \left(\boldsymbol{I}_n - \boldsymbol{U}_{l-1}(\boldsymbol{U}_{l-1}^t\boldsymbol{U}_{l-1})^{-1}\boldsymbol{U}_{l-1}^t\right)\boldsymbol{u}_l^*$$

and are orthogonal to the columns of $\boldsymbol{U}_{l-1}$. Therefore we set $\tilde{\boldsymbol{u}}_l = \hat{\varepsilon}_6$.

The final canonical variates $\boldsymbol{u}_l$ will now to be expressed as linear combinations of $\boldsymbol{X}_0$ with coefficients $\boldsymbol{\alpha}_l$. Therefore, we consider

$$\tilde{\boldsymbol{u}}_l = \boldsymbol{X}_0\boldsymbol{f} + \varepsilon_7, \tag{3.13}$$

and obtain the estimated coefficients

$$\hat{\boldsymbol{f}} = (\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t\tilde{\boldsymbol{u}}_l,$$

which are set to the estimated canonical vectors $\hat{\boldsymbol{\alpha}}_l = \hat{\boldsymbol{f}}$. The fitted values $\hat{\tilde{\boldsymbol{u}}}_l = \boldsymbol{X}_0\hat{\boldsymbol{f}}$ are the final canonical variates, i.e. $\boldsymbol{u}_l = \hat{\tilde{\boldsymbol{u}}}_l = \boldsymbol{X}_0\hat{\boldsymbol{\alpha}}_l$. The procedure for obtaining $\boldsymbol{v}_l$ and $\hat{\boldsymbol{\beta}}_l$ is analogous. In fact, using the property that Least Squares residuals are orthogonal to the regressors, it is immediate to check that $\boldsymbol{u}_l = \tilde{\boldsymbol{u}}_l = \hat{\tilde{\boldsymbol{u}}}_l$. To make the generalization to robust regression methods more directly, we did not use these equality in the description above. But in the Appendix, where a summary of the complete algorithm is presented, these relations are used.

3.2 Robust Alternating Regressions

The Robust Alternating Regression (RAR) procedure consist in estimating all regression models listed in Section 3.1 by robust regression instead of least squares regression. A popular estimator is the Least Trimmed Squares (LTS) estimator of (Rousseeuw, 1984), being highly robust, fast to compute, and available in several software packages. Besides the aspect of robustness, another issue is relevant, namely the convergence of the algorithm. For that we have to use a regression method which uses a smooth function for down weighting outliers in order to ensure that the estimates from one iteration to the other will not change drastically. Now LTS regression has a zero/one weighting of outliers: if different observations are detected as outliers in consecutive iterations, instabilities will be created and there is a high risk of non-convergence of the RAR algorithm. Instead of LTS, L_1 regression could

be used, but the latter is not robust against leverage points. A compromise is weighted L_1 regression, where the weights are defined as smooth functions of an appropriate measure for the leverage of each observation. An iterative weighted regression scheme was also used in the context of factor analysis by Croux et al (2003). For a multiple linear regression model

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \qquad (i = 1, \ldots, n) \tag{3.14}$$

the weighted L_1 regression estimator is defined as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^t \boldsymbol{\beta}| w(\boldsymbol{x}_i), \tag{3.15}$$

where $w(\boldsymbol{x}_i)$ is a weight attached to the i-th observation of the x-space.

For RAR, one needs weights for outlying observations in the x-part and in the y-part, since regressions are being alternated. Suppose we want to extract the l-th canonical variates ($l = 1, \ldots, k$), then we have to compute the weights according to outliers in the matrices $\boldsymbol{X}_{l-1}$ and $\boldsymbol{Y}_{l-1}$. However, both matrices have reduced rank for $l > 1$, and since the weight will be computed from robust Mahalanobis distances this creates a problem. Therefore a singular value decomposition (3.7) is performed and the matrices $\boldsymbol{X}_{l-1}^* = \boldsymbol{X}_{l-1}\boldsymbol{V}_X^* = \boldsymbol{U}_X^*\boldsymbol{D}_X^*$ and $\boldsymbol{Y}_{l-1}^* = \boldsymbol{Y}_{l-1}\boldsymbol{V}_Y^* = \boldsymbol{U}_Y^*\boldsymbol{D}_Y^*$ are considered. Similar to Croux et al. (2003) weights $w_i(\boldsymbol{X}_{l-1}^*)$ are defined for an outlying row $\boldsymbol{x}_i^{*(l-1)}$ ($i = 1, \ldots, n$) of the matrix $\boldsymbol{X}_{l-1}^*$ by

$$w_i(\boldsymbol{X}_{l-1}^*) = \min\left(1, \frac{\chi^2_{p^*,0.95}}{\mathrm{RD}_i^2(\boldsymbol{X}_{l-1}^*)}\right) \qquad \text{for } i = 1, \ldots, n. \tag{3.16}$$

Here $\chi^2_{p^*,0.95}$ is the upper 5% critical value of a chi-squared distribution with p^* degrees of freedom, $p^* = k - l + 1$ is the number of columns of $\boldsymbol{X}_{l-1}^*$, and

$$\mathrm{RD}_i(\boldsymbol{X}_{l-1}^*) = \sqrt{(\boldsymbol{x}_i^{*(l-1)} - T(\boldsymbol{X}_{l-1}^*))^t C(\boldsymbol{X}_{l-1}^*)^{-1}(\boldsymbol{x}_i^{*(l-1)} - T(\boldsymbol{X}_{l-1}^*))}$$

for $i = 1, \ldots, n$, are robust distances (Rousseeuw and Van Zomeren, 1990). The robust multivariate location and scatter estimators T and C are taken as the location and scatter part of the MVE estimator (Rousseeuw, 1985) computed from $\boldsymbol{X}_{l-1}^*$. The MVE estimator was chosen here since it performs well as an outlier identifier (see Becker and Gather, 2001), but other robust covariance matrix estimates can also be taken here. In an analogous way weights $w_i(\boldsymbol{Y}_{l-1}^*)$ for outlying rows of $\boldsymbol{Y}_{l-1}^*$ for $i = 1, \ldots, n$ are defined.

As mentioned before, the RAR algorithm mimics now the Least Squares alternating regression scheme, but now using weighted regressions in the iterative part of the algorithm and LTS regressions elsewhere. Note that some of the explicit formulas, applicable for Least Squares regression, are not valid for other regression estimators. The outline of the RAR algorithm is presented in the Appendix.

4 Simulation Study

In this Section the methods are compared by means of a simulation study. We consider:

Class: Classical CCA based on eigenvectors/values of the matrices (1.5), which are estimated by the sample covariance matrix. This approach gives the same results (up to numerical imprecision) as the least squares alternating regression method of Section 3.1.

M: CCA based on eigenvectors/values of the matrices (1.5), now using a $(p+q)$-dimensional M estimator decomposed according to (1.1).

MCD: As **M**, but now using the Minimum Covariance Determinant estimator.

PP-MCD, PP-M, PP-SPM: as defined in Section 2. Recall that *bivariate* MCD and M, and also Spearman rank correlation coefficient are used as projection indices.

RAR: The robust alternating regression from Section 3.2 is implemented as outlined in the Appendix.

We considered several dimensions p and q for each group of variables $\boldsymbol{x}$ and $\boldsymbol{y}$. The number of observations is $n = 500$ and the number of simulations within each setup was $m = 300$. Due to equivariance properties of CCA, the covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$ may be taken as identity matrices, and the choices for $\boldsymbol{\Sigma}_{xy}$ are summarized in Table 1.

Table 1: Simulation setup. $\boldsymbol{\Sigma}_{xx} = \boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_{yy} = \boldsymbol{I}_q$

p	q	$\boldsymbol{\Sigma}_{xy}$
2	2	$\begin{bmatrix} 0.9 & 0 \\ 0 & 1/2 \end{bmatrix}$
2	4	$\begin{bmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \end{bmatrix}$
4	4	$\begin{bmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}$

The following sampling distributions are considered:

- normal distribution (NOR), $N_{p+q}(\boldsymbol{0}, \boldsymbol{\Sigma})$, with zero mean and covariance matrix $\boldsymbol{\Sigma}$

- multivariate t distribution with 3 degrees of freedom (T3). If $\boldsymbol{z} = (z_1, \ldots, z_{p+q})^t \sim N_{p+q}(\mathbf{0}, \boldsymbol{\Sigma})$, and w has a chi-square distribution with r degrees of freedom, then $\boldsymbol{t} = \boldsymbol{z}/\sqrt{w/r}$ has a multivariate t distribution with r degrees of freedom and scatter parameter $\boldsymbol{\Sigma}$
- symmetric contamination (SCN): there is a probability of 0.95 that an observation is generated from $N_{p+q}(\mathbf{0}, \boldsymbol{\Sigma})$ and a 0.05 probability that it is generated from $N_{p+q}(\mathbf{0}, 9\boldsymbol{\Sigma})$.
- asymmetric contamination (ACN): 95% of the data are generated from the $N_{p+q}(\mathbf{0}, \boldsymbol{\Sigma})$, and 5% of the observations equals the point $tr(\boldsymbol{\Sigma})\,\mathbf{1}^t$ (where $tr(\boldsymbol{\Sigma})$ is the trace of $\boldsymbol{\Sigma}$).

The contamination introduced in the sampling schemes SCN and ACN will generate outliers not being extremely far away from the vast majority of the data. This is a more realistic type of outliers. We believe that simulations with very extreme outliers are not of practical interest because these types of outliers could easily be identified by preliminary data analysis. For details on the simulation see also the `Splus` program at the before mentioned website.

The estimated parameters for a replication j $(j = 1, \ldots, m)$ of a specific sampling distribution are denoted by $\hat{\rho}_l^j$, $\hat{\boldsymbol{\alpha}}_l^j$, and $\hat{\boldsymbol{\beta}}_l^j$ for $l = 1, \ldots, k$. These values are compared with the "true" parameters ρ_l, $\boldsymbol{\alpha}_l$ and $\boldsymbol{\beta}_l$ which were derived from the specified matrix $\boldsymbol{\Sigma}$. The following measures of mean squared error (MSE) are computed:

$$\mathrm{MSE}(\hat{\rho}_l) = \frac{1}{m}\sum_{j=1}^{m}(\phi(\hat{\rho}_l^j) - \phi(\rho_l))^2, \tag{4.1}$$

where $\phi(\rho_l) = \tanh^{-1}(\rho_l)$ is the Fisher transformation of ρ_l (which is classically applied to turn a distribution of correlation coefficients towards normality), and for the canonical coefficients

$$\mathrm{MSE}(\hat{\boldsymbol{\alpha}}_l) = \frac{1}{m}\sum_{j=1}^{m}\cos^{-1}\left(\frac{|\boldsymbol{\alpha_l}^t\hat{\boldsymbol{\alpha}}_l^j|}{||\hat{\boldsymbol{\alpha}}_l^j|| \cdot ||\boldsymbol{\alpha}_l||}\right), \tag{4.2}$$

and similarly for $\mathrm{MSE}(\hat{\boldsymbol{\beta}})$. The measure (4.2) is the average value of the positive angles between the vectors $\hat{\boldsymbol{\alpha}}_l^j$ and $\boldsymbol{\alpha}_l$. The use of angles makes the MSE invariant to the choice of the normalization constraint for the canonical coefficients.

The results of the simulation are presented in Figures 1 and 2. Figure 1 shows the mean squared errors for dimensions $p = 2$ and $q = 2$. Pictures 1(a) and 1(b) present the MSEs for the canonical vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$, 1(c) and 1(d)

for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, 1(e) and 1(f) for the transformed canonical correlations $\phi(\rho_1)$ and $\phi(\rho_2)$. The horizontal axes labels the different methods. For a better comparison the resulting values have been connected to lines, and the four line types in the pictures reflect the different sampling schemes models. In all pictures it is visible that asymmetric contamination (ACN) leads to the largest MSEs. The classical method and the M estimator are clearly worse than the other methods. For higher order canonical variates the robust alternating regression (RAR) method is getting worse. All other tested methods, those based on PP and MCD lead to comparable results, and have low MSE over all sampling schemes.

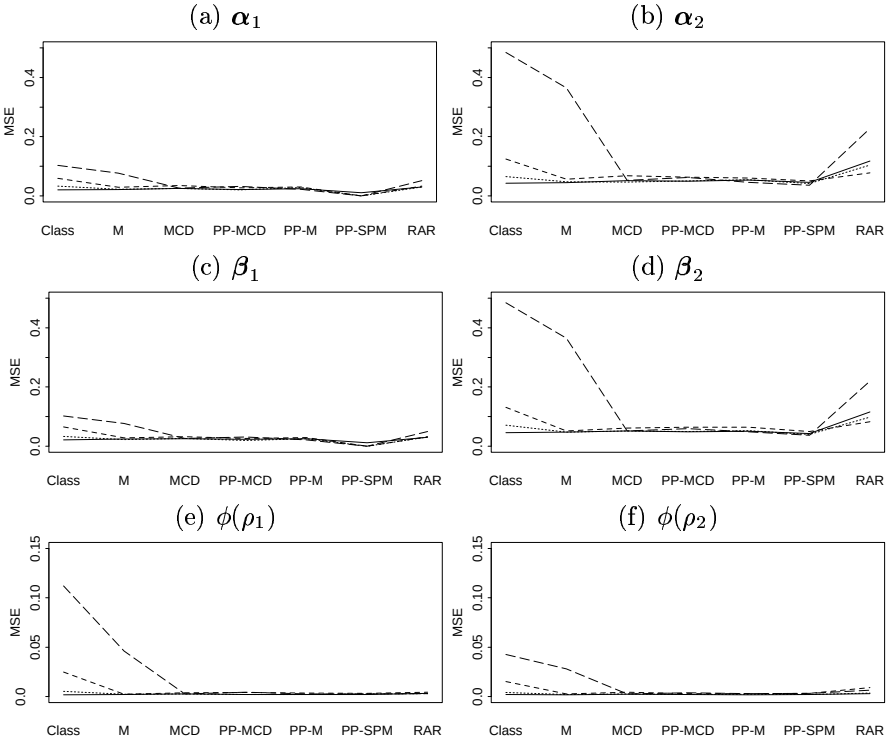

Figure 1: Mean squared error for canonical correlations and vectors, for 7 different estimation procedures and under 4 different sampling schemes for $p = 2$, $q = 2$. The lines correspond to the sampling schemes, as indicated in the legend.

For the practitioner it might be of interest whether the different methods overestimate or underestimate the true correlation, and therefore we also

computed the bias. The largest bias was observed for the classical method with asymmetric contamination with a value of 0.33 for the first transformed correlation and 0.2 for the second. In all other cases the absolute value of the bias was smaller than 0.01 for $\phi(\rho_1)$ and smaller than 0.03 for $\phi(\rho_2)$. An exception was RAR for the second transformed correlation where the bias for the different contamination schemes was somewhat higher with values in the range 0.03-0.08 (we observed a similar effect for simulations with larger dimensions p and q).

Since the number of simulations was rather low we also computed the standard errors of the simulated MSEs, hereby giving an idea about the precision. For example, the largest value of the standard error of $\phi(\rho_1)$ was 0.005, and for $\phi(\rho_2)$ lower than 0.002. Also for the following simulations with higher dimensions p and q the standard errors around the simulated MSEs were comparably low.

For the dimensions $p = 2$ and $q = 4$ we obtained very similar results and therefore no graphs are presented. For $p = 4$ and $q = 4$ resulting MSEs for the canonical vectors $\boldsymbol{\alpha}_1$ to $\boldsymbol{\alpha}_4$ are shown in Figure 2(a)-(d). The pictures for the canonical vectors $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_4$ are very similar and hence not shown. Figure 2(e)-(h) presents the results for the first to the fourth transformed canonical correlation. In general, asymmetric contamination leads to the largest MSEs, followed by T3 and SCN. As before, the classical method performs worst, followed by the method based on the M estimator. The increase for RAR at higher order canonical vectors is again visible. Very good results are obtained with the projection pursuit method based on the Spearman correlation. In fact, the resulting MSEs for the canonical vectors are lower for PP-SPM than for all other considered methods under all four contamination schemes.

Finally, note that the classical method is the most precise under the NOR sampling scheme. It looses quickly this optimality property when deviating from the normal model.

Breakdown Plots

Another feature that is worthwhile to study is the sensitivity of the proposed estimators to increasing amounts of contamination. With this in mind, a simulation study was carried out, where each of the two groups of variables has 3 variables ($p = q = 3$) and samples were generated from a normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$, with $\boldsymbol{\Sigma}_{xx} = \boldsymbol{I}_3$ and $\boldsymbol{\Sigma}_{yy} = \boldsymbol{I}_3$ and

$$\boldsymbol{\Sigma}_{xy} = \begin{bmatrix} 0.9 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}.$$

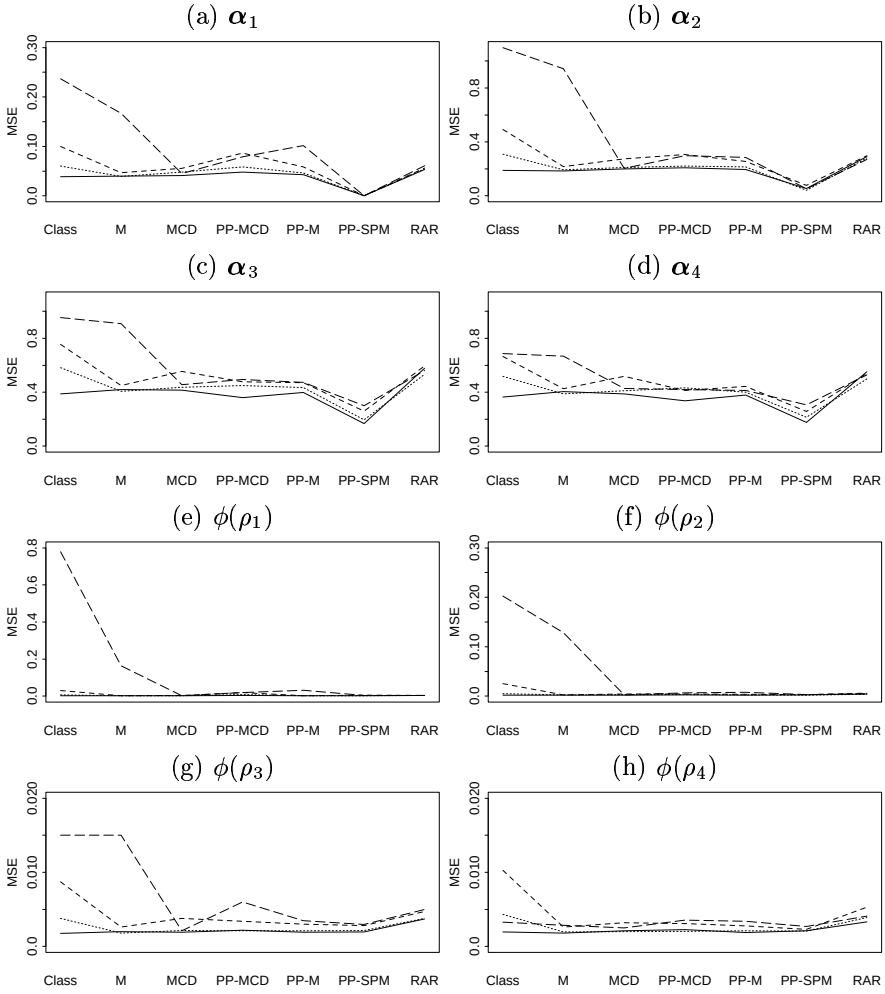

Figure 2: Mean squared error for canonical correlations and vectors, for 7 different estimation procedures and under 4 different sampling schemes for $p = 4$, $q = 4$. The lines correspond to the sampling schemes, as indicated in the legend.

However, ϵ percent of the observations are put equal to a specific outlying position, $c\mathbf{1}^t$, where $c = tr(\mathbf{\Sigma}) = 6$. The values of ϵ were chosen from zero (no contamination) to 25 (25% of contamination). As before we chose $n = 500$. For every value of $\epsilon \in \{0, 1, \ldots, 25\}$, mean squared errors were computed as before, now over 200 simulations.

The results are summarized in Figures 3 and 4. On the horizontal axes we have the percentage of contamination ϵ, the vertical axes presents the mean squared errors. Different lines correspond now to different estimators. The plots can be called breakdown plots, since they indicate how resistant an estimation procedure is under increasing percentages of contamination. Figure 3 shows the resistance of the MSE of the canonical vectors $\boldsymbol{\alpha}_1$ to $\boldsymbol{\alpha}_3$ for the different methods (see legend at the bottom right of the figure). Since the results for $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_3$ were very similar, they are not presented. It is clearly visible that the MSE of the classical method is rapidly increasing in presence of contamination. The classical method is very sensitive with respect to outlying observations, and the results confirm the behavior in Figures 1 and 2. Like in the previous simulation we can see the non-robustness of the method based on the M estimator, even in presence of relatively small amounts of contamination. The projection pursuit method based on the M estimator is more stable, the MSE remains small up to about 12% of contamination, but then it also goes up. The MSE for the robust alternating regression procedure (RAR) is steadily increasing for increasing contamination, showing again that it cannot compete with PP-based methods. The PP method using the MCD estimator is less stable than the procedure based on the full MCD estimate of the covariance matrix. The latter method has a low MSE for up to 20% of contaminated data. The clear favorite when looking at the breakdown plots for the canonical vectors is the projection pursuit method based on the Spearman correlation (PP-SPM). The MSEs for the first order canonical vectors remain very low for the whole considered range of contamination. For $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ (and hence also for $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$) we observe very low MSEs which increase only after adding more than 20% of contamination.

Figure 4 shows the breakdown plots for the canonical correlations. Obviously, the MSEs are getting smaller in general for higher order canonical correlations (see different scales of the vertical axes). The increase of the MSEs already at small amounts of contamination for the classical method and the method based on the M estimator resembles the breakdown plots of Figure 3. A rather unexpected behavior shows PP-M: For the first canonical correlation the MSE remains very small for the complete contamination range, for higher order canonical correlations we observe an increase at about 10% contamination.

While the results of the RAR method were not among the best for the canonical vectors (Figure 3), they are better for the canonical correlations. Also PP-MCD gives low MSEs for small and moderate contamination. Similar to the results of Figure 3, the MCD-based method is very stable up to 20% of contamination and turns out to give, on the whole, the best for the breakdown plot of the canonical correlations. Finally, PP-SPM shows again a very good behavior for 0-20% contamination, and is only slightly worse than MCD.

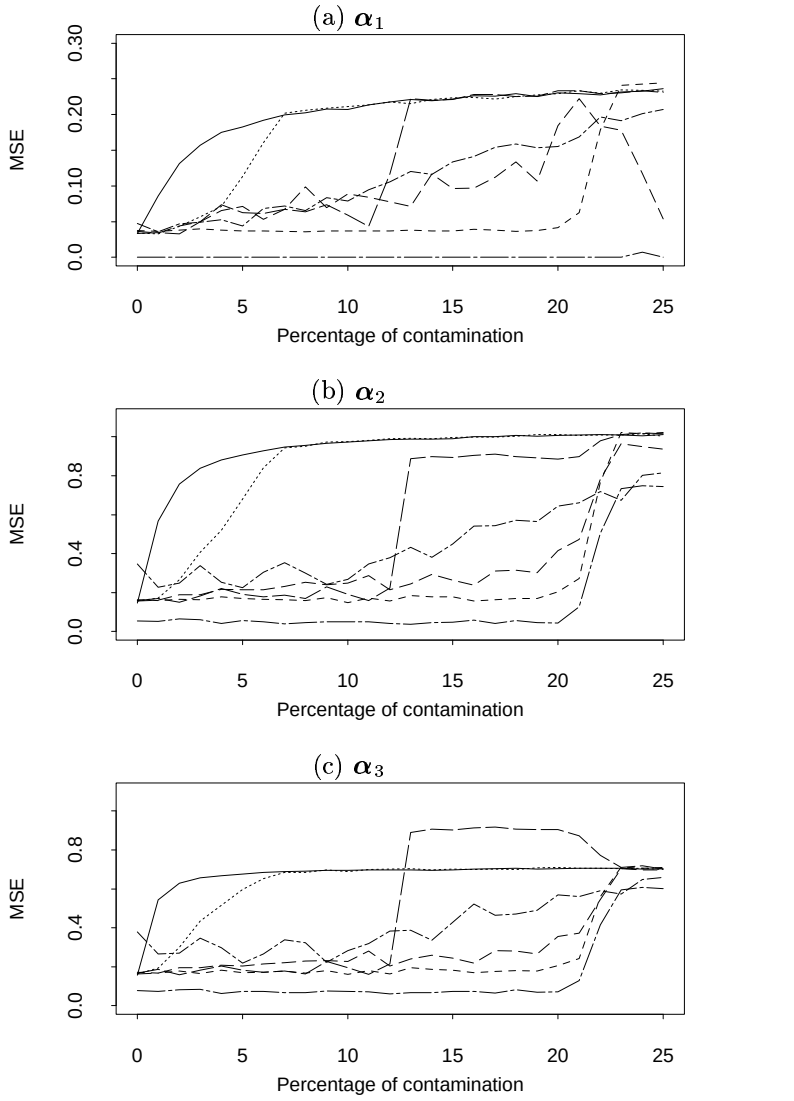

Figure 3: Breakdown plot: Mean squared error for canonical vectors as a function of the percentage of contamination, ranging from 0 to 25%. The lines correspond to different estimation methods, as indicated in the legend.

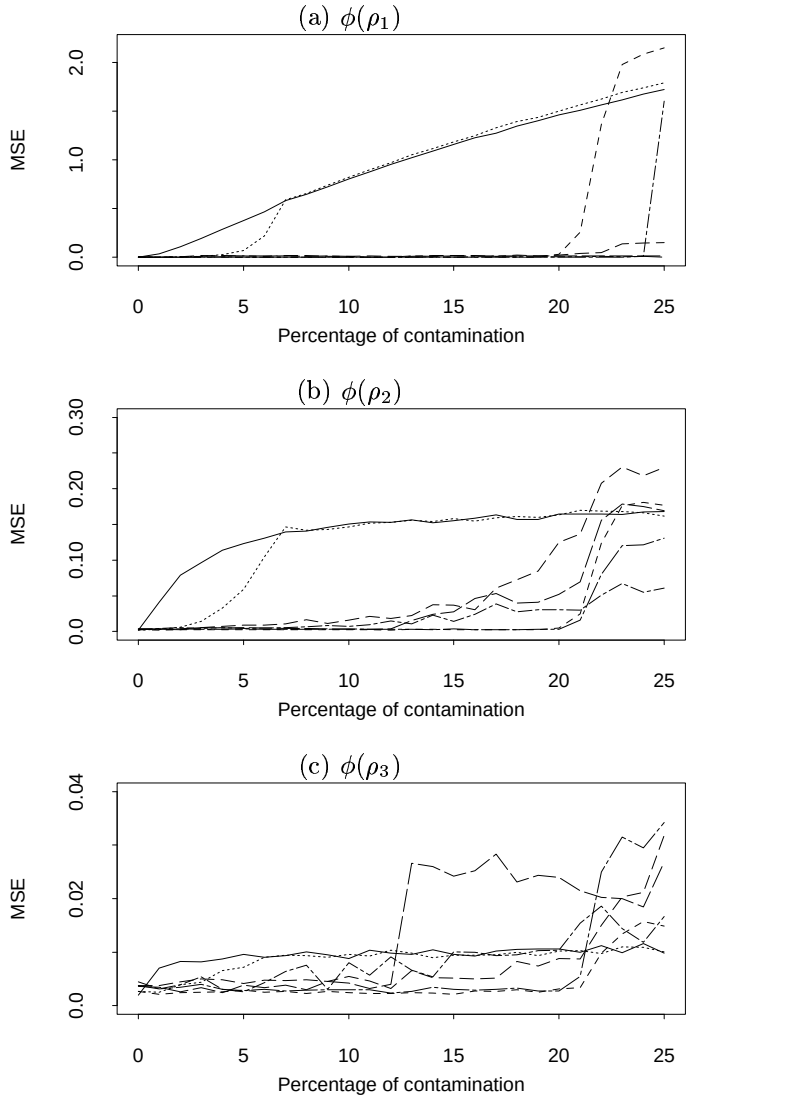

Figure 4: Breakdown plot: Mean squared error for canonical correlations as a function of the percentage of contamination, ranging from 0 to 25%. The lines correspond to different estimation methods, as indicated in the legend.

Test of Independence

The study of associations between two groups of variables only makes sense when these groups are not independent. Assuming that $(\boldsymbol{x}, \boldsymbol{y})$ is multivariate normally distributed, the hypothesis of independence can be formulated as $H_0 : \boldsymbol{\Sigma}_{xy} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\Sigma}_{xy} \neq \mathbf{0}$. If H_0 holds then all the canonical correlations are equal to zero, thus the hypothesis of independence is equivalent to $H_0 : \rho_1 = \ldots = \rho_p = 0$. Bartlett proved that under the null hypothesis,

$$T_0 = -\left(n - \frac{1}{2}(p+q+3)\right) \ln \Lambda \overset{a}{\sim} \chi^2_{pq}, \tag{4.3}$$

where $\Lambda = \prod_{i=1}^{p} \left(1 - \hat{\rho}_i^2\right)$ (see Rencher, 1998). We will also consider other tests of independence like the tests of Lawley-Hotelling, Lawley, and F approximation (see Rencher, 1998).

In order to study the effect of atypical observations in these tests of independence a small simulation study was performed. We consider a situation where the two groups of variables are independent and compute the frequency of rejecting the null hypothesis at the 5% significance level and using the critical value obtained from the asymptotic distribution in (4.3). We consider $p = q = 2$, $\boldsymbol{\Sigma}_{xx} = \boldsymbol{\Sigma}_{yy} = \boldsymbol{I}_2$ and $\boldsymbol{\Sigma}_{xy} = Diag\{0.05, 0.01\}$. In this case, the population canonical correlations are $\rho_1 = 0.05$ and $\rho_2 = 0.01$. We generate data from the sample distributions NOR, SCN and ACN. The estimation methods considered are the classical estimator (Class), the projection pursuit estimator that previously led to the better results (PP-SPM) and the robust estimates based on alternating regression (RAR). The simulation study was done for $m = 1000$ replications. We calculated the p-values associated with the 4 tests (Bartlett, Lawley, Lawley-Hotelling and F) using the classical and the robust estimates.

Table 2: Percentage of rejecting the null hypothesis in 1000 simulations for the Bartlett test.

	Class	PP-SPM	RAR
NOR	0.14	0.23	0.24
SCN	0.50	0.24	0.23
ACN	1.00	0.18	0.22

The results obtained for the 4 tests are very similar, thus only the results for the Bartlett test are presented (see Table 2). As expected, the test with the classical estimates gives good results for normally distributed data and is very sensitive to contamination. In the case of ACN the test rejects in all 1000 simulations. When we plug-in the robust estimates in the considered

test statistics, the percentage of rejecting the null hypothesis is stable even in the case of heavy contamination.

Although these results seem to be quite promising, further studies have to be done in order to evaluate the asymptotic distribution of the null hypothesis when robust estimators are used (see Taskinen et al., 2004).

5 Conclusions

Several methods for robust canonical correlation analysis are available. The most commonly used is to robustly estimate the joint covariance matrix of $\boldsymbol{x}$ and $\boldsymbol{y}$ and perform the usual eigenvalue analysis. Here we considered two estimators of the joint covariance matrix: the M estimator and the MCD estimator. The latter is often preferred due to its high breakdown point. The simulations clearly indicated that the MCD estimator is to be preferred, even for relatively small levels of contamination.

Another type of robustified CCA is based on projection pursuit, and has been discussed in this paper. It is inspired on the initial definition of CCA, namely maximizing a bivariate correlation between linear combinations of both random variables, taking orthogonality restrictions into account. Here we considered three different possibilities for the projection index: a bivariate M or MCD correlation estimator, and Spearman's rank correlation. The projection pursuit method based on the Spearman correlation clearly leads to smaller MSEs. The breakdown plots confirmed that PP-SPM is the most preferable projection pursuit estimator considered here.

The third method for robustifying CCA is based on the alternating regression technique, as proposed by Wold (1966). A clear outline of the Least Squares and the robust alternating regression scheme has been presented. The RAR algorithm seems to be rather complicated, especially for higher order variates, but it is very flexible with respect to the statistical method and model. In all simulation experiments it was observed that this method is robust, but gives on the whole less good results than the best of its competitors.

Besides robustness and efficiency of an estimation procedure, also computation time needs to be taken into account. Figure 5 shows how the methods behave with respect to computation time. Median computation times for one canonical analysis on data being generated according to the sampling schemes considered in the simulation study are presented. Computation times for Class, M, and MCD are not given since they are substantially lower than for the other procedures. It is obvious that PP-MCD and PP-M are time consuming. Using the Spearman correlation projection index results in a substantially lower computing time. Notice that the computing time for the RAR algorithm also remains very reasonable. Moreover, the RAR algorithm

has the advantage that if one only wants to have the first k canonical variates, then the algorithm can be stopped after step k, hereby reducing computing time (the same remark applies in fact for the PP-based methods). Figure 5 shows that computation time for PP-SPM and RAR stays clearly below 1 minute, for all considered sampling schemes. Hence applying robust CCA on a single data set poses no problem, but repeated application of robust CCA in a simulation study leads to a considerable total computation time. This explains why the number of simulation runs in Section 4 was rather limited.

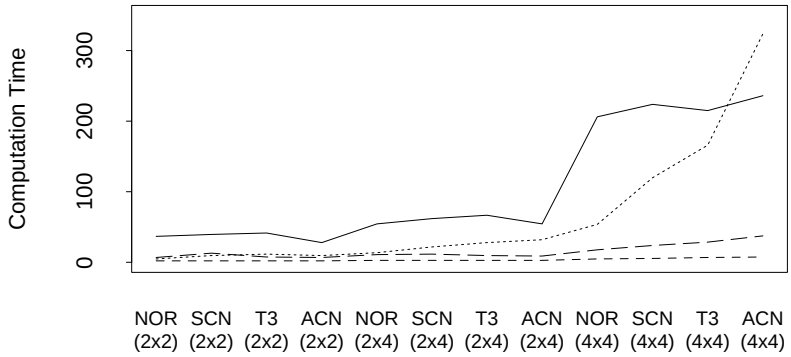

Figure 5: Median of the computation times, measured in seconds, for different estimation procedures. The lines correspond to the estimator, as indicated in the legend. The horizontal axis shows the various sampling schemes and the number of variables ($p \times q$) under study.

Choosing between the studied estimation procedures is difficult. The projection pursuit method based on the Spearman index turned out, somehow surprisingly, to be a highly robust estimator, with good efficiency properties in presence and absence of contamination. If fast computation is an important issue, than the more simple approach based on the MCD-estimator could be advised.

6 Appendix

Least Squares Alternating Regression Scheme (using the notations of Section 3):

Step 1: $\boldsymbol{X}_0 = \boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^t$, $\boldsymbol{Y}_0 = \boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^t$

Step 2: For $l = 1, \ldots, k$:

Step 2.1: Residual spaces (only if $l > 1$):

$$\boldsymbol{X}_{l-1} = \left(\boldsymbol{I}_n - \frac{\boldsymbol{u}_{l-1}\boldsymbol{u}_{l-1}^t}{\boldsymbol{u}_{l-1}^t\boldsymbol{u}_{l-1}}\right)\boldsymbol{X}_{l-2}$$

$$\boldsymbol{Y}_{l-1} = \left(\boldsymbol{I}_n - \frac{\boldsymbol{v}_{l-1}\boldsymbol{v}_{l-1}^t}{\boldsymbol{v}_{l-1}^t\boldsymbol{v}_{l-1}}\right)\boldsymbol{Y}_{l-2}$$

Step 2.2: Starting values (using first principal component $\boldsymbol{z}_1^{l-1}$ of $\boldsymbol{X}_{l-1}$):

$$\hat{\boldsymbol{b}}_l^{(0)} = (\boldsymbol{Y}_{l-1}^t\boldsymbol{Y}_{l-1})^-\boldsymbol{Y}_{l-1}^t\boldsymbol{z}_1^{l-1}$$

$$\boldsymbol{\beta}_l^{(0)} = \frac{\hat{\boldsymbol{b}}_l^{(0)}}{\|\hat{\boldsymbol{b}}_l^{(0)}\|}$$

$$\boldsymbol{v}_l^{(0)} = \boldsymbol{Y}_{l-1}\boldsymbol{\beta}_l^{(0)}$$

Step 2.3: From iteration $s = 1$ to convergence:

$$\hat{\boldsymbol{a}}_l^{(s)} = (\boldsymbol{X}_{l-1}^t\boldsymbol{X}_{l-1})^-\boldsymbol{X}_{l-1}^t\boldsymbol{v}_l^{(s-1)}$$

$$\boldsymbol{\alpha}_l^{(s)} = \frac{\hat{\boldsymbol{a}}_l^{(s)}}{\|\hat{\boldsymbol{a}}_l^{(s)}\|}$$

$$\boldsymbol{u}_l^{(s)} = \boldsymbol{X}_{l-1}\boldsymbol{\alpha}_l^{(s)}$$

$$\hat{\boldsymbol{b}}_l^{(s)} = (\boldsymbol{Y}_{l-1}^t\boldsymbol{Y}_{l-1})^-\boldsymbol{Y}_{l-1}^t\boldsymbol{u}_l^{(s)}$$

$$\boldsymbol{\beta}_l^{(s)} = \frac{\hat{\boldsymbol{b}}_l^{(s)}}{\|\hat{\boldsymbol{b}}_l^{(s)}\|}$$

$$\boldsymbol{v}_l^{(s)} = \boldsymbol{Y}_{l-1}\boldsymbol{\beta}_l^{(s)}$$

Step 2.4: After convergence, resulting in $\boldsymbol{u}_l^*$, $\boldsymbol{v}_l^*$, $\boldsymbol{\alpha}_l^*$, $\boldsymbol{\beta}_1^*$:

$$|r_l = \mathrm{Corr}(\boldsymbol{u}_l^*, \boldsymbol{v}_l^*)|$$

Step 2.4.1: If $l = 1$:

$$\boldsymbol{u}_1 = \boldsymbol{u}_1^*,\ \boldsymbol{v}_1 = \boldsymbol{v}_1^*,\ \hat{\boldsymbol{\alpha}}_1 = \boldsymbol{\alpha}_1^*,\ \hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1^*$$

Step 2.4.2: If $l > 1$:

$$\boldsymbol{u}_l = \boldsymbol{u}_l^*$$

$$\hat{\boldsymbol{\alpha}}_l = (\boldsymbol{X}_0^t\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^t\boldsymbol{u}_l$$

$$\boldsymbol{v}_l = \boldsymbol{v}_l^*$$

$$\hat{\boldsymbol{\beta}}_l = (\boldsymbol{Y}_0^t\boldsymbol{Y}_0)^{-1}\boldsymbol{Y}_0^t\boldsymbol{v}_l$$

Robust Alternating Regression Scheme (using the notations of Section 3):

Step 1: $\boldsymbol{X}_0 = \boldsymbol{X} - \mathbf{1}\tilde{\boldsymbol{x}}^t$, $\boldsymbol{Y}_0 = \boldsymbol{Y} - \mathbf{1}\tilde{\boldsymbol{y}}^t$

$\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{y}}$ are the column-wise medians of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

Step 2: For $l = 1, \ldots, k$:

Step 2.1: Residual spaces (only if $l > 1$):

$\boldsymbol{X}_{l-1}$ are the estimated residuals of $\boldsymbol{X}_{l-2} = \boldsymbol{u}_{l-1}\boldsymbol{c}^t + \boldsymbol{\varepsilon}_1$ using weighted L_1 regressions with weights $w_i(\boldsymbol{u}_{l-1})$

$\boldsymbol{Y}_{l-1}$ are the estimated residuals of $\boldsymbol{Y}_{l-2} = \boldsymbol{v}_{l-1}\boldsymbol{d}^t + \boldsymbol{\varepsilon}_2$ using weighted L_1 regressions with weights $w_i(\boldsymbol{v}_{l-1})$

Step 2.2: Starting values:

Compute the first robust principal component $\boldsymbol{z}_1^{l-1}$ of $\boldsymbol{X}_{l-1}$
using the algorithm of Croux and Ruiz-Gazen (1996)
$\hat{\boldsymbol{b}}_l^{(0)}$ are the estimated coefficients of $\boldsymbol{z}_1^{l-1} = \boldsymbol{Y}_{l-1}\boldsymbol{b}_l^{(0)} + \varepsilon_3$
using weighted L_1 regression with weights $w_i(\boldsymbol{Y}_{l-1}^*)$
$\boldsymbol{\beta}_l^{(0)} = \frac{\hat{\boldsymbol{b}}_l^{(0)}}{\|\hat{\boldsymbol{b}}_l^{(0)}\|}$
$\boldsymbol{v}_l^{(0)} = \boldsymbol{Y}_{l-1}\boldsymbol{\beta}_l^{(0)}$

Step 2.3: From iteration $s = 1$ upto convergence:
$\hat{\boldsymbol{a}}_l^{(s)}$ are the estimated coefficients of $\boldsymbol{v}_l^{s-1} = \boldsymbol{X}_{l-1}\boldsymbol{a}_l^{(s)} + \varepsilon_4$
using weighted L_1 regression with weights $w_i(\boldsymbol{X}_{l-1}^*)$
$\boldsymbol{\alpha}_l^{(s)} = \frac{\hat{\boldsymbol{a}}_l^{(s)}}{\|\hat{\boldsymbol{a}}_l^{(s)}\|}$
$\boldsymbol{u}_l^{(s)} = \boldsymbol{X}_{l-1}\boldsymbol{\alpha}_l^{(s)}$
$\hat{\boldsymbol{b}}_l^{(s)}$ are the estimated coefficients of $\boldsymbol{u}_l^{s-1} = \boldsymbol{Y}_{l-1}\boldsymbol{b}_l^{(s)} + \varepsilon_5$
using weighted L_1 regression with weights $w_i(\boldsymbol{Y}_{l-1}^*)$
$\boldsymbol{\beta}_l^{(s)} = \frac{\hat{\boldsymbol{b}}_l^{(s)}}{\|\hat{\boldsymbol{b}}_l^{(s)}\|}$
$\boldsymbol{v}_l^{(s)} = \boldsymbol{Y}_{l-1}\boldsymbol{\beta}_l^{(s)}$

Step 2.4: After convergence, resulting in $\boldsymbol{u}_l^*$, $\boldsymbol{v}_l^*$, $\boldsymbol{\alpha}_l^*$, $\boldsymbol{\beta}_1^*$:
$r_l = \mathrm{Corr}(\boldsymbol{u}_l^*, \boldsymbol{v}_l^*)$; Corr is a robust correlation measure like
the bivariate MCD correlation discussed in Section 2

Step 2.4.1: If $l = 1$:
$\boldsymbol{u}_1 = \boldsymbol{u}_1^*$, $\boldsymbol{v}_1 = \boldsymbol{v}_1^*$, $\hat{\boldsymbol{\alpha}}_1 = \boldsymbol{\alpha}_1^*$, $\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1^*$

Step 2.4.2: If $l > 1$:
$\boldsymbol{U}_{l-1} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{l-1}]$
$\tilde{\boldsymbol{u}}_l$ are the estimated residuals of $\boldsymbol{u}_l^* = \boldsymbol{U}_{l-1}\boldsymbol{e} + \varepsilon_6$
using robust LTS regression
$\hat{\boldsymbol{\alpha}}_l$ are the estimated coefficients of $\tilde{\boldsymbol{u}}_l = \boldsymbol{X}_0\boldsymbol{f} + \varepsilon_7$
using robust LTS regression
$\boldsymbol{u}_l = \boldsymbol{X}_0\hat{\boldsymbol{\alpha}}_l$
$\boldsymbol{V}_{l-1} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{l-1}]$
$\tilde{\boldsymbol{v}}_l$ are the estimated residuals of $\boldsymbol{v}_l^* = \boldsymbol{V}_{l-1}\boldsymbol{g} + \varepsilon_8$
using robust LTS regression
$\hat{\boldsymbol{\beta}}_l$ are the estimated coefficients of $\tilde{\boldsymbol{v}}_l = \boldsymbol{X}_0\boldsymbol{h} + \varepsilon_9$
using robust LTS regression
$\boldsymbol{v}_l = \boldsymbol{Y}_0\hat{\boldsymbol{\beta}}_l$

Acknowledgements

The authors are grateful for helpful comments of two anonymous referees.

7 References

Becker, C. and Gather, U. (2001). The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistics and Data Analysis*, 36, 119-127.

Croux, C., and Dehon, C. (2002). Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. *La Revue de Statistique Appliquée*, 2, 5-26.

Croux, C., Filzmoser, P., Pison, G., and Rousseeuw, P.J. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 13, 23-36.

Croux, C., and Ruiz-Gazen, A. (1996). A fast algorithm for robust principal components based on projection pursuit. In: A. Prat (ed.), *COMPSTAT: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pp. 211-216.

Das, S. and Sen, P.K. (1998). Canonical correlations. In: P. Armitage and T. Colton (eds.), *Encyclopedia of Biostatistics, Vol. 1*, Wiley, New York, pp. 468-482.

Filzmoser, P., Dehon, C., and Croux, C. (2000). Outlier resistant estimators for canonical correlation analysis. In: J.G. Betlehem and P.G.M. van der Heijden (eds.), *COMPSTAT: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pp. 301-306.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321-377.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Huber, P.J. (1985). Projection pursuit. *The Annals of Statistics*, 13, 435-525.

Johnson, R. A., and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*. Prentice-Hall, London.

Karnel, G. (1991). Robust canonical correlation and correspondence analysis. In: *The Frontiers of Statistical Scientific and Industrial Applications*, (Volume II of the proceedings of ICOSCO-I, The First International Conference on Statistical Computing), American Sciences Press, Strassbourg, pp. 335-354.

Lyttkens, E. (1972). Regression aspects of canonical correlation. *Journal of Multivariate Analysis*, 2, 418-439.

Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4, 51-67.

Oliveira, M.R., and Branco, J.A. (2000). Projection pursuit approach to robust canonical correlation analysis. In: J.G. Betlehem and P.G.M. van der Heijden (eds.), *COMPSTAT: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, pp. 415-420.

Rencher, A.C. (1998). *Multivariate Statistical Inference and Applications*, John Wiley, New York.

Romanazzi, M. (1992). Influence in canonical correlation analysis. *Psychometrika*, 57, 237-259.

Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-880.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In: W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (eds.), *Mathematical Statistics and Applications, Vol. B*, Reidel, Dordrecht, pp. 283-297.

Rousseeuw, P.J., and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212-223.

Rousseeuw, P.J. and Van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633-651.

Taskinen, S., Croux, C., Kankainen, A., Ollila, E., and Oja, H. (2003). Canonical Analysis based on Scatter Matrices. Manuscript.

Taskinen, S., Oja, H., and Randles, R.H. (2004). Multivariate nonparametric tests of independence. Manuscript, conditionally accepted.

Tenenhaus, M. (1998). *La Régression PLS. Théorie et pratique*, Éditions Technip, Paris.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. In: F.N. David (ed.), *A Festschrift for J. Neyman*, Wiley, New York, pp. 411-444.