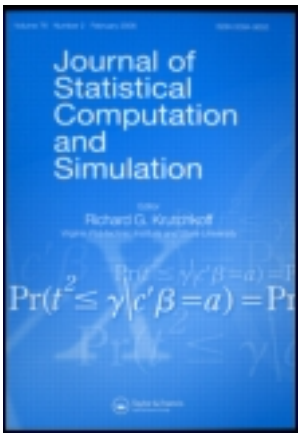

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

A comparative study for robust canonical correlation methods

Ali Alkenani [a] & Keming Yu [a]

[a] School of Information System, Computing and Mathematics, Mathematical Science Department, Brunel University, Kingston Lane, Uxbridge, Middlesex, UB8 3PH, London, UK

Published online: 21 Nov 2011.

To cite this article: Ali Alkenani & Keming Yu (2013): A comparative study for robust canonical correlation methods, Journal of Statistical Computation and Simulation, 83:4, 690-718

To link to this article: http://dx.doi.org/10.1080/00949655.2011.632775

Journal of Statistical Computation and Simulation, 2013
Vol. 83, No. 4, 690–718, http://dx.doi.org/10.1080/00949655.2011.632775

Taylor & Francis
Taylor & Francis Group

A comparative study for robust canonical correlation methods

Ali Alkenani* and Keming Yu

School of Information System, Computing and Mathematics, Mathematical Science Department, Brunel University, Kingston Lane, Uxbridge, Middlesex UB8 3PH, London, UK

(*Received 18 August 2011; final version received 13 October 2011*)

The aim of this study is to obtain robust canonical vectors and correlation coefficients based on the percentage bend correlation and winsorized correlation in the correlation matrix and fast consistent high breakdown (FCH), reweighted fast consistent high breakdown (RFCH), and reweighted multivariate normal (RMVN) estimators to estimate the covariance matrix and then compare these estimators with the existing estimators. In the correlation matrix of canonical correlation analysis (CCA), we present an approach that substitutes the percentage bend correlation and the winsorized correlation in place of the widely employed the Pearson correlation. Moreover, we employ the FCH, RFCH, and RMVN estimators to estimate the covariance matrix in the CCA. We conduct a simulation study and employ real data with the objective of comparing the performance of the different estimators for canonical vectors and correlation with that of our proposed approaches. The breakdown plots and independent tests are employed as differentiating criteria of the robustness and performance of the estimators. Based on our computational and real data studies, we propose suggestions and guidelines on the practical implications of our findings.

Keywords: dimension reduction; robust canonical correlation

1. Introduction

Canonical correlation analysis (CCA), originally proposed by Hotelling [1], is a method used for measuring the linear relationship between two multidimensional variables. This method can be seen as the problem of finding basis vectors for two sets of variables such that the correlations between the projections of the variables into these basis vectors are mutually maximized.

Suppose that X is a p-dimensional random variable and Y is a q-dimensional random variable, with $p \leq q$. Furthermore, suppose that X and Y have the covariance matrix (if it exists)

$$\sum = \begin{pmatrix} \sum_{XX} & \sum_{XY} \\ \sum_{YX} & \sum_{YY} \end{pmatrix}, \tag{1}$$

where $\sum_{XX}$ and $\sum_{YY}$ are non-singular. The objective of the CCA is to study the linear relationship between X and Y as measured by the correlation between the linear combination of both sets of

*Corresponding author. Email: mapgaja@brunel.ac.uk

variables. Specifically, we look for

$$(\alpha_1, \beta_1) = \text{argmax}_{a,b}\, \text{Corr}(a^t X, b^t Y), \tag{2}$$

where Corr is the Pearson correlation and the vectors $\alpha_1 \in \mathrm{R}^p$ and $\beta_1 \in \mathrm{R}^q$ are the resulting first pair of canonical vectors. The linear combinations $U_1 = \alpha_1^t X$ and $V_1 = \beta_1^t Y$ are called the first pair of canonical variates. Note that according to Equation (2), the vectors α_1 and β_1 are only determined up to a multiple, and in order to identify them uniquely (up to a sign), the maximizing linear combinations are normalized as $\text{Var}(\alpha_1^t X) = \text{Var}(\beta_1^t Y) = 1$.

While the first canonical vectors are useful, they do not capture the complete dependency structure between X and Y. To this end, higher order canonical vectors defined for $k = 2, 3, \ldots, p$ as

$$(\alpha_k, \beta_k) = \text{argmax}_{a,b}\, \text{Corr}(a^t X, b^t Y) \tag{3}$$

are used where the pairs of canonical variates of order k are $U_k = \alpha_k^t X$ and $V_k = \beta_k^t Y$ and

$$\text{Cov}(U_k, U_j) = \alpha_k^t \sum_{XX} \alpha_j = \text{Cov}(V_k, V_j) = \beta_k^t \sum_{XX} \beta_j = \begin{cases} 1 & \text{if } k = j, \\ 0 & \text{if } 1 \le j < k. \end{cases} \tag{4}$$

The correlation ρ_k between the canonical variates of the kth pair, $\rho_k = \text{Corr}(U_k, V_k)$, is the kth canonical correlation. Moreover, the canonical vectors α_k and α_k are the eigenvectors corresponding to the eigenvalues $\rho_1^2 \ge \cdots \ge \rho_p^2 > 0$ of the matrices

$$\sum_A = \sum_{XX}^{-1} \sum_{XY} \sum_{YY}^{-1} \sum_{YX} \text{ and } \sum_B = \sum_{YY}^{-1} \sum_{YX} \sum_{XX}^{-1} \sum_{XY} \tag{5}$$

or

$$R_A = R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX} \text{ and } R_E = R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY}, \tag{6}$$

where $\mathbf{R} = \begin{pmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{pmatrix}$ is the correlation matrix. The matrices in Equations (5) and (6) have the same eigenvalues, ρ_k^2, which correspond to the squared canonical correlations.

A major attraction of the CCA is its application for dimension reduction and thus it acts as a valuable tool that facilitates the understanding of complex relationships among sets of variables [2].

To estimate the population canonical correlations and canonical vectors, we first estimate $\sum$ by the sample covariance matrix followed by the computation of the eigenvalues and eigenvectors of the matrices $\sum_A$ and $\sum_B$ as given by Equation (5). This procedure is best when X and Y are from a multivariate normal distribution; however, it appears to be less efficient with respect to outlying observations. From a practical point of view, it is well known that the sample covariance matrix is not resistant to outliers and thus a canonical analysis based on this matrix will result in uncertain and misleading results. Similarly, Romanazzi [3] showed that the classical canonical vectors and correlations are also sensitive to outlying observations. Consequently, in order to obtain accuracy and robustness, there is a need to estimate the population covariance matrix using robust approaches.

An apparent approach to 'robustify' canonical correlation is to estimate sample covariance or correlation matrix using methods that can account for outliers. One such approach was presented by Karnel [4], who considered M-estimators of multivariate location and scatter as robust estimators of $\sum$ and then followed the classical approach. However, the robustness properties of the M-estimators are poor in high dimensions.

There are many robust estimators for robust multivariate location and dispersion. The fastest estimator of multivariate location and dispersion that has been shown to be both consistent and having a high breakdown point is the minimum covariance determinant (MCD) estimator with $O(n^v)$ complexity, where $v = 1 + (p(p + 3)/2)$ [5]. The complexity of the minimum volume ellipsoid (MVE) estimator is far higher, and there may be no known method for computing S, τ, projection-based, constrained M, M-estimate of the scale of the residuals and M-estimate of the parameters, and Stahel–Donoho estimators [6].

Since the above estimators are computationally time consuming, they have been replaced by practical estimators which strike a balance between accuracy and computing cost. However, none of the practical estimators have been shown to be consistent and having a high breakdown point. For example, the Rousseeuw and Van Driessen [7] fast minimum covariance determinant (FMCD) estimator is used to replace the MCD estimator. So, the robust multivariate techniques (one of which is the robust canonical correlation) that claim to use the impractical MCD estimator actually use the Rousseeuw and Van Driessen [7] FMCD estimator.

Croux and Dehon [8] used the FMCD estimator (see [7,9] for a fast algorithm). Taskinen *et al.* [10] obtained the influence function and asymptotic distributional properties for CCA based on robust estimates of the covariance matrix. Following the approach suggested by Wold [11], Filzmoser *et al.* [12] devised a robust method for obtaining the first canonical variates using robust alternating regressions (RARs).

Branco *et al.* [13] compared and discussed a number of approaches for robust canonical correlation analysis (RCCA), and they proposed a robust method for obtaining all the canonical variates using RARs. They suggested that the canonical correlation estimators based on the FMCD estimator for the covariance matrix are often preferred due to their high breakdown point. Furthermore, the simulations which they conducted clearly indicate that the FMCD estimator is preferable, even for relatively small levels of contamination.

Jiao and Jian [14] studied the association between two sets of random variables based on the projection pursuit (pp) method, and they derived the asymptotic normal distributions of estimators of the pp based on canonical correlations and canonical vectors. Recently, Kudraszow and Maronna [15] proposed a method for the RCCA based on the prediction approach.

Olive and Hawkins [6] showed that the FMCD estimator is not a high breakdown estimator. They proposed practical $\sqrt{n}$ consistent outlier resistant estimators for multivariate location and dispersion. They suggested that the fast consistent high breakdown (FCH) estimator is fast, consistent, and highly outlier resistant and that the reweighted fast consistent high breakdown (RFCH) estimator is the reweighted FCH estimator, and they used the reweighted multivariate normal (RMVN) estimator for CCA, discrimination, factor analysis, principal components, and regression. The RMVN estimator uses a slightly modified method for reweighting such that it gives good estimates of $(\mu, \sum)$ for multivariate normal data, even when certain types of outliers are present.

Zhang and Olive [16] used the RMVN estimator with principle component analysis. They suggested the application of the classical multivariate procedures to the RMVN subset. Zhang [17, Ch. 5] used the RMVN estimator for CCA.

The computational complexity of the FCH, RFCH, and RMVN estimators is $O[p^3 + np^2 + \text{nplog}(n)]$, and these estimators are roughly 100 times faster than the FMCD estimator [18].

Cannon and Hsieh [19] suggested the use of robust variants of nonlinear canonical correlation analysis (NLCCA) to improve performance on data sets with low signal-to-noise ratios. To achieve this, they employed a neural network model architecture of standard NLCCA; however, the cost functions used to set the model parameters were replaced with more robust variants, and in the double-barreled network, the Pearson correlation was replaced with a biweight midcorrelation.

Wilcox [20] studied the percentage bend correlation which is motivated in part by asymptotic results associated with the M-estimators of location and the percentage bend measure of scale studied by Shoemaker and Hettmansperger [21].

Wilcox [22] stated that robust versions of the Pearson correlation are divided into two types. The first type are those that 'robustify' against outliers without taking into account the general structure of the data, whereas the second type take into account the general structure of the data when dealing with outliers. In the literature, the first and second types are, respectively, referred to as the M correlation and the O correlation. Moreover, Wilcox [22] described the four types of M correlations as the percentage bend correlation, the biweight midcorrelation, the winsorized correlation, and Kendall's tau correlation. Similarly, he also presented a number of O correlation methods such as the fast minimum volume ellipsoid (FMVE), FMCD, and skipped measures of correlations. The FMVE and FMCD measures employ the central half of the data to estimate location, scatter, covariance, and correlation. For instance, one can simply compute the Pearson correlation based on the central half of the data. Skipped correlations are obtained by detecting the outliers using one of the multivariate outlier detection methods (for details, see [22], Section 6.4) and then removing these outliers and applying some of the correlation coefficients to the remaining data.

To our knowledge, there is no research paper that has focused on replacing the Pearson correlation in the correlation matrix of canonical correlation with the percentage bend correlation and the winsorized correlation. However, Olive and Hawkins [6] suggested using the FCH, RFCH, and RMVN estimators for CCA, discrimination, factor analysis, principal components, and regression, and Zhang [17, CH.5] used the RMVN estimator for CCA, but until now, no research has employed the FCH and RFCH estimators to estimate the covariance matrix in the CCA. To this end, the goal of this paper is to obtain robust canonical vectors and correlation coefficients that depend on percentage bend correlation and the winsorized correlation in the correlation matrix. Furthermore, we aim to employ the FCH and RFCH estimators to estimate the covariance matrix and then compare these estimators with other known estimators.

In this paper, we conduct a comparative study to explore the performance of 13 different estimators for canonical vectors and correlation. Simulation studies are used to compare the numerical performances of the 13 different estimators under different sampling schemes similar to that done in [13]. To assess the robustness of the estimators, we make use of the breakdown plots and apply the test of independence.

In Section 2, 12 different robustifications of CCA are discussed. In Section 3, the different estimators are compared using a simulation study. In Section 4, we use the breakdown plots to study the robustness of the estimators. In Section 5, tests of independence are done for the different estimators. The conclusions are summarized in Section 6.

2. RCCA based on robust correlation coefficients and robust covariance matrix

2.1. *The percentage bend correlation*

Let a special case of Huber's function be defined as

$$\psi(\chi) = \max[-1, \min(1, \chi)].$$

Furthermore, let θ_x and θ_y be the respective population medians for the random variables X and Y and define W_x as the solution of the following equation:

$$P(|X - \theta_x| < W_x) = 1 - \beta. \tag{7}$$

Let Φ_{pbx} and Φ_{pby} denote the percentage bend measure of location for X and Y, respectively. Furthermore, let $U = (X - \Phi_{\text{pbx}})/W_x$ and $V = (Y - \Phi_{\text{pby}})/W_y$ such that $[\psi(U)] = E[\psi(V)] = 0$.

The percentage bend correlation between X and Y is

$$\rho_{\text{pb}} = \frac{E\{\psi(U)\psi(V)\}}{\sqrt{E\{\psi^2(U)\}E\{\psi^2(V)\}}}, \tag{8}$$

where $-1 \leq \rho_{\text{pb}} \leq 1$ and ρ_{pb} is a robust measure of the linear association between X and Y such that the variables X and Y are said to be independent when $\rho_{\text{pb}} = 0$. It can be noted that ρ_{pb} depends in part on the measure of scale, W_x, which is a generalization of median absolute deviation (MAD). Similarly, W_x is a measure of dispersion when $\psi(X) = \max[-1, \min(1, \chi)]$.

The Huber's function is selected to be used in the percentage bend correlation for a number of reasons. First, Huber's function is a monotonic function. Second, Huber's function gives a consistent estimator of location. Third, it is has the convenient feature of a single iteration being sufficient in the applied work. Finally, when $\psi(X) = \max[-1, \min(1, x)]$, the resulting measure of scale is a measure of dispersion [20].

To estimate the percentage bend correlation,

(1) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$, be a random sample. Let M_x be the sample median for the observations $X_1, \ldots, X_n$. Select a value for β, where $0 \leq \beta \leq 0.5$.
(2) Compute $W_i = |X_i - M_x|$ and $m = [(1-\beta)n]$, and let $\hat{w}_x = W_{(m)}$, where $W_{(1)} \leq \cdots \leq W_{(n)}$ are the W_i values written in ascending order.
(3) Compute $S_x = \sum_{i=i_1+1}^{n-i_2} X_{(i)}$, $\Phi_x = (\hat{w}_x(i_2 - i_1) + S_x)/(n - i_1 - i_2)$, where i_1 is the number of X_t values such that $(X_i - M_x)/\hat{w}_x < -1$ and i_2 is the number of X_t values such that $(X_i - M_x)/\hat{w}_x > 1$.
(4) Set $U_i = (X_i - \Phi_x)/\hat{w}_x$. Repeat these computations for the Y_i values, $V_i = (Y_i - \Phi_y)/\hat{w}_y$.
(5) The estimated percentage bend correlation (r_{pb}) between X and Y is

$$r_{\text{pb}} = \frac{\sum A_i B_i}{\sqrt{\sum A_i{}^2 \sum B_i{}^2}}, \tag{9}$$

where $A_i B_i = \psi(U_1)\psi(V_i)$, $B_i = \psi(V_i)$ and $\psi(x) = \max[1-, \min(1, x)]$.

To test the hypothesis

$$H_0 : \rho_{\text{pb}} = 0, \tag{10}$$

when X and Y are independent, we need to compute

$$T_{\text{pb}} = r_{\text{pb}} \sqrt{\frac{n-2}{1 - r_{\text{pb}}^2}}. \tag{11}$$

We reject H_0 if $|T_{\text{pb}}| > t_{1-\alpha}$, the $1-\alpha$ quantile of Student's t distribution with $v = n - 2$ degrees of freedom.

2.2. *The biweight midcorrelation*

Let ψ be any odd function and let μ_x and μ_y be any measure of location for random variables X and Y, respectively. Let τ_x, and τ_y be some measure of scale for random variables $X$and Y, respectively. Let K be some constant and let $U = (X - \mu_x)/(K\tau_x)$ and $V = (Y - \mu_y)/(K\tau_y)$.

Then, a measure of covariance between X and Y is

$$\gamma_{xy} = \frac{nK^2\tau_x\tau_y E\{\psi(U)\psi(V)\}}{E\{\psi(U)\psi(V)\}} \tag{12}$$

with a corresponding measure of correlation given by

$$\rho_b = (\gamma_{xy})/(\sqrt{\gamma_{xx}\gamma_{yy}}) - 1 \leq \rho_b \leq 1. \tag{13}$$

Wilcox [22] chose ψ as the biweight function and $K = 9$, where the biweight function is defined as follows:

$$\psi(x) = \begin{cases} x(1-x^2)^2 & \text{if } |\chi| < 1, \\ 0 & \text{if } |\chi| \geq 1. \end{cases} \tag{14}$$

Let M_x and M_y denote the respective medians calculated from the random samples $(X_1, Y_1), \ldots, (X_n, Y_n)$. Define $U_i = (X_i - M_x)/(9\,\mathrm{MAD}_x)$ and $V_i = (Y_i - M_y)/(9\,\mathrm{MAD}_y)$, then the MAD_x and MAD_y are the values of MAD for the X and Y values.

Let

$$a_i = \begin{cases} 1 & \text{if } -1 \leq U_i \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$b_i = \begin{cases} 1 & \text{if } -1 \leq V_i \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that the sample biweight midcovariance between X and Y is

$$\mathrm{bicov}(x,y) = \frac{n\sum a_i(X_i - M_x)(1-U_i^2)^2 b_i(Y_i - M_y)(1-V_i^2)^2}{(\sum a_i(1-U_i^2)(1-5U_i^2))(\sum b_i(1-V_i^2)(1-5V_i^2))}, \tag{15}$$

and the biweight midcorrelation is then given by

$$r_b = \frac{\mathrm{bicov}(x,y)}{\sqrt{\mathrm{bicov}(x,x)\mathrm{bicov}(y,y)}}. \tag{16}$$

To test the null hypothesis

$$H_0 : \rho_b = 0 \tag{17}$$

when x and y are independent, we need to compute the test statistic

$$T_b = r_b\sqrt{\frac{n-2}{1-r_b^2}}, \tag{18}$$

and we reject H_0 if $|T_b| > t_{1-\alpha/2}$, the $1-\alpha/2$ quantile of Student's t distribution with $v = n-2$.

2.3. *The winsorized correlation*

The population winsorized correlation between two random variables X_1 and X_2 is given by

$$\rho_w = \frac{E\{(X_1 = \mu_{w_1})(X_2 - \mu_{w_2})\}}{\sigma_{w_1}\sigma_{w_2}} = \frac{\sigma_{w12}}{\sigma_{w1}\sigma_{w2}}, \quad -1 \leq \rho_w \leq 1. \tag{19}$$

where σ_{w_j} is the population winsorized standard deviation of X_j and $E_w(X)$ is the winsorized expected value of X.

To estimate ρ_w, based on the random sample $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$, first winsorize the observations by computing the Y_{ij} values as follows:

$$Y_{ij} = \begin{cases} X_{(g+1)_j} & \text{if } X_{ij} \le X_{(g+1)_j}, \\ X_{ij} & \text{if } X_{(g+1)_j} < X_{ij} < X_{(n-g)_j}, \\ X_{(n-g)_j} & \text{if } X_{ij} \ge X_{(n-g)_j}, \end{cases} \tag{20}$$

where g is the number of observations trimmed or Winsorized from each end of the distribution corresponding to the jth group. Then ρ_w is estimated by computing the Pearson's correlation with the Y_{ij} values:

$$r_w = \frac{\sum (Y_{i1} - \bar{Y}_1)(Y_{i1} - \bar{Y}_2)}{\sqrt{\sum (Y_{i1} - \bar{Y}_1)^2 \sum (Y_{i2} - \bar{Y}_2)^2}}. \tag{21}$$

To test the null hypothesis

$$H_0 : \rho_w = 0 \tag{22}$$

we need to compute

$$T_w = r_w \sqrt{\frac{n-2}{1-r_w^2}} \tag{23}$$

and we reject H_0 if $|T_w| > t_{1-\alpha/2}$, $1 - \alpha/2$ quantile of Student's t distribution with $v = h - 2$ degrees of freedom, where h, the effective sample size, is the numbher of pairs of observations that are not winsorized.

2.4. *Kendall's tau correlation*

Kendall's tau correlation is a non-parametric M-type correlation. Because of being resistant to outlying observations, it is often said to be robust. Let two pairs of observations (X_1, Y_1) and (X_2, Y_2) be such that $X_1 < X_2$ and assuming that tied values never occur. If $Y_1 < Y_2$, then (X_1, Y_1) and (X_2, Y_2) will be concordant; otherwise (X_1, Y_1) and (X_2, Y_2) are discordant.

For n pairs of points, let

$$S_{ij} = \begin{cases} 1 & \text{if } i\text{th and } j\text{th are concordant}, \\ -1 & \text{otherwise}. \end{cases}$$

Kendall's tau correlation formula is

$$r_\tau = \frac{2 \sum_{i<j} S_{ij}}{n(n-1)}. \tag{24}$$

Although Kendall's tau correlation provides resistance against outliers, the presence of outliers can substantially change its value.

Under independence, the population Kendall's tau correlation $\rho_\tau = 0$.

To test the null hypothesis

$$H_0 : \rho_\tau = 0 \tag{25}$$

we compute

$$z = \frac{6 \sum_{i<j} S_{ij}}{\sqrt{2n(n-1)(2n+5)}}. \tag{26}$$

If $|Z| > Z_{1-\alpha/2}$, our decision will be to reject H_0.

To compare the canonical correlation estimators based on Kendall's tau correlation with other canonical correlation estimators, we apply the transformation $\sin\left(\frac{\pi}{2}\rho_\tau\right)$ to get a consistent estimation under normality.

2.5. *Spearman's rho correlation*

Spearman's rank correlation ρ_s is the most popular non-parametric correlation, which is just a Pearson correlation based on the ranks of the observations. This correlation provides resistance against outliers; however, outliers that are properly placed can alter its value considerably. In applications where ties are known to be absent, a simpler procedure can be used to calculate r_s. The differences, $d_i = x_i = y_i$, between the ranks of each observation on the two variables were calculated by Myers and Arnold [23] and ρ_s was given by

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}. \tag{27}$$

When sampling from a bivariate normal distribution, r_s does not estimate the same quantity as the Pearson correlation. To compare the estimators of canonical correlations based on Spearman's rho correlation with other estimators, we need to apply the transformation $\sin\left(\frac{\pi}{2}\rho_\tau\right)$ to get a consistent estimation under normality.

The population Spearman's tau correlation $\rho_s = 0$ under independence. To test

$$H_0 : \rho_s = 0, \tag{28}$$

we need to calculate the statistic

$$T_s = r_s\sqrt{\frac{n-2}{1-r_s^2}}. \tag{29}$$

Our decision will be to reject H_0 if $|T_s| > t_{1-\alpha/2}$, the $1-\alpha/2$ quantile of Student's t distribution with $v = n-2$ degrees of freedom.

2.6. *The MVE estimator*

The MVE estimator is one of the affine-equivariant estimators having a high breakdown point (for details, see [24]). Assume any ellipsoid containing half of the data. The idea is to find the ellipsoid having the smallest volume among all the ellipsoids. When we find this ellipsoid, the mean and covariance matrix of its points are taken as the estimated measures of location and scatter, respectively. In the multivariate normal model, the covariance matrix needs to be rescaled to get consistency. In general, the group of all ellipsoids containing half of the data is very large, so the approximation must be used to find the MVE.

Let $h = (n/2) + 1$, rounded down to the nearest integer. An approach to computing the FMVE estimator is to randomly select h points, without replacement, from the n points available, compute the volume of the ellipse containing these points and then repeat this process many times. The FMVE ellipsoid will be the set of points giving the smallest volume.

2.7. *The MCD estimator*

The MCD estimator is also one among the affine-equivariant estimators having a high breakdown point .The difference between the MCD and MVE estimators is that rather than searching for the subset of half the data that has the smallest volume, the MVE estimator searches for the half that

has the smallest generalized variance. The MCD estimator searches for the half of the data that are most tightly clustered together among all the subsets containing half the data, as measured by the generalized variance. Like the MVE estimator, the group of all subsets of half the data is very large, hence an approximate method must be used. Rousseeuw and Van Driessen [7] described an FMCD algorithm employed to achieve this aim. After we find an approximation of the subset of half the data that minimize the generalized variance, we can obtain the MCD estimate of location and scatter by computing the usual mean and covariance matrix based on this subset. In our comparative study, we used the FMCD and reweighted MCD (WMCD) measures as practical approximations for the MCD.

2.8. *The constrained M-estimators*

Rocke [25] suggested a modified biweight estimator, which is basically a constrained M-estimator, where for values of g and a to be determined, the non-decreasing function $\xi(d)$ is defined as

$$\xi(d) = \begin{cases} \dfrac{g^2}{2} - \dfrac{g^2(g^4 - 5g^2a^2 + 15a^4)}{30a^4} + d^2\left(0.5 + \dfrac{g^4}{2a^4} - \dfrac{g^2}{a^2}\right) & \\ \quad + d^3\left(\dfrac{4g}{3a^2} - \dfrac{4g^3}{3a^4}\right) + d^4\left(\dfrac{3g^2}{2a^4} - \dfrac{1}{2a^2}\right) - \dfrac{4gd^5}{5a^4} + \dfrac{d^6}{6a^4} & \text{if } g \le d \le g + a, \\ \dfrac{d^2}{2} & \text{if } 0 \le d < g, \\ \dfrac{g^2}{2} + \dfrac{a(5a + 16g)}{30} & \text{if } d > g + a. \end{cases} \tag{30}$$

The values of g and a can be chosen by an investigator to acquire the desired breakdown point and the asymptotic rejection probability, approximately referring to the probability that a point will get zero weight when the sample size is large. If the asymptotic rejection probability is to be γ, say, then g and a are determined by $E_{\chi_p^2}(\xi(d)) = b_0$ and

$$g + a = \sqrt{\chi^2_{p,1-\gamma}},$$

where $\chi^2_{p,1-\gamma}$ is the $1 - \gamma$ quantile of a chi-squared distribution with p degrees of freedom. Rocke (1996) showed that this estimator can be computed iteratively.

2.9. *The FCH estimator*

Olive and Hawkins [6] proposed a robust $\sqrt{n}$ consistent estimator. The FCH estimator uses the $\sqrt{n}$ consistent estimator, DGK (Devlin, Gnanadesikan, and Kettenring [26]) estimator and the high breakdown estimator the Olive [27] median ball (MB) estimator as attractors. The FCH estimator also uses a location criterion to choose the attractors. If the DGK location estimator $T_{K,D}$ has a greater Euclidean distance from MED(X) than half the data, then the FCH estimator uses the MB attractor. The FCH estimator uses only the attractor with the smallest determinant if

$$\|T_{K,D} - \text{MED}(X)\| \le \text{MED}(D_i(\text{MED}(X), I_p)). \tag{31}$$

Let T_A, C_A be the attractor used. Then, the estimator (T_F, C_F) takes T_F, T_A and

$$C_F = \frac{\text{MED}(D_i^2(T_A, C_A))}{\chi^2_{p,0.5}} C_A, \tag{32}$$

where $\chi^2_{p,0.5}$ is the0.50th percentile of a chi-squared distribution with p degrees of freedom and F is the FCH estimator. Olive and Hawkins [6] showed that T_{F} is a high breakdown estimator and C_{F} is non-singular even with up to nearly 50% outliers.

2.10. *The RFCH breakdown estimator*

Olive and Hawkins [6] used a standard method of reweighting to produce the RFCH estimator. The RFCH estimator uses two standard reweighting steps. Let $(\hat{\mu}, \tilde{\sum}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{\mathrm{FCH}}, C_{\mathrm{FCH}}) \leq \chi^2_{p,0.975}$ and let

$$\hat{\sum}_1 = \frac{\mathrm{MED}(D_i^2(\hat{\mu}_1, \tilde{\sum}_1))}{\chi^2_{p,0.5}} \tilde{\sum}_1. \tag{33}$$

Then, let $(T_{\mathrm{RFCH}}, \tilde{\sum}_2)$ be the classical estimator applied to the cases with $D_i^2(\hat{\mu}_1, \tilde{\sum}_1) \leq \chi^2_{p.0.975}$ and let

$$C_{\mathrm{RECH}} = \frac{\mathrm{MED}(D_i^2(T_{\mathrm{RFCH}}, \tilde{\sum}_2))}{\chi^2_{p,0.5}} \tilde{\sum}_2. \tag{34}$$

Olive and Hawkins [6] showed that the RFCH estimator is a $\sqrt{n}$ consistent estimator of $(\mu, c \sum)$.

2.11. *The RMVN estimator*

Olive and Hawkins [6] suggested the RMVN estimator as a robust multivariate location and dispersion estimator, and they showed that this estimator is a $(\mu, d \sum)$ consistent estimator of $(\mu, d \sum)$. The RMVN estimator uses a slight modification for a standard reweighting method so that the RMVN estimator gives good estimates of $(\mu, \sum)$ for multivariate normal data, even when certain types of outliers are present (for details, see [6]).

The FCH, RFCH, and RMVN methods of RCCA produce consistent estimators of the kth canonical correlation ρ_k on a large class of elliptically contoured distributions. To see this, suppose $\mathrm{Cov}(x) = c_x \sum$ and $C \equiv C(X) \xrightarrow{P} c \sum$, where $c_x > 0$ and $c > 0$ are some constants. Then, $C_{XY}^{-1} C_{XY} C_{YY}^{-1} C_{YX} \xrightarrow{P} \sum_A = \sum_{XX}^{-1} \sum_{XY} \sum_{YY}^{-1} \sum_{YX}$, and $C_{YY}^{-1} C_{YX} C_{XX}^{-1} C_{XY} \xrightarrow{P} \sum_B = \sum_{YY}^{-1} \sum_{YX} \sum_{XX}^{-1} \sum_{XY}$. Note that $\sum_A$ and $\sum_E$ only depend on $\sum$ and do not depend on the constant c or c_x.

(If C is also the classical covariance matrix applied to some subset of the data, then the correlation matrix $G \equiv R_C$ applied to the same subset satisfies $G_{XX}^{-1} G_{XY} G_{YY}^{-1} G_{YX} \xrightarrow{P} R_A = R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX}$ and $G_{YY}^{-1} G_{YX} G_{XX}^{-1} G_{XY} \to^P R_E = R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY}$). Since eigenvalues are continuous functions of the associated matrix, and the FCH, RFCH, and RMVN estimators are consistent estimators of $c_1 \sum, c_2 \sum$, and $c_3 \sum$ on a large class of elliptically contoured distributions, these three RCCA methods produce consistent estimators of the kth canonical correlation ρ_k on that class of distributions.

Eigenvectors are not continuous functions of the associated matrix, hence it may not be true that the three RCCA methods produce consistent estimators of the canonical vectors σ_k and β_k. In principal component analysis, the eigenvectors from two different methods often roughly differ by a factor of -1. This may be the case for CCA, too.

3. Simulation study

In this section, we employ a simulation study to compare the different methods. We considered the following:

CL: Classical CCA based on eigenvalues and eigenvectors of the matrices (5), which were estimated using the sample covariance matrix.

RP, RM, RW, RK, RS: CCA based on eigenvalues and eigenvectors of the matrices (6) after we used the percentage bend correlation, the biweight midcorrelation, the winsorized correlation, Kendall's tau correlation, and Spearman's rho correlation, respectively, instead of the Pearson correlation.

MV, MC, WM, CM, FC, RF, RMV: CCA based on eigenvalues and eigenvectors of the matrices (5), which are estimated using the FMVE, FMCD, WMCD, CM, FCH, RFCH, RMVN estimators instead of the classical sample covariance matrix.

We used the functions *pball* and *winall* from the Wilcox package at http://www.unt.edu/rss/class/mike/Rallfun-v9_2.txt to compute the percentage bend and the winsorized correlation matrices, respectively. We used the function *bicor* from the package (weighted gene co-expression network analysis (WGCNA)) to compute the midcorrelation matrix and the base functions *cor(,method = c("kendall"))* and *cor(, method = c("spearman"))* to compute the Kendall and the Spearman correlation matrices, respectively.

We used the base functions *cov.mve* and *cov.mcd* to compute the FMVE and FMCD covariance matrices and *covRob(,estim="weighted")* and *covRob(,estim="M")* from the package (robust) to compute the weighted MCD (WM) and constrained M (CM) covariance matrices, respectively. The function *covfch* from the package (rpack.txt) at www.math.siu.edu/olive/rpack.txt was used to compute the FCH and RFCH covariance matrices and the function *covrmvn* was used to compute the RMVN covariance matrix.

We followed the simulation settings given in [13]. We generated $m = 500$ samples with size $n = 500$ and assumed $\sum_{XX} = I_p$ and $\sum_{YY} = I_q$. We summarize the choices for $\sum_{XY}$ in Table 1.

Following the work of Branco *et al.* [13], the following sampling distributions were assumed:

(1) Normal distribution (NOR), $N_{p+q}(0, \sum)$.
(2) Multivariate t distribution with three degrees of freedom.
(3) Symmetric contamination (SCN): there is a probability of 0.95 that an observation is generated from $N_{pq}(0, \sum)$ and a 0.05 probability that it is generated from $N_{p+q}(0.9\Sigma)$.
(4) Asymmetric contamination (ACN): 95% of the data are generated from the $N_{p+q}(0, \sum)$, and 5% of the observations equal the point $tr(\sum)1^t$ (where $tr(\sum)$ is the trace of $\sum$).

The estimated parameters for a replication $j (j = 1, \ldots, m)$ of a specific sampling distribution are denoted by $\hat{\rho}_k^j, \hat{\alpha}_k^j$, and $\hat{\beta}_k^j$ for $k = 1, \ldots, l$. These values were compared with the 'true' parameters ρ_k, α_k, and β_k, which were derived from the specific matrix $\sum$. The measures of mean squared error (MSE) were computed as follows:

$$\mathrm{MSE}(\hat{\rho}_k) = \frac{1}{m}\sum_{i=1}^{m}(\phi(\hat{\rho}_k^j) - \phi(\rho_k))^2, \tag{35}$$

Table 1. Simulation setup: $\sum_{XX} = I_p$ and $\sum_{YY} = I_1$.

p	q	$\sum_{XY}$
2	2	$\begin{bmatrix} 0.9 & 0 \\ 0 & 1/2 \end{bmatrix}$
4	4	$\begin{bmatrix} 0.9 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/4 \end{bmatrix}$

Table 2. The MSEs of $\alpha_1, \alpha_2, \beta_1, \beta_2, \phi(\rho_1)$, and $\phi(\rho_2)$ multiplied by 1000 for 13 different methods – NOR – $p = 2$ and $q = 2$.

	α_1	α_2	β_1	β_2	$\phi(\rho_1)$	$\phi(\rho_2)$
CL	22.33	44.07	22.41	44.29	2.02	2.21
RP	23.96	45.83	24.08	43.67	5.46	3.22
RM	22.92	44.71	23.23	43.62	2.44	2.39
RW	25.74	47.34	27.04	45.49	21.19	6.74
RK	21.91	39.74	21.75	38.69	2.30	2.57
RS	23.75	45.72	23.82	43.97	4.37	2.89
MV	28.48	56.15	28.73	54.79	3.35	3.68
MC	27.78	54.07	28.78	53.02	2.76	3.32
WM	27.12	55.78	27.99	54.04	3.11	2.90
CM	28.12	52.62	29.52	56.39	3.24	3.24
FC	62.89	123.70	60.29	121.65	16.26	16.72
RF	26.59	50.94	24.88	50.03	2.66	2.35
RMV	26.57	51.30	24.98	50.14	2.64	2.36

Table 3. The MSEs of $\alpha_1, \alpha_2, \beta_1, \beta_2, \phi(\rho_1)$, and $\phi(\rho_2)$ multiplied by 1000 for 13 different methods – SCN – $p = 2$ and $q = 2$.

	α_1	α_2	β_1	β_2	$\phi(\rho_1)$	$\phi(\rho_2)$
CL	35.38	69.18	35.32	70.82	5.28	5.09
RP	25.11	46.32	25.82	46.21	9.04	3.46
RM	26.14	47.38	26.99	46.46	8.57	3.48
RW	26.24	46.69	27.79	47.76	25.73	7.09
RK	22.71	41.48	23.55	41.28	2.93	2.45
RS	25.29	46.71	26.01	46.71	7.93	3.29
MV	29.54	56.49	27.89	55.89	3.49	3.27
MC	28.59	55.61	28.34	54.68	3.11	3.03
WM	27.89	54.50	28.13	56.11	3.16	3.37
CM	28.82	58.54	26.75	56.45	3.39	2.99
FC	62.54	119.45	60.79	124.31	18.30	7.94
RF	25.73	49.93	25.97	49.13	2.42	2.74
RMV	26.25	51.18	26.08	49.89	2.43	2.88

Table 4. The MSEs of $\alpha_1, \alpha_2, \beta_1, \beta_2, \phi(\rho_1)$, and $\phi(\rho_2)$ multiplied by 1000 for 13 different methods – T – $p = 2$ and $q = 2$.

	α_1	α_2	β_1	β_2	$\phi(\rho_1)$	$\phi(\rho_2)$
CL	67.91	124.76	65.26	125.41	21.35	17.01
RP	27.50	48.27	28.14	49.69	14.96	5.02
RM	29.64	48.29	28.94	48.96	22.78	7.09
RW	28.19	46.66	27.46	47.53	37.32	8.62
RK	25.41	43.36	24.98	44.46	3.61	3.05
RS	27.54	48.21	28.01	49.45	14.89	4.89
MV	35.49	71.66	37.39	72.03	5.73	5.63
MC	32.12	61.98	32.30	64.74	4.12	4.38
WM	35.68	68.63	34.08	66.88	4.93	4.65
CM	33.85	66.66	36.30	67.89	5.20	4.39
FC	54.34	107.443	53.76	105.34	11.87	10.95
RF	33.89	68.71	34.26	66.93	4.65	4.06
RMV	34.81	69.36	35.27	67.73	4.78	4.41

where $\phi(\rho_k) = \tanh^{-1}(\rho_k)$ is the Fisher transformation of ρ_k

$$\mathrm{MSE}(\hat{\alpha}_k) = \frac{1}{m}\sum_{j=1}^{m} \cos^{-1}\left(\frac{|\alpha_k^t \hat{\alpha}_k^j|}{\|\hat{\alpha}_k^j\| \cdot \|\alpha_k\|}\right), \mathrm{MSE}(\hat{\beta}_k) = \frac{1}{m}\sum_{j=1}^{m} \cos^{-1}\left(\frac{|\beta_k^t \hat{\beta}_k^j|}{\|\hat{\beta}_k^j\| \cdot \|\beta_k\|}\right). \quad (36)$$

Table 5. The MSEs of $\alpha_1, \alpha_2, \beta_1, \beta_2, \phi(\rho_1)$, and $\phi(\rho_2)$ multiplied by 1000 for 13 different methods – ACN – $p = 2$ and $q = 2$.

	α_1	α_2	β_1	β_2	$\phi(\rho_1)$	$\phi(\rho_2)$
CL	103.36	482.04	103.80	483.60	113.22	44.62
RP	37.09	159.47	37.49	163.25	3.94	5.11
RM	39.72	175.93	39.07	179.34	8.54	8.16
RW	33.89	118.39	34.30	122.33	7.52	2.92
RK	70.08	162.09	70.95	165.02	15.96	12.00
RS	39.64	174.70	40.14	178.36	4.71	5.63
MV	29.47	56.70	29.55	55.85	3.32	3.29
MC	29.58	55.49	28.29	53.65	3.16	2.89
WM	27.53	55.40	27.54	53.51	3.12	3.00
CM	29.14	55.79	28.23	55.02	3.17	2.93
FC	66.19	133.87	64.49	136.49	19.01	19.50
RF	25.64	50.06	26.24	48.01	2.46	2.66
RMV	26.59	50.89	27.01	49.27	2.56	2.79

Table 6. The MSEs of $\alpha_1, \alpha_2, \beta_1, \beta_2, \phi(\rho_1)$, $\phi(\rho_2)$, $\phi(\rho_3)$, and $\phi(\rho_4)$ multiplied by 1000 for 13 different methods – SCN – $p = 4$ and $q = 4$.

	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4	$\phi(\rho_1)$	$\phi(\rho_2)$	$\phi(\rho_3)$	$\phi(\rho_4)$
CL	40.15	189.28	369.99	350.81	42.11	189.49	367.54	345.01	2.04	2.27	1.74	2.03
RP	42.95	203.04	396.64	372.65	45.01	203.64	395.71	369.29	5.10	2.64	1.73	2.60
RM	40.97	194.49	379.41	357.41	43.43	195.57	377.49	353.03	2.40	2.24	1.69	2.13
RW	46.02	223.12	431.24	398.82	48.87	220.10	429.79	397.56	20.74	5.56	2.63	3.96
RK	38.28	196.35	391.94	366.69	40.36	197.40	388.87	362.13	2.38	2.47	1.94	2.15
RS	42.02	201.21	393.09	368.89	44.80	201.54	391.33	364.59	4.26	2.49	1.71	2.43
MV	47.67	223.02	445.73	419.86	47.85	224.81	439.87	411.03	2.87	3.35	2.55	2.86
MC	45.66	212.69	412.24	388.61	47.65	213.21	412.51	387.79	2.66	2.91	2.25	2.50
WM	45.81	219.42	413.98	376.50	45.45	219.86	418.74	383.53	2.85	2.77	2.41	2.74
CM	47.15	222.10	434.49	411.44	45.75	222.22	434.86	406.45	2.29	2.60	2.17	2.60
FC	86.01	440.42	744.16	651.38	90.43	446.29	746.01	660.04	12.95	9.85	7.03	10.79
RF	43.64	205.93	427.14	402.41	43.44	207.57	426.55	401.69	2.54	2.49	2.08	2.01
RMV	43.86	206.75	426.44	401.35	43.49	208.57	426.90	402.02	2.57	2.49	2.08	2.02

Table 7. The MSEs of $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4, \phi(\rho_1), \phi(\rho_2), \phi(\rho_3)$, and $\phi(\rho_4)$ multiplied by 1000 for 13 different methods – SCN – $p = 4$ and $q = 4$.

	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4	$\phi(\rho_1)$	$\phi(\rho_2)$	$\phi(\rho_3)$	$\phi(\rho_4)$
CL	63.62	322.68	594.32	523.05	61.61	321.48	585.41	515.44	5.29	5.69	4.14	4.54
RP	46.32	217.61	435.31	400.07	45.25	213.39	428.93	395.49	7.14	2.93	1.98	2.62
RM	47.59	218.25	438.98	406.41	46.52	214.81	431.98	400.96	6.82	2.92	1.95	2.73
RW	49.22	232.68	450.93	412.94	48.60	228.24	443.27	407.85	22.66	5.28	2.87	3.80
RK	42.07	209.78	432.99	398.58	41.14	209.54	425.92	392.86	2.57	2.55	2.25	2.15
RS	46.59	218.19	435.19	398.99	45.29	215.10	427.74	394.17	6.09	2.85	1.97	2.57
MV	48.36	232.67	465.5	428.65	46.68	228.35	459.11	423.46	2.83	2.74	2.54	2.63
MC	45.64	224.39	454.15	422.36	45.97	218.40	449.79	417.72	2.58	2.65	2.41	2.48
WM	46.48	212.17	448.65	423.27	48.26	213.44	451.16	423.24	3.04	2.85	2.19	2.29
CM	47.57	221.75	448.16	411.73	46.977	223.21	448.45	417.50	2.32	2.41	2.27	2.55
FC	88.49	453.17	707.26	615.99	88.71	452.50	707.51	628.79	10.93	10.39	6.79	10.88
RF	44.22	200.74	393.69	369.01	42.95	200.71	395.05	374.93	2.48	2.43	2.08	2.16
RMV	44.56	203.59	400.11	373.51	43.35	203.99	401.85	379.48	2.52	2.54	2.11	2.24

The results of the simulation are presented in Tables 2–9 and Figures 1–4.

From Table 2, where the data are from NOR, we can observe that the best three estimators for the canonical variates α_1 and α_2 are RK, CL, and RM, respectively, and the worst three estimators are FC, CM, and MV, respectively, for α_1; and FC, MV, and WM, respectively, for α_2. Also,

Table 8. The MSEs of $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4, \phi(\rho_1), \phi(\rho_2), \phi(\rho_3)$, and $\phi(\rho_4)$ multiplied by 1000 for 13 different methods – SCN – $p = 4$ and $q = 4$.

	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4	$\phi(\rho_1)$	$\phi(\rho_2)$	$\phi(\rho_3)$	$\phi(\rho_4)$
CL	102.26	499.41	767.89	694.97	104.46	489.55	761.29	676.90	44.41	23.98	9.23	10.73
RP	48.33	226.85	474.58	445.26	49.98	223.58	476.29	448.82	15.44	3.37	2.40	3.38
RM	51.55	242.47	510.29	472.53	53.45	239.23	511.10	475.12	7.89	4.68	2.84	4.30
RW	48.95	231.89	479.51	450.85	50.71	228.88	487.78	456.43	38.65	6.18	3.44	4.48
RK	43.81	222.98	461.69	435.89	45.22	218.88	466.13	439.97	3.59	3.22	2.47	2.55
RS	48.28	227.47	468.55	441.38	49.74	223.65	472.04	445.08	15.59	3.40	2.36	3.42
MV	58.30	282.65	548.73	505.16	57.76	289.23	559.46	510.15	4.22	5.02	3.95	4.48
MC	54.79	267.12	528.93	493.06	54.83	271.66	533.95	493.19	4.05	4.33	3.51	3.78
WM	60.57	293.19	533.81	484.42	59.09	290.98	532.47	477.83	4.25	5.02	3.58	3.96
CM	60.01	298.43	555.98	511.50	56.65	393.52	547.92	494.39	4.69	4.68	3.55	4.09
FC	78.63	380.48	686.59	612.51	77.45	295.59	685.76	614.86	23.74	7.78	5.54	6.95
RF	59.15	267.28	555.07	506.06	57.58	271.41	557.49	517.87	4.14	4.52	3.24	3.83
RMV	59.86	273.32	549.40	505.33	57.89	273.87	548.34	507.93	4.25	4.80	3.54	3.83

Table 9. The MSEs of $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4, \phi(\rho_1), \phi(\rho_2), \phi(\rho_3)$, and $\phi(\rho_4)$ multiplied by 1000 for 13 different methods – SCN – $p = 4$ and $q = 4$.

	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4	$\phi(\rho_1)$	$\phi(\rho_2)$	$\phi(\rho_3)$	$\phi(\rho_4)$
CL	237.72	1101.3	962.36	693.29	238.47	1101.1	960.81	690.75	777.12	198.49	15.391	3.14
RP	61.99	497.71	711.55	579.35	63.45	497.79	710.77	576.68	4.09	12.65	6.05	2.24
RM	40.91	190.51	404.49	383.77	41.51	190.18	403.03	379.29	2.42	2.04	1.86	2.08
RW	57.56	429.79	666.64	566.08	59.183	431.94	665.12	564.07	7.23	3.16	2.58	2.25
RK	117.89	583.91	756.69	595.23	118.91	583.92	755.21	592.90	17.91	32.53	10.15	2.98
RS	66.04	529.45	735.64	592.06	67.36	530.35	734.56	590.09	4.86	15.01	6.37	2.31
MV	45.54	219.39	454.23	428.75	46.82	214.26	449.97	427.47	2.71	2.77	2.4	2.61
MC	45.23	211.48	436.61	410.57	46.06	210.11	433.29	409.14	2.71	2.71	2.29	2.43
WM	46.96	211.37	434.43	411.74	47.51	211.50	437.98	415.38	2.73	2.79	2.26	2.35
CM	46.16	221.53	440.88	409.34	47.13	225.23	444.34	411.39	2.54	2.86	2.49	2.54
FC	91.31	461.02	742.70	645.59	92.33	456.84	734.41	643.08	11.44	11.41	7.70	10.01
RF	44.51	196.56	406.58	384.62	44.51	198.57	405.11	387.70	2.48	2.58	2.06	2.19
RMV	44.72	199.59	414.11	391.04	44.91	201.03	413.33	395.49	2.51	2.65	2.21	2.22

the best three estimators for the canonical variate β_1 are RK, CL, and RM, respectively, and the worst three estimators are FC, CM, and MC, respectively. While the best three estimators for the canonical variate β_2 are RK, RM, and RP, respectively, and the worst three estimators are FC, CM, and MV, respectively. For canonical correlations, the best three estimators for the transformed canonical correlation $\phi(\rho_1)$ are CL, RK, and RM, respectively, and the worst three estimators are RW, FC, and RP, respectively. The best three estimators for the transformed canonical correlation $\phi(\rho_2)$ are CL, RF, and RMV, respectively, and the worst three estimators are FC, RW, and MV, respectively.

From Table 3, where the data are from SCN, we can see that the best three estimators for the canonical variate α_1 are RK, RP, and RS, respectively, and the worst three estimators are FC, CL, and MV, respectively. For α_2, the best three estimators are RK, RP, and RW and the worst are FC, CL, and CM, respectively. Also, the best three estimators for the canonical variate β_1 are RK, RP, and RF, respectively, and the worst three estimators are FC, CL, and MC, respectively. While the best three estimators for the canonical variate β_2 are RK, RP, and RM, respectively, the worst three estimators are FC, CL, and MV, respectively. For canonical correlations, the best three estimators for the transformed canonical correlation $\phi(\rho_1)$ are RF, RMV, and RK, respectively, and the worst three estimators are RW, FC, and RP, respectively. The best three estimators for the transformed canonical correlation $\phi(\rho_2)$ are RK, RF, and RMV, respectively, and the worst three estimators are FC, RW, and CL, respectively.

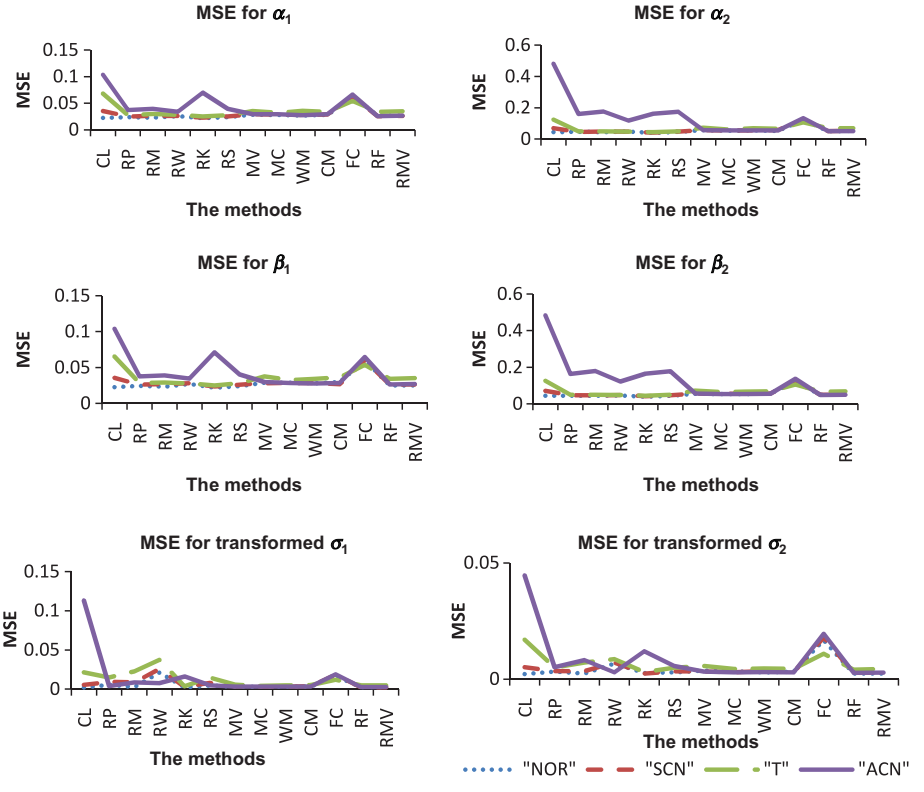

Figure 1. The MSEs for the canonical correlations and vectors for 13 different estimators and under 4 different sampling schemes for $p = 2$ and $q = 2$.

From Table 4, where the data are from T, we can see that the best three estimators for the canonical variate α_1 are RK, RP, and RS, respectively, and the worst three estimators are CL, FC, and WM, respectively. For α_2, the best three estimators are RK, RW, and RS and the worst are CL, FC, and MV, respectively. Also, the best three estimators for the canonical variate β_1 are RK, RW, and RS, respectively, and the worst three estimators are CL, FC, and MV, respectively. While the best three estimators for the canonical variate β_2 are RK, RW, and RM, respectively, the worst three estimators are CL, FC, and MV, respectively. For canonical correlations, the best three estimators for the transformed canonical correlation $\phi(\rho_1)$ are RK, MC, and RF, respectively, and the worst three estimators are RW, RM, and CL, respectively. The best three estimators for the transformed canonical correlation $\phi(\rho_2)$ are RK, RF, and MC, respectively, and the worst three estimators are CL, FC, and RW, respectively.

From Table 5, where the data are from ACN, we can see that the best three estimators for the canonical variates $\alpha_1, \alpha_2, \beta_1$, and β_2 are RF, RMV, and WM, respectively. The worst three estimators for α_1 and β_1 are CL, RK, and FC, respectively, and for α_2 and β_2, they are CL, RM, and RS. For canonical correlations, the best three estimators for the transformed canonical correlation $\phi(\rho_1)$ are RF, RMV, and WM, respectively, and the best three estimators for the transformed canonical correlation $\phi(\rho_2)$ are RF, RMV, and MC, respectively. The worst three estimators for $\phi(\rho_1)$ and $\phi(\rho_2)$ are CL, FC, and RK, respectively.

Figure 1 shows the MSEs for dimensions $p = 2$ and $q = 2$. The first picture from the left and that from the right present the MSEs for the canonical vectors α_1 and α_2. The second picture from the left and that from the right present the MSEs for β_1 and β_2. The third picture from the left and that from the right present the MSEs for the transformed canonical correlations $\phi(\rho_1), \phi(\rho_2)$.

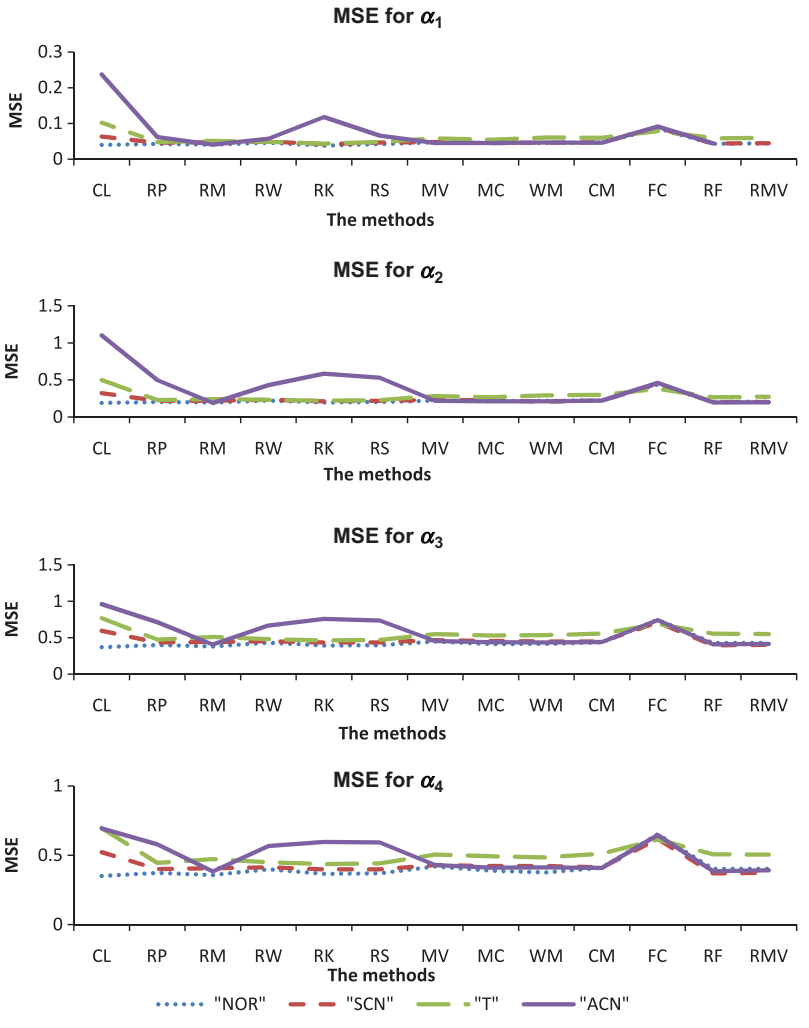

Figure 2. The MSEs for canonical vectors α for 13 different estimators and under 4 different sampling schemes for $p = 4$ and $q = 4$.

The horizontal axis refers to the 13 different methods and the vertical axis refers to the MSEs of the estimators. From Figure 1, it is clear that the largest MSEs are for the estimators in the case of ACN and then for those in the case of T distribution. In the case of ACN, the best estimators for the canonical variates $\alpha_1, \alpha_2, \beta_1$, and β_2 and transformed canonical correlations $\phi(\rho_1)$ and $\phi(\rho_2)$ are RF and RMV and the worst are CL and RK for α_1 and β_1 or CL and RM for α_2 and β_2 or CL and FC for $\phi(\rho_1)$ and $\phi(\rho_2)$. In the case of T distribution, the best estimators for the canonical variates α_2, β_1, and β_2 are RK and RW, while the best estimators for $\alpha_1, \phi(\rho_1)$ and $\phi(\rho_2)$ are RK and RP, RK and MC, and Rk and RF, respectively. The worst estimators for $\alpha_1, \alpha_2, \beta_1, \beta_2$, and $\phi(\rho_2)$ are CL and FC, and for $\phi(\rho_1)$, they are RW and RM.

From Table 6, where the dimensions $p = 4$ and $q = 4$ and the data are from NOR, we can see that the best three estimators for the canonical variates α_1, α_2, and β_1 are RK, CL, and RM, respectively. The worst three estimators for α_1 are FC, MV, and CM, respectively, and for α_2, they are FC, CM, and RW, respectively, and for β_1, they are FC, RW, and MV, respectively. The best three estimators for $\alpha_2, \alpha_4, \beta_2, \beta_3$, and β_4 are CL,RM, and RK, and the worst three estimators are FC, MV, and CM. For canonical correlations, the best three estimators for the transformed

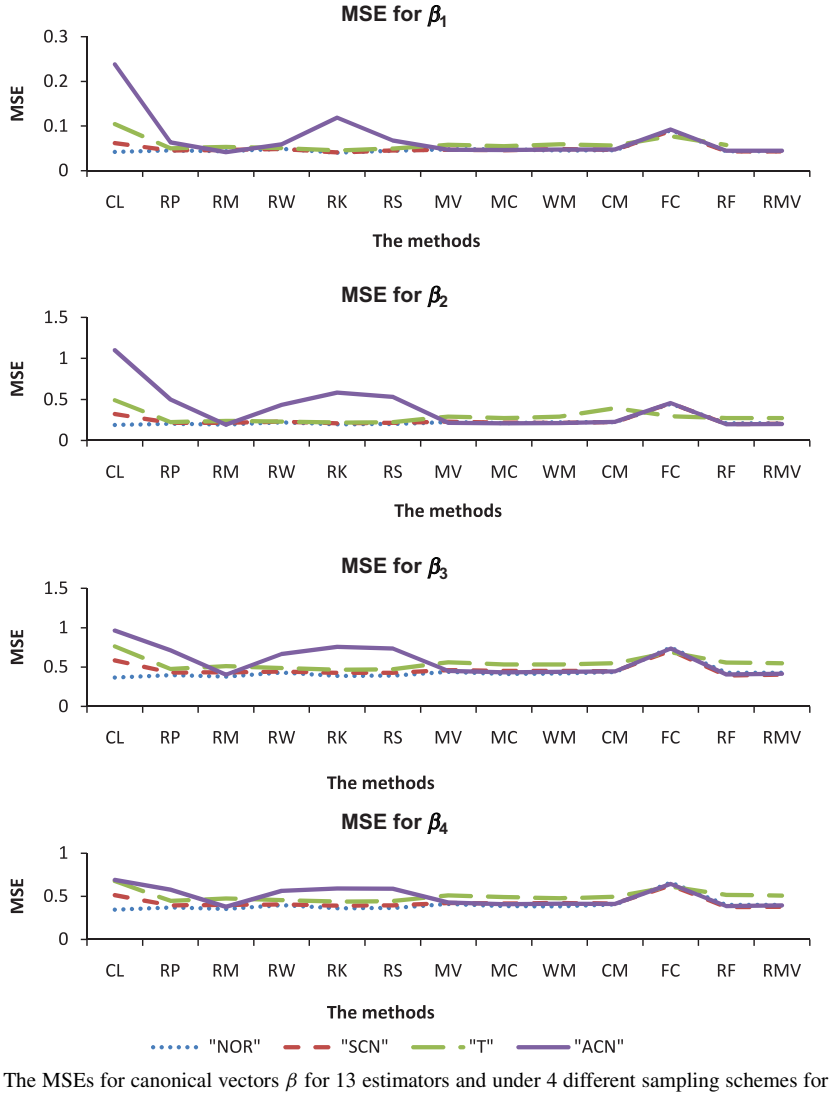

Figure 3. The MSEs for canonical vectors β for 13 estimators and under 4 different sampling schemes for $p = 2$ and $q = 2$.

canonical correlation $\phi(\rho_1)$ are CL, CM, and RK, respectively, and the worst three estimators are RW, FC, and RP, respectively. The best three estimators for the transformed canonical correlation $\phi(\rho_2)$ are RM, CL, and RK, respectively. The best three estimators for the transformed canonical correlation $\phi(\rho_3)$ are RM, RS, and RP. The best three estimators for the transformed canonical correlation $\phi(\rho_4)$ are RM, RS, and RP, respectively. The worst three estimators for $(\rho_2), \phi(\rho_2)$, and $\phi(\rho_4)$ are FC, RW, and MV, respectively.

From Table 7, where the dimensions $p = 4$ and $q = 4$ and the data are from SCN, we can see that the best three estimators for the canonical variates α_1 and β_1 are RK, RF, and RMV, respectively, and the worst three estimators are FC, CL, and RW, respectively. The best three estimators for $\alpha_2, \alpha_3, \alpha_4, \beta_2, \beta_3$, and β_4 are RF, RMV, and RK, and the worst three estimators are FC, CL, and MV. For canonical correlations, the best three estimators for the transformed canonical correlations $\phi(\rho_1)$ and $\phi(\rho_2)$ are CM, RF, and RMV, respectively, and the worst three estimators for $\phi(\rho_1)$ are RW, FC, and RP, respectively. The best three estimators for the transformed canonical correlation $\phi(\rho_3)$ are RM, RS, and RP, respectively. The best three estimators for the transformed canonical

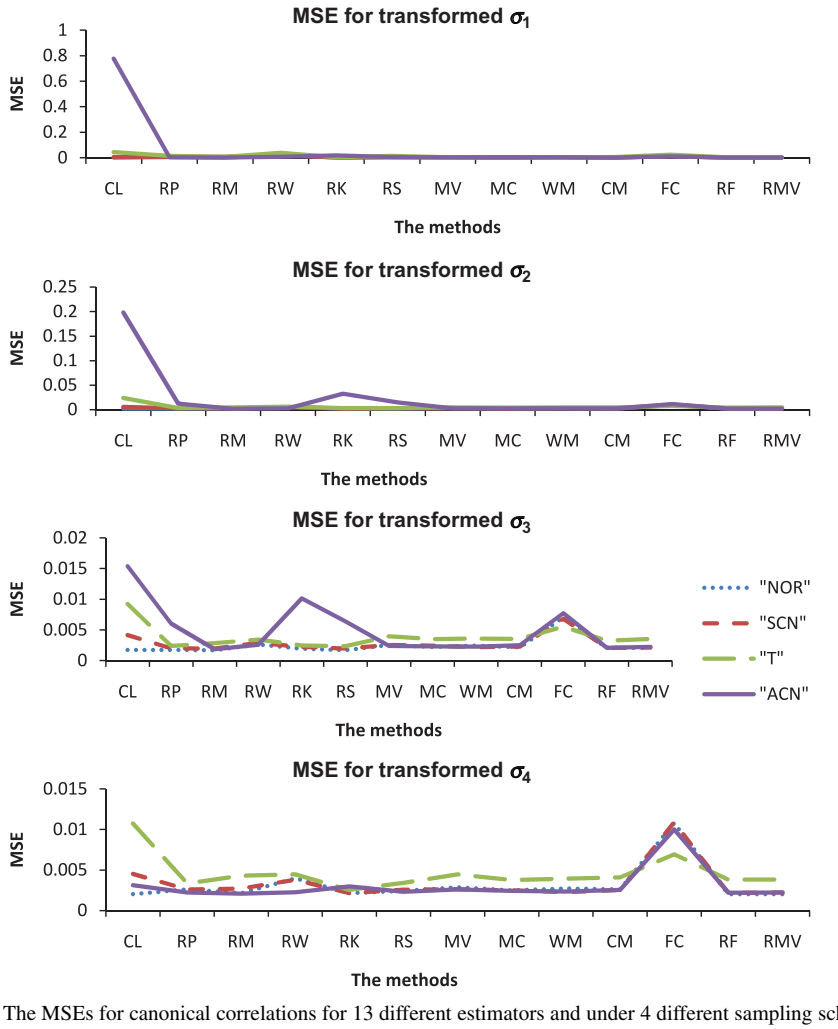

Figure 4. The MSEs for canonical correlations for 13 different estimators and under 4 different sampling schemes for $p = 4$ and $q = 4$.

correlation $\phi(\rho_4)$ are RK, RF, and RMV. The worst three estimators for (ρ_2), $\phi(\rho_3)$, and $\phi(\rho_4)$ are FC, CL, and RW, respectively.

From Table 8, where the dimensions $p = 4$ and $q = 4$ and the data are from T, we can see that the best three estimators for the canonical variates α_1 and β_1 are RK, RS, and RP, respectively, and the worst three estimators are CL, FC, and WM, respectively. The best three estimators for α_2 and β_2 are RK, RP, and RS, and the worst three estimators are CL, FC, and CM. The best three estimators for $\alpha_3, \alpha_4, \beta_2$, and β_4 are RK, RS, and RP, respectively, and the worst three estimators are CL, FC, and CM for α_3 and α_4 or CL, FC, and MV for β_3 and β_4. For canonical correlations, the best three estimators for the transformed canonical correlation $\phi(\rho_1)$ are RK, MC, and RF, respectively, and the worst three estimators are CL, RW, and FC, respectively. The best three estimators for the transformed canonical correlation $\phi(\rho_2)$ are RS, RP, and RK, respectively, and the worst three estimators are CL, FC, and MV, respectively. The best three estimators for the transformed canonical correlations $\phi(\rho_2)$ and $\phi(\rho_4)$ are RK, RP, and RS, respectively, and the worst three estimators are CL, FC, and RW, respectively.

From Table 9, where the dimensions $p = 4$ and $q = 4$ and the data are from ACN, we can see that the best three estimators for the canonical variates $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3$, and β_4 and the

canonical correlations are RM, RF, and RMV, respectively. The worst three estimators for α_1 and β_1 are CL, RK, and FC, respectively; for α_2 and β_2, they are CL, RK, and RS, respectively; for α_3 and β_3, they are CL, RK, and FC, respectively; and for α_4 and β_4, they are CL, FC, and RK, respectively. For the canonical correlations, the worst three estimator for $\phi(\rho_1)$ and $\phi(\rho_3)$ are CL, RK, and FC, respectively. The worst three estimators for $\phi(\rho_2)$ are CL, RK, and RS, respectively, and for $\phi(\rho_4)$, they are FC, CL, and RK, respectively.

4. Breakdown plots

A simulation was carried out to study the sensitivity of the proposed estimators to increasing amounts of contamination. Each of the two groups of variables has three variables ($p = q = 3$),

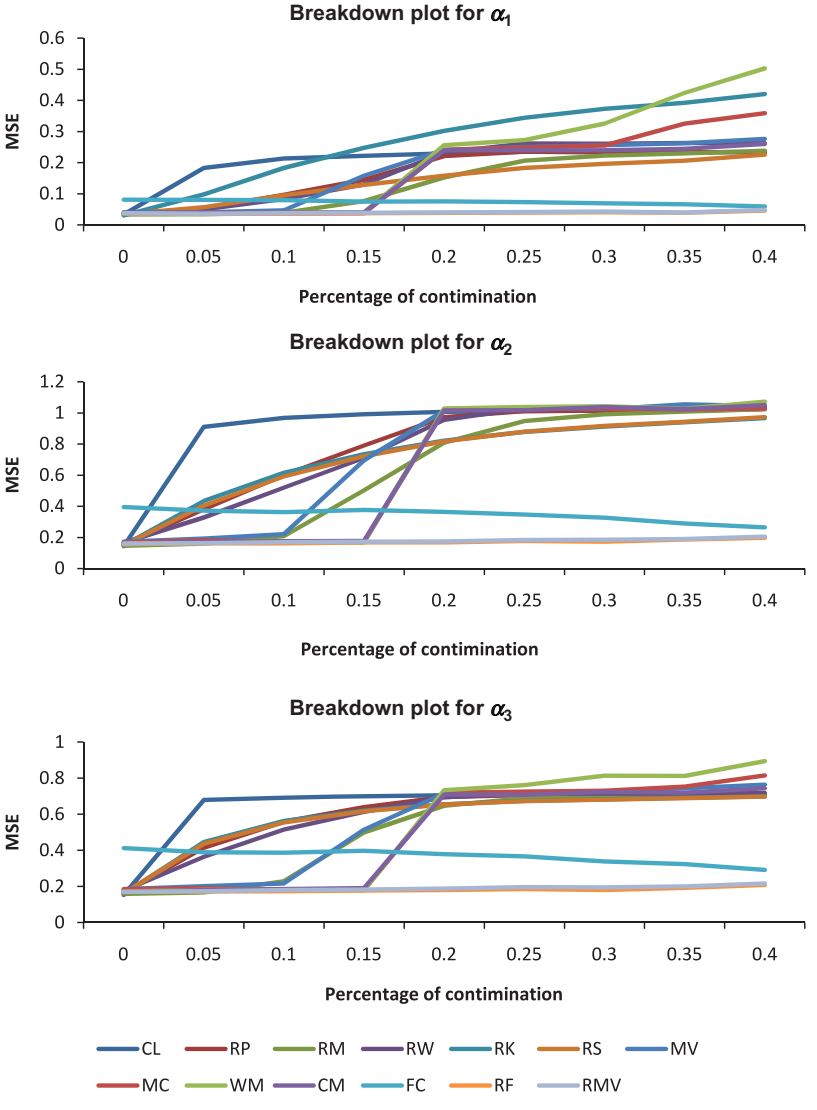

Figure 5. Breakdown plot: MSE for canonical α vectors as a function of the percentage of contamination, ranging from 0% to 40%. The lines represent the different estimation methods.

and the samples were generated from a normal distribution with zero mean and covariance matrix $\sum$, with $\sum_{XX} = I_3$, $\sum_{YY} = I_3$, and

$$\sum_{XY} = \begin{bmatrix} 0.9 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}$$

The values of $\in$ were 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, and 0.40, where $\in$ is the percentage of contamination. The contaminated observations were from the ACN distribution. We chose $n = 500$, and the MSEs were computed over $m = 500$. The results are summarized in Figures 5–7.

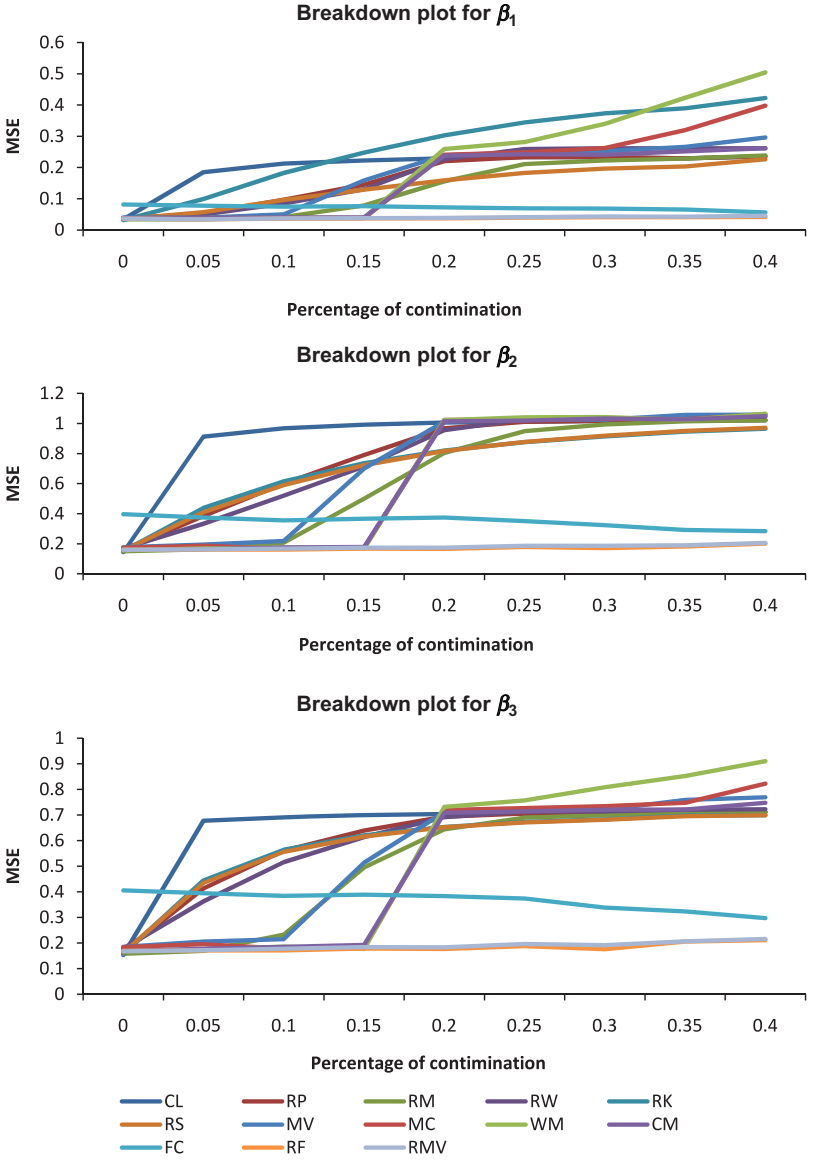

Figure 6. Breakdown plot: MSE for canonical β vectors as a function of the percentage of contamination, ranging from 0% to 40%. The lines represent the different estimation methods.

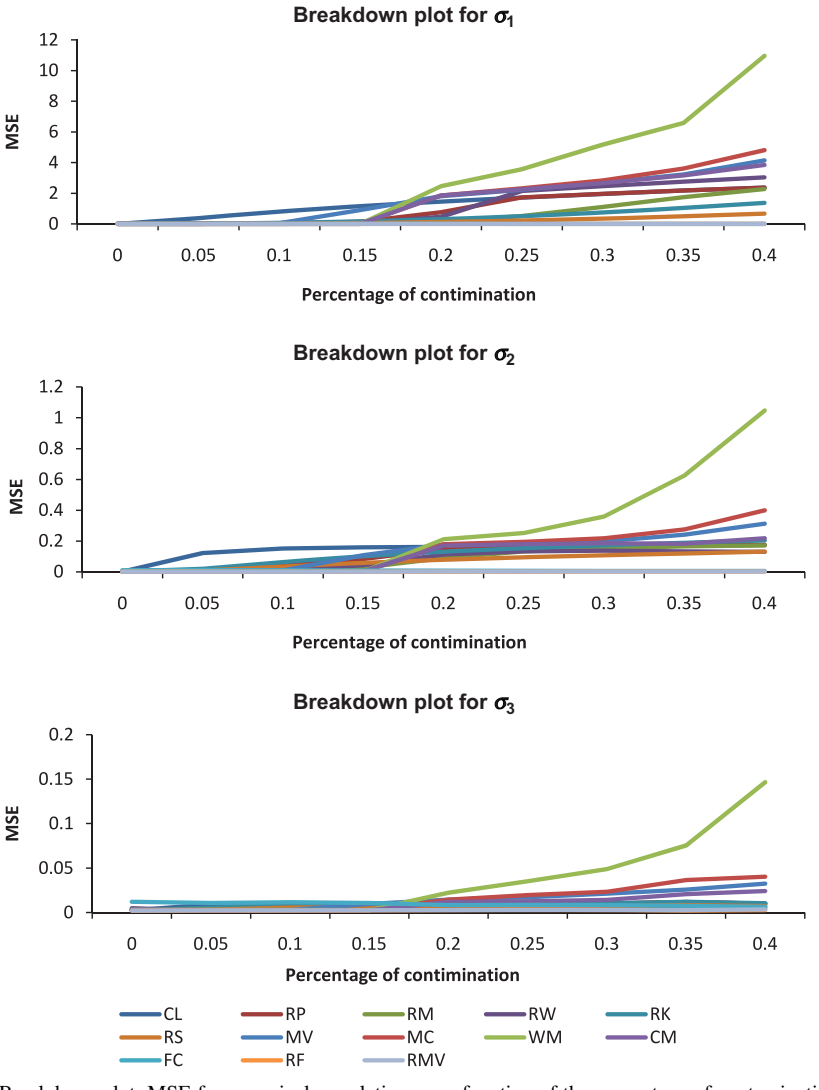

Figure 7. Breakdown plot: MSE for canonical correlations as a function of the percentage of contamination, ranging from 0% to 40%. The lines represent the different estimation methods.

In the figures, different lines correspond to different estimators. The breakdown plots indicate how resistant an estimation procedure is under increasing percentages of contamination.

Figures 5 and 6 show the resistance of the MSE of the canonical vectors α_1, α_2, and α_3 and β_1, β_2 and β_3 for the different methods. It is clearly visible that the MSE of the classical method rapidly increases in the presence of contamination. The classical method is very sensitive with respect to the outlying observations, and the results confirm this behaviour in Figures 1–4. We can see that the robustness of the methods based on the RP, RW, RK, and RS estimators is less than that of other robust methods, where the performance of these estimators worsens as the percentage of contamination is increased beyond 0.05. Similarly, it can be noted that the performance of the RM and MV estimators worsens as the percentage of contamination increases beyond 0.10.

The performance of the methods based on the MC, WM, and CM estimators becomes worst when the percentage of contamination is 0.15 or more, while that of the methods based on the RF

and RMV estimator is still the best for all the percentages of contamination. The performance of the method based on the FC estimator is better when the percentage of contamination increases.

Figure 6 shows the breakdown plots for the canonical correlations. Clearly, the MSEs become smaller, in general, for higher order canonical correlations. The classical method is very sensitive with respect to the outlying observations and its performance is the worst for all canonical correlations and at all percentages of contamination. For the first canonical correlations, the performance of the method based on the RK estimator becomes worst when the percentage of contamination is 0.05 or more. The performance of the methods based on the RP, RS, and MV estimators becomes worst when the percentage of contamination is 0.10 or more. The performance of the methods based on the RM and RW estimators becomes worst when the percentage of contamination is 0.15 or more. The performance of the methods based on the MC, WM, and CM estimators becomes worst when the percentage of contamination is 0.20 or more. The performance of the methods based on the RF and RMV estimators is still the best for all the percentages of contamination, while the performance of the method based on the FC estimator becomes better when the percentage of contamination increases. For the second canonical correlation, the performance of the method based on the RW estimator becomes worst when the percentage of contamination is 0.10 or more. The performance of the methods based on the RK, RP, RS, MV, RM, MC, WM, CM, RF, RMV, and FC estimators is still similar to their performance in the case of the first canonical correlation. For the third canonical correlation, the performance of the methods based on the RP and RK estimators becomes worst when the percentage of contamination is 0.25 or more. The performance of the method based on the RM estimator becomes worst when the percentage of contamination is 0.35 or more. The performance of the methods based on the RF, RMV, RW and RS estimators is still good for all the percentages of contamination. The performance of the methods based on the MV, MC, WM, and CM estimators becomes worst when the percentage of contamination is 0.20 or more. The performance of the method based on the FC estimator becomes better when the percentage of contamination increases.

5. Tests of independence

Assuming that (X,Y) is multivariate normally distributed, the hypothesis of independent can be formulated as

$$H_0 : \sum_{XY} = 0 \quad \text{against} \quad H_1 : \sum_{XY} = 0.$$

If H_0 holds, then all the canonical correlations are equal to zero, thus $H_0 : \rho_1 = \cdot = \rho_p = 0$.

A simulation study was *implemented* to study the effect of the outlier observations in tests of independence. We assumed a situation where the two groups of variables are independent and compute the frequency of rejecting H_0 at the 0.05 significance level. We assumed that each of the two groups of variables has two variables $p = q = 2$, $\sum_{XX} = \sum_{YY} = I_2$ and $\sum_{XY} = \text{Diag}(0.05, 0.01)$. The generated data were from the sample distributions NOR, SCN, and ACN. The estimation methods considered are the classical estimator (CL), percentage bend correlation (RP),

Table 10. The percentage of rejection of the null hypothesis in 1000 simulations.

	CL	RP	RM	RW	RK	RS
NOR	0.007	0.011	0.010	0.010	0.008	0.009
SCN	0.089	0.014	0.009	0.009	0.017	0.014
ACN	1	0.785	0.447	0.447	0.928	0.939

Table 11. The estimated canonical vectors $\hat{\rho}_1$, $\hat{\rho}_2$, and $\hat{\rho}_3$ for the non-contaminated and contaminated data.

	CL			RM			MCD			RFCH			RMVN		
	No contamination	10% contamination	$\lvert D\rvert$	No contamination	10% contamination	$\lvert D\rvert$	No contamination	10% contamination	$\lvert D\rvert$	No contamination	10% contamination	$\lvert D\rvert$	No contamination	10% contamination	$\lvert D\rvert$
$\hat{\rho}_1$	0.464	0.334	0.130	0.465	0.445	0.020	0.444	0.495	0.051	0.482	0.472	0.010	0.479	0.478	0.001
$\hat{\rho}_2$	0.168	0.110	0.085	0.168	0.164	0.004	0.187	0.379	0.192	0.184	0.179	0.005	0.187	0.174	0.013
$\hat{\rho}_3$	0.104	0.060	0.044	0.080	0.019	0.061	0.076	0.033	0.043	0.079	0.102	0.023	0.083	0.106	0.023

Table 12. The estimated canonical vectors $\hat{\alpha}_1, \hat{\alpha}_2$, and $\hat{\alpha}_3$ for the non-contaminated and contaminated data.

			X_1	X_2	X_3	$\sum \lvert D \rvert$
CL	$\hat{\alpha}_1$	No contamination	−0.84	0.25	−0.43	
		10% contamination	−0.08	−0.07	−0.98	
		$\lvert D \rvert$	0.76	0.32	0.55	1.63
	$\hat{\alpha}_2$	No contamination	−0.42	−0.84	0.69	
		10% contamination	0.69	−0.77	−0.06	
		$\lvert D \rvert$	1.11	0.07	0.75	1.93
	$\hat{\alpha}_3$	No contamination	−0.44	0.58	0.69	
		10% contamination	0.74	0.64	−0.25	
		$\lvert D \rvert$	1.18	0.06	0.94	2.18
RM	$\hat{\alpha}_1$	No contamination	−0.84	0.28	−0.43	
		10% contamination	−0.55	0.19	−0.75	
		$\lvert D \rvert$	0.29	0.09	0.32	0.7
	$\hat{\alpha}_2$	No contamination	0.50	0.71	−0.77	
		10% contamination	0.53	0.81	−0.49	
		$\lvert D \rvert$	0.03	0.10	0.28	0.41
	$\hat{\alpha}_3$	No contamination	−0.37	0.71	0.59	
		10% contamination	0.69	−0.61	−0.54	
		$\lvert D \rvert$	1.06	1.32	1.13	3.51
MCD	$\hat{\alpha}_1$	No contamination	−1.46	0.52	−1.30	
		10% contamination	0.52	−0.24	−1.69	
		$\lvert D \rvert$	1.98	0.76	0.39	3.13
	$\hat{\alpha}_2$	No contamination	0.58	1.39	−2.03	
		10% contamination	−1.25	0.04	−0.43	
		$\lvert D \rvert$	1.83	1.35	1.60	4.78
	$\hat{\alpha}_3$	No contamination	−0.68	0.94	2.23	
		10% contamination	0.09	−1.21	0.89	
		$\lvert D \rvert$	0.77	2.15	1.34	4.26
RFCH	$\hat{\alpha}_1$	No contamination	−1.24	0.52	−1.36	
		10% contamination	−1.23	0.47	−1.13	
		$\lvert D \rvert$	0.01	0.05	0.23	0.29
	$\hat{\alpha}_2$	No contamination	0.78	1.23	−2.06	
		10% contamination	0.57	1.33	−1.65	
		$\lvert D \rvert$	0.21	0.10	0.41	0.72
	$\hat{\alpha}_3$	No contamination	−0.62	1.04	1.88	
		10% contamination	−0.66	0.72	2.25	
		$\lvert D \rvert$	0.04	0.32	0.37	0.73
RMVN	$\hat{\alpha}_1$	No contamination	−1.24	0.50	−1.39	
		10% contamination	−1.28	0.42	−1.17	
		$\lvert D \rvert$	0.04	0.08	0.22	0.34
	$\hat{\alpha}_2$	No contamination	0.76	1.25	−2.04	
		10% contamination	0.55	1.43	−1.73	
		$\lvert D \rvert$	0.21	0.18	0.31	0.7
	$\hat{\alpha}_3$	No contamination	−0.66	1.03	1.89	
		10% contamination	−0.72	0.76	2.33	
		$\lvert D \rvert$	0.06	0.27	0.44	0.77

midcorrelation (RM), winsorized correlation (RW), Kendall tau's correlation (RK), and Spearman's rho correlation (RS). We used the functions *pball, winall,* and *spear* from Wilcox package (http://www.unt.edu/rss/class/mike/Rallfun-v9_2.txt) to conduct the test for RP, RW, and RS in Equations (10), (22), and (28), respectively. We used the functions *bicorAndPvalue* from the package WGCNA, and *Kendall* from the package Kendall to test RM in Equation (17) and RK in Equation (25), respectively. We calculated p values associated with the above functions for $m = 1000$ replications.

In the case of NOR data, the test with the classical estimates (CL) gave good results. In the case of SCN and ACN, the test with RM and RW gave the best results. The test with CL estimates was rejected in all 1000 simulations in the case of ACN data.

6. Real data

A researcher has collected data on three psychological variables, four academic variables (standardized test scores), and gender for 600 college freshman (www.ats.ucla.edu/stat/R/dae/canonical.htm). The psychological variables are X_1, locus of control; X_2, self-concept; and X_3, motivation. The academic variables are Y_1, standardized tests in reading; Y_2, writing; Y_3, math; and Y_4, science. Additionally, Y_5, the sex variable, is a zero–one indicator variable with one indicating a female student.

The goal of the researcher is to determine how the set of psychological variables is related to the academic variables and gender.

In the first case, we computed the canonical correlation methods based on the RM, FMCD, RFCH, RMVN, and classical estimators with the above data. In the second case, we contaminated the data with 10% data from multivariate t distribution with three degrees of freedom. Then, all the previous methods were computed.

From Tables 10–13 and Figure 8, we can observe that the results of the methods based on the RFCH, RMVN, and RM estimators are stable and less sensitive to the outliers. However, the results of the method based on the FMCD estimator are changeable and unstable. The performance of the method based on the RM estimator was low than the performance of the methods based on

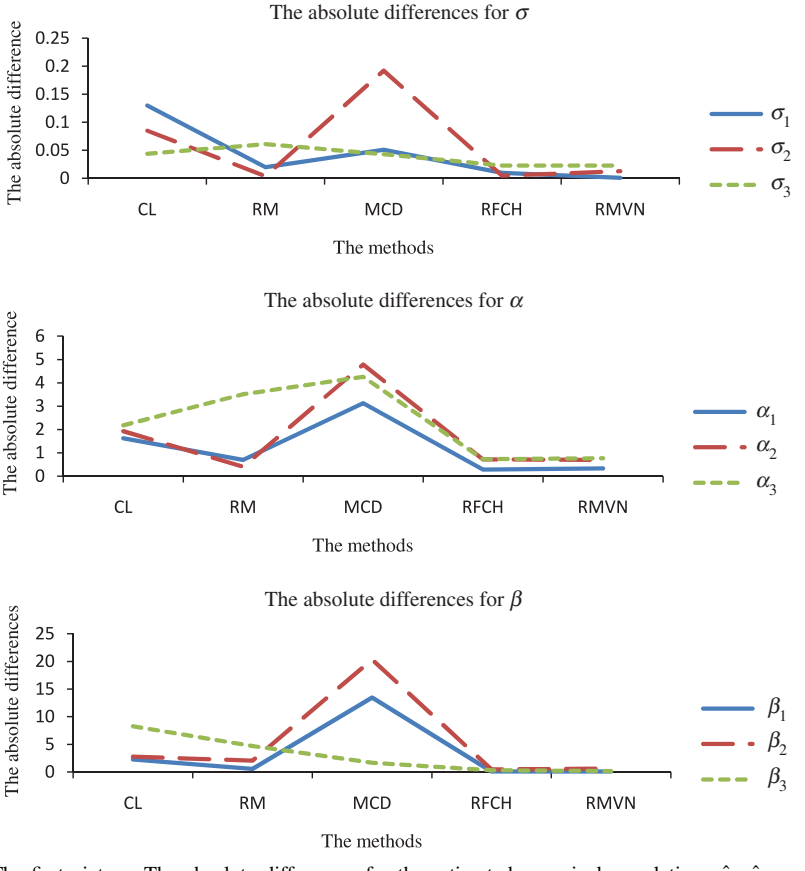

Figure 8. The first picture: The absolute differences for the estimated canonical correlations $\hat{\rho}_1, \hat{\rho}_2$, and $\hat{\rho}_3$ for the non-contaminated and contaminated data. The second picture: The absolute differences for the estimated canonical vectors $\hat{\alpha}_1, \hat{\alpha}_2$, and $\hat{\alpha}_3$ for the non-contaminated and contaminated data. The third picture: The absolute differences for the estimated canonical vectors $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ for the non-contaminated and contaminated data.

Table 13. The estimated canonical vectors $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ for the non-contaminated and contaminated data.

			Y_1	Y_2	Y_3	Y_4	Y_5	$\sum \lvert D\rvert$
CL	$\hat{\beta}_1$	No contamination	−0.45	−0.35	−0.22	−0.05	−0.32	
		10% contamination	−0.38	−1.11	−0.03	0.47	0.41	
		$\lvert D\rvert$	0.07	0.76	0.19	0.52	0.73	2.27
	$\hat{\beta}_2$	No contamination	−0.05	0.41	0.04	−0.83	0.54	
		10% contamination	0.90	0.24	−0.82	−0.39	0.91	
		$\lvert D\rvert$	0.95	0.17	0.86	0.44	0.44	2.79
	$\hat{\beta}_3$	No contamination	0.22	0.89	0.09	−1.07	−0.89	
		10% contamination	2.46	−1.67	−1.08	0.38	−0.05	
		$\lvert D\rvert$	2.24	2.56	1.17	1.45	0.84	8.26
RM	$\hat{\beta}_1$	No contamination	−0.44	−0.41	−0.15	−0.07	−0.29	
		10% contamination	−0.39	−0.60	−0.13	0.11	−0.17	
		$\lvert D\rvert$	0.05	0.19	0.02	0.18	0.12	0.56
	$\hat{\beta}_2$	No contamination	−0.07	−0.48	−0.21	1.13	−0.34	
		10% contamination	0.17	−0.26	0.19	0.43	−0.82	
		$\lvert D\rvert$	0.24	0.22	0.40	0.70	0.48	2.04
	$\hat{\beta}_3$	No contamination	0.41	0.42	0.34	−0.97	−0.92	
		10% contamination	−0.44	−0.60	1.66	−0.61	0.21	
		$\lvert D\rvert$	0.85	1.02	1.32	0.36	1.13	4.68
MCD	$\hat{\beta}_1$	No contamination	−0.06	−0.02	−0.02	−0.005	−0.68	
		10% contamination	−0.007	−0.037	0.006	0.001	−14.070	
		$\lvert D\rvert$	0.053	0.017	0.026	0.006	13.39	13.49
	$\hat{\beta}_2$	No contamination	−0.02	−0.04	−0.01	0.11	−0.89	
		10% contamination	−0.041	0.004	−0.019	−0.018	19.240	
		$\lvert D\rvert$	0.021	0.044	0.009	0.128	20.13	20.33
	$\hat{\beta}_3$	No contamination	0.11	0.009	−0.02	−0.10	−1.44	
		10% contamination	−0.143	−0.00003	0.058	0.103	−2.569	
		$\lvert D\rvert$	0.253	0.00903	0.078	0.203	1.129	1.67
RFCH	$\hat{\beta}_1$	No contamination	−0.04	−0.04	−0.02	−0.004	−0.68	
		10% contamination	−0.04	−0.04	−0.01	−0.01	−0.61	
		$\lvert D\rvert$	0.000	0.000	0.010	0.006	0.070	0.09
	$\hat{\beta}_2$	No contamination	−0.02	−0.03	−0.02	0.11	−0.83	
		10% contamination	−0.01	−0.02	−0.004	0.08	−1.25	
		$\lvert D\rvert$	0.010	0.010	0.016	0.030	0.420	0.49
	$\hat{\beta}_3$	No contamination	0.05	0.09	−0.05	−0.08	−1.79	
		10% contamination	0.04	0.11	−0.06	−0.09	−1.53	
		$\lvert D\rvert$	0.01	0.02	0.01	0.01	0.26	0.31
RMVN	$\hat{\beta}_1$	No contamination	−0.05	−0.04	−0.02	−0.002	−0.67	
		10% contamination	−0.04	−0.04	−0.01	−0.01	−0.61	
		$\lvert D\rvert$	0.010	0.000	0.010	0.008	0.060	0.09
	$\hat{\beta}_2$	No contamination	−0.02	−0.03	−0.02	0.12	−0.80	
		10% contamination	−0.01	−0.02	−0.01	0.08	−1.35	
		$\lvert D\rvert$	0.01	0.01	0.01	0.04	0.55	0.62
	$\hat{\beta}_3$	No contamination	0.05	0.11	−0.08	−0.07	−1.67	
		10% contamination	0.05	0.12	−0.07	−0.09	−1.53	
		$\lvert D\rvert$	0.00	0.01	0.01	0.02	0.14	0.18

the RFCH and RMVN estimators and more stable than that of the method based on the FMCD estimator. As expected, the results of the classical method were highly affected by the outliers. We have used the absolute differences $|D|$ between the estimated values $\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ to measure the changes between the non-contaminated and contaminated data.

7. Conclusions

In this study, we have theoretically investigated and numerically compared a number of methods for comparing canonical correlations. From our simulation study and real data, we can conclude

Table 14. The MSEs of $\alpha_1, \alpha_2, \beta_1, \beta_2, \phi(\rho_1)$, and $\phi(\rho_2)$ multiplied by 1000 for the FMCD method using cov.mcd and covMcd functions – NOR, SCN, T, and ACN – $p = 2$ and $q = 2$, and the computing time, measured in seconds, for $m = 500$ samples with size $n = 500$.

	NOR		SCN		T		ACN	
	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd
α_1	27.78	27.24	28.59	26.53	32.12	32.41	29.58	28.29
α_2	54.07	54.32	55.61	52.26	61.98	65.97	55.49	53.25
β_1	28.78	27.58	28.34	28.08	32.30	36.39	28.29	26.37
β_2	53.02	56.70	54.68	55.94	64.74	69.76	53.65	52.55
$\phi(\rho_1)$	2.76	3.03	3.11	2.84	4.12	4.279	3.16	2.92
$\phi(\rho_2)$	3.32	3.17	3.03	2.96	4.38	4.50	2.89	3.23
Computing time	702	60	1011	368	703	60	703	60

Table 15. The MSEs of $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4, \phi(\rho_1), \phi(\rho_2), \phi(\rho_3)$, and $\phi(\rho_4)$ multiplied by 1000 for the FMCD method using cov.mcd and covMcd functions – NOR, SCN, T, and can – $p = 4$ and $q = 4$, and the computing time, measured in seconds, for $m = 500$ samples with size $n = 500$.

	NOR		SCN		T		ACN	
	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd	cov.mcd	covMcd
α_1	45.66	45.89	45.64	44.99	54.79	58.92	45.23	46.31
α_2	212.69	215.43	224.39	218.43	267.12	284.09	211.48	217.86
α_3	412.24	444.68	454.15	437.79	528.93	547.40	436.61	433.56
α_4	388.61	412.85	422.36	409.68	493.06	513.04	410.57	409.35
β_1	47.65	46.05	45.97	44.78	54.83	59.34	46.06	45.16
β_2	213.21	214.77	218.40	218.35	271.66	282.99	210.11	213.12
β_3	412.51	437.16	449.79	445.43	533.95	543.88	433.29	428.15
β_4	387.79	412.50	417.72	412.08	493.19	502.59	409.14	404.48
$\phi(\rho_1)$	2.66	2.42	2.58	2.66	4.05	5.07	2.71	2.56
$\phi(\rho_2)$	2.91	2.77	2.65	2.78	4.33	5.00	2.71	2.95
$\phi(\rho_3)$	2.25	2.32	2.41	2.52	3.51	3.62	2.29	2.28
$\phi(\rho_4)$	2.50	2.44	2.48	2.15	3.78	4.09	2.43	2.47
Computing time	1808	124	2152	428	1808	123	1800	125

Table 16. The computing time, measured in seconds, for different estimation procedures for $m = 500$ samples with size $n = 500$.

	NOR 2×2	SCN 2×2	T 2×2	ACN 2×2	NOR 4×4	SCN 4×4	T 4×4	ACN 4×4
RP	25	325	26	27	110	437	111	112
RM	27	327	28	29	89	418	91	100
RW	30	329	31	30	70	401	71	86
RK	81	381	81	81	284	618	285	282
RS	9	311	8	8	10	347	10	10
MV	107	411	106	106	256	611	256	249
MC	60	368	60	60	124	428	123	125
WM	63	386	63	63	124	474	124	129
CM	78	396	78	78	132	448	132	138
FC	16	345	16	16	19	341	19	21
RF	16	348	16	16	19	348	19	22
RMV	16	343	16	16	20	349	20	22

that the canonical variates and correlations based on the RFCH and RMVN estimators perform better than the canonical variates and correlations based on the FMCD estimator or the weighted FMCD estimator. Furthermore, from studying the breakdown plots of different estimators, we clearly observed that the performance of the methods based on the RFCH and RMVN estimators to be unrivalled for all percentages of contamination.

Moreover, in the case when the data were from ACN, the simulation study indicated that the performance of the canonical variates and canonical correlation based on the RM estimator is very promising; this fact is especially emphasised in the case when $p = q = 4$ than in that when $p = q = 2$. Additionally, the breakdown plot indicated that the canonical variates and canonical correlations based on the RM estimator are higher than those of other M-type correlations. We also observed that although the breakdown plot showed that the FCH estimator had a high breakdown point, this estimator was one of the worst estimators for all cases.

Later, we took into account the computation time besides robustness and efficiency of estimation. Table 16 shows the computation time, measured in seconds, for different estimation methods for $m = 500$ samples with size $n = 500$. From this table, we can see that the computing time for the RS, FC, RF, and RMV methods is significantly lower than that for the other methods. Also, it is obvious that the MV, CM, WM, and MC methods are time consuming.

From Tables 14 and 15, we can see that the *covMcd* estimator from *the roustbase library* is a much faster implementation of FMCD than *cov.mcd* from *the MASS library*, but the MSEs for the canonical coefficients and canonical correlations are larger in many cases. So, we can recommend the use of the *covMcd* function from the roubustbase library to compute FMCD if we take the computation time into account.

From examining the simulation results of the study, we make a number of practical recommendations. First, in the presence of outliers, we advise the usage of CCA based on the RFCH and RMVN estimators. Second, when the percentage of outliers is pre-determined to be less that 15%, we suggest the employment of CCA based on the RM estimators due to the fact that it has performed very well and that the computing time remains very reasonable. Finally, in the case of contamination above 20%, we do not recommend the usage of the FMCD estimator.

References

[1] H. Hotelling, *Relations between two sets of variates*, Biometrika 28 (1936), pp. 321–377.

[2] S. Das and P.K. Sen, *Canonical correlations,* in *Encyclopedia of Biostatistics*, Vol. 1, P. Armitage and T. Colton, eds., Wiley, New York, 1998, pp. 468–482.

[3] M. Romanazzi, *Influence in canonical correlation*, Psychometrika 57 (1992), pp. 237–259.

[4] G. Karnel, *Robust canonical correlation and correspondence analysis*, in *The Frontiers of Statistical Scientific and Industrial Applications*, *Proceeding of ICOSCO-I, The First International Conference on Statistical Computing*, Vol. II, P.R. Nelson, ed., American Sciences Press, Syracuse, NY, 1991, pp. 335–354.

[5] T. Bernholt and P. Fischer, *The complexity of computing the MCD-estimator*, Theor. Comput. Sci. 326 (2004), pp. 383–398.

[6] D.J. Olive and D.M. Hawkins, *Robust Multivariate Location and Dispersion*, 2010. Available at www.math.siu.edu/olive/pphbmld.pdf.

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

[8] C. Croux and C. Dehon, *Analyse canonique basee sur des estimateurs robustes de la matrice de covariance*, La Revue de Statistique Appliquee 2 (2002), pp. 5–26.

[9] P.J. Rousseeuw, *Multivariate estimation with high breakdown point*, in *Mathematical Statistics and Applications*, W. Grossman, G. Pflug, I. Vincze, and W. Wertz, eds.,Vol. B, Reidel, Dordrecht, 1985, pp. 283–297.

[10] S. Taskinen, C. Croux, A. Kankainen, E. Ollila, and H. Oja, *Influence functions and efficiencies of the canonical correlation and vector estimates based on scatter and shape matrices*, J. Multivariate Anal. 97(2) (2006), pp. 359–384.

[11] H. Wold, *Nonlinear estimation by iterative least squares procedures*, in *A Festschrift for J. Neyman*, F.N. David, ed., Wiley, New York, 1966, pp. 411–444.

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

[13] J.A. Branco, C. Croux, P. Filzmoser, and M.R. Olivera, *Robust canonical correlations: A comparative study*, Comput. Statist. 20(2) (2005), pp.203–229.

[14] J. Jiao and C.H. Jian, *Asymptotic distributions in the projection pursuit based canonical correlation analysis,* Science China press and Springer, Berlin, Heidelberg, 2010. Doi: 10.1007/s11425-010-0035-5.

[15] N.L. Kudraszow and R.A. Maronna, *Robust canonical correlation analysis: A predictive approach*. Preprint submitted to Elsevier, 2011.

[16] J. Zhang and D.J. Olive, *Applications of a Robust Dispersion Estimator*, 2009. Available at www.math.siu.edu/olive/pprcovm.pdf.
[17] J. Zhang, *Applications of a robust dispersion estimator*, Ph.D. thesis, Southern Illinois University, 2011. Available at www.math.siu.edu/olive/szhang.pdf.
[18] S.S. Reyen, J.J. Miller, and E.J. Wegman, *Separating a mixture of two normals with proportional covariances*, Metrika 70 (2009), pp. 297–314.
[19] M. Cannon and W. Hsieh, *Robust nonlinear canonical correlation analysis: Application to seasonal climate forecasting*, Nonlinear Process. Geophys. 15 (2008), pp. 221–232.
[20] R.R. Wilcox, *The percentage bend correlation coefficient*, Psychometrika 59 (1994), pp. 601–616.
[21] L.H. Shoemaker and T.P. Hettmansperger, *Robust estimates and tests for the one-and two-sample scale models*, Biometrika 69 (1982), pp. 47–54.
[22] R.R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing,* 2nd ed., Academic Press, San Diego, CA, 2005.
[23] J.L. Myers and D.W. Arnold, *Research Design and Statistical Analysis,* 2nd ed., Lawrence Erlbaum, Mahwah, NJ, 2003, pp. 1–508.
[24] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection,* Wiley, New York, 1987.
[25] D.M. Rocke, *Robustness properties of S-estimators of multivariate location and shape in high dimension*, Annals of Statistics 24 (1996), pp. 1327–1345.
[26] S.J. Devlin, R. Gnanadesikan, and J.R. Kettenring, *Robust estimation of dispersion matrices and principal components*, Journal of the American Statistical Association 76 (1981), pp. 354–362.
[27] D.J. Olive, A resistant estimator of multivariate location and dispersion, *Computational Statistics and Data Analysis* 46 (2004), pp. 99–102.