

Robust Covariance Matrix Estimation with Canonical Correlation Analysis

Jianfeng Zhang¹, David J. Olive² & Ping Ye³

¹ Department of Mathematics, Chattanooga State Community College, Chattanooga, TN, USA

² Department of Mathematics, Southern Illinois University, Carbondale, IL, USA

³ Department of Mathematics, Quincy University, Quincy, IL, USA

Correspondence: Jianfeng Zhang, Department of Mathematics, Chattanooga State Community College, Chattanooga, TN, USA. Tel: 1-423-697-2404. E-mail: jianfeng.zhang@chattanooga.state.edu

Received: July 23, 2012 Accepted: August 13, 2012 Online Published: September 18, 2012

doi:10.5539/ijsp.v1n2p119

URL: <http://dx.doi.org/10.5539/ijsp.v1n2p119>

Abstract

This paper gives three easily computed highly outlier resistant robust $\sqrt{n}$ consistent estimators of multivariate location and dispersion for elliptically contoured distributions with fourth moments. When the data is from a multivariate normal distribution, the dispersion estimators are also consistent estimators of the covariance matrix. Outlier detection and robust canonical correlation analysis are presented as applications.

Keywords: minimum covariance determinant estimator, multivariate location and dispersion, outliers, canonical correlation analysis, projection pursuit approach, robust projection index

1. Introduction

This paper gives three robust estimators of multivariate location and dispersion and then uses one of the estimators to create a robust method of canonical correlation analysis. The FCH estimator is so named because it is fast, consistent and highly outlier resistant. The reweighted FCH (RFCH) estimator is the second estimator while the RMVN estimator is so named because it is a reweighted FCH estimator that can give useful estimates of the population covariance matrix when the data is from a multivariate normal distribution, even when certain types of outliers are present. This claim will be illustrated in Section 3.1.

Creating a robust estimator and applying it to create a robust method of canonical correlation analysis is not new. See Zhang (2011) and Alkenani and Yu (2012) for references. Typically highly outlier resistant estimators that are backed by theory are impractical to compute while the practical algorithm estimator used to approximate the impractical estimator is not backed by theory. For example, the theoretical robust projection pursuit estimator, discussed in Section 2.4, may not be possible to compute since it is defined on all possible projections. Practical robust projection pursuit algorithms (e.g., that use n randomly chosen projections) typically are not backed by large sample theory. Similarly, the impractical Rousseeuw (1984) minimum covariance determinant (MCD) estimator was shown to be $\sqrt{n}$ consistent by Cator and Lopuhaä (2010), but no large sample theory was provided by Rousseeuw and Van Driessen (1999) for the Fast-MCD (FMCD) estimator. The practical FCH, RFCH and RMVN estimators have been shown to be $\sqrt{n}$ consistent by Olive and Hawkins (2010). These three estimators satisfy Theorem 7.1 in Olive (2012), hence replacing the classical estimator by the RMVN estimator to create a robust method of canonical correlation analysis results in consistent estimators of the population canonical correlations on a large class of elliptically contoured distributions.

A *multivariate location and dispersion (MLD) model* is a joint distribution for a $p \times 1$ random vector $\mathbf{x}$ that is completely specified by a $p \times 1$ population *location* vector $\boldsymbol{\mu}$ and a $p \times p$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. The observations $\mathbf{x}_i$ for $i = 1, \dots, n$ are collected in an $n \times p$ matrix $\mathbf{X}$ with n rows $\mathbf{x}_1^T, \dots, \mathbf{x}_n^T$. An important MLD model is the elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function

$$f(\mathbf{z}) = k_p |\boldsymbol{\Sigma}|^{-1/2} g[(\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})]$$

where $\mathbf{z}$ is a $p \times 1$ dummy vector, $k_p > 0$ is some constant and g is some known function. The multivariate normal (MVN) $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is a special case, and $\mathbf{x}$ is “spherical about $\boldsymbol{\mu}$ ” if $\mathbf{x}$ has an $EC_p(\boldsymbol{\mu}, c\mathbf{I}_p, g)$ distribution

where $c > 0$ is some constant and $\mathbf{I}_p$ is the $p \times p$ identity matrix. Many classical procedures originally meant for the MVN distribution are semiparametric in that the procedures also perform well on a much larger class of EC distributions. See Olive (2012) for examples and references.

For EC distributions, let constants $d > 0$ and $c_X > 0$. Then a dispersion estimator estimates $d \Sigma$, and a covariance matrix estimator estimates the covariance matrix $\text{Cov}(\mathbf{x}) = c_X \Sigma$. Notice that a covariance matrix estimator is also a dispersion estimator. For multivariate analysis, the classical estimator $(\bar{\mathbf{x}}, \mathbf{S})$ of $(E(\mathbf{x}), \text{Cov}(\mathbf{x}))$ is the sample mean and sample covariance matrix where

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad \text{and} \quad \mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T. \quad (1)$$

The following assumptions will be used frequently.

Assumptions (E1): i) The $\mathbf{x}_1, \dots, \mathbf{x}_n$ are iid $EC_p(\mu, \Sigma, g)$ with nonsingular $\text{Cov}(\mathbf{x}_i)$. ii) Assume g is continuously differentiable with finite 4th moment.

Let the $p \times 1$ column vector $T(\mathbf{X})$ be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\mathbf{C}(\mathbf{X})$ be a dispersion estimator. The notation $(T, \mathbf{C})$ will be often be used, suppressing $\mathbf{X}$. Then the i th squared sample Mahalanobis distance is the scalar

$$D_i^2 = D_i^2(T(\mathbf{X}), \mathbf{C}(\mathbf{X})) = (\mathbf{x}_i - T(\mathbf{X}))^T \mathbf{C}^{-1}(\mathbf{X})(\mathbf{x}_i - T(\mathbf{X})) \quad (2)$$

for each observation $\mathbf{x}_i$. Notice that the Euclidean distance of $\mathbf{x}_i$ from the estimate of center $T(\mathbf{X})$ is $D_i(T(\mathbf{X}), \mathbf{I}_p)$. The classical Mahalanobis distance uses $(T, \mathbf{C}) = (\bar{\mathbf{x}}, \mathbf{S})$. The population squared Mahalanobis distance

$$U \equiv D^2 = D^2(\mu, \Sigma) = (\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu). \quad (3)$$

For EC distributions, Johnson (1987, pp. 107-108) states that U has density

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u) \quad (4)$$

for $u > 0$.

Section 2 describes the FCH, RFCH, and RMVN estimators and some competitors. Several methods of canonical correlation analysis are also studied. Section 3 presents simulation studies.

2. Method

The FCH, RFCH, and RMVN estimators are described along with some important competitors such as FMCD. Section 2.1 describes the FCH estimator and some competitors. Section 2.2 uses RMVN to estimate (μ, Σ) when the data are $N_p(\mu, \Sigma)$ even when certain types of outliers are present. As an application, Sections 2.3 and 2.4 describe methods for canonical correlation analysis. Zhang (2011) is followed closely.

2.1 Practical Robust Estimators

Many of the most used practical “robust estimators” generate a sequence of K trial fits called *attractors*: $(T_1, \mathbf{C}_1), \dots, (T_K, \mathbf{C}_K)$. Then some criterion is evaluated and the attractor $(T_A, \mathbf{C}_A)$ that minimizes the criterion is used as the final estimator. One way to obtain attractors is to generate trial fits called *starts*, and then use the following *concentration* technique. Let $(T_{0,j}, \mathbf{C}_{0,j})$ be the j th start and compute all n Mahalanobis distances $D_i(T_{0,j}, \mathbf{C}_{0,j})$. At the next iteration, the classical estimator $(T_{1,j}, \mathbf{C}_{1,j})$ is computed from the $c_n \approx n/2$ cases corresponding to the smallest distances. This iteration can be continued for k steps resulting in the sequence of estimators $(T_{0,j}, \mathbf{C}_{0,j}), (T_{1,j}, \mathbf{C}_{1,j}), \dots, (T_{k,j}, \mathbf{C}_{k,j})$. Then $(T_{k,j}, \mathbf{C}_{k,j}) = (\bar{\mathbf{x}}_{k,j}, \mathbf{S}_{k,j})$ is the j th attractor. The quantities c_n and k depend on the concentration estimator. The Fast-MCD estimator use the classical estimator applied to $K = 500$ randomly drawn elemental sets of $p + 1$ cases as starts. Then the attractor with the smallest determinant $\det(\mathbf{C}_{k,j})$ is used in the final estimator. Hawkins and Olive (1999) have a similar estimator. These are the widely used elemental concentration algorithms. For the estimators in the following paragraph, $k = 5$ concentration steps are used, and $c_n = (n + 1)/2$ for odd n since the distances that are less than or equal to the median distance are used.

The FCH and Olive (2004) median ball algorithm (MBA) estimators use the same two attractors. The first attractor is the Devlin, Gnanadesikan and Kettenring (1981) DGK estimator that uses the classical estimator as the start.

The second attractor is the median ball (MB) estimator that uses the classical estimator computed from the cases with $D_i(\text{MED}(X), I_p) \leq \text{MED}(D_i(\text{MED}(X), I_p))$ as a start where $\text{MED}(X)$ is the coordinatewise median. Thus the start $(T_{0,M}, C_{0,M}) = (\bar{x}_{0,M}, S_{0,M})$ is the classical estimator applied after trimming $M\%$ of the cases furthest in Euclidean distance from $\text{MED}(X)$ for $M \in \{0, 50\}$. The M th attractor is $(T_{k,M}, C_{k,M}) = (\bar{x}_{k,M}, S_{k,M})$. The median ball estimator $(\bar{x}_{k,50}, S_{k,50})$ is also the attractor of $(T_{-1,50}, C_{-1,50}) = (\text{MED}(X), I_p)$. The MBA estimator uses the attractor with the smallest determinant as does the FCH estimator if $\|\bar{x}_{k,0} - \text{MED}(X)\| \leq \text{MED}(D_i(\text{MED}(X), I_p))$. If the DGK location estimator $\bar{x}_{k,0}$ has a greater Euclidean distance from $\text{MED}(X)$ than half the data, then FCH uses the median ball attractor. Let (T_A, C_A) be the attractor used. Then the estimator (T_F, C_F) takes $T_F = T_A$ and

$$C_F = \frac{\text{MED}(D_i^2(T_A, C_A))}{\chi_{p,0.5}^2} C_A \quad (5)$$

where $\chi_{p,0.5}^2$ is the 50th percentile of a chi-square distribution with p degrees of freedom and F is the MBA or FCH estimator.

Olive (2008, § 10.7) and Olive and Hawkins (2010) prove that the MBA and FCH estimators are highly outlier resistant $\sqrt{n}$ consistent estimators of $(\mu, d \Sigma)$ when (E1) holds where $d = 1$ for MVN data. Also C_A and C_{MCD} are $\sqrt{n}$ consistent estimators of $d_{MCD} \Sigma$ where (T_{MCD}, C_{MCD}) is the minimum covariance determinant (MCD) estimator. The proofs use two results. First, Lopuhaä (1999) shows that if a start (T, C) is a consistent estimator of $(\mu, s \Sigma)$, then the attractor is a consistent estimator of $(\mu, a \Sigma)$ where $a, s > 0$ are some constants. Also the constant a does not depend on s , and the attractor and the start have the same rate. If the start is inconsistent, then so is the attractor. Second, the proofs need the result from Butler, Davies and Jhun (1993) and Cator and Lopuhaä (2010) that the MCD estimator is consistent and high breakdown (HB).

2.2 A Practical Robust Covariance Matrix Estimator

This subsection presents the RFCH and RMVN estimators. Since the FCH estimator is a $\sqrt{n}$ consistent estimator, RFCH and RMVN are, too, by Lopuhaä (1999). See Olive and Hawkins (2010).

It is important to note that if (T, C) is a $\sqrt{n}$ consistent estimator of $(\mu, d \Sigma)$, then

$$\begin{aligned} D^2(T, C) &= (x - T)^T C^{-1} (x - T) = (x - \mu + \mu - T)^T [C^{-1} - d^{-1} \Sigma^{-1} + d^{-1} \Sigma^{-1}] (x - \mu + \mu - T) \\ &= d^{-1} D^2(\mu, \Sigma) + O_p(n^{-1/2}). \end{aligned}$$

Thus the sample percentiles of $D_i^2(T, C)$ are consistent estimators of the percentiles of $d^{-1} D^2(\mu, \Sigma)$. For MVN data, $D^2(\mu, \Sigma) \sim \chi_p^2$. Similarly, suppose (T_A, C_A) is a consistent estimator of $(\mu, d \Sigma)$, and that $P(U \leq u_\alpha) = \alpha$ where U is given by (3). Then the scaling in (5) makes C_F a consistent estimator of $d_F \Sigma$ where $d_F = u_{0.5} / \chi_{p,0.5}^2$, and $d_F = 1$ for MVN data.

The RFCH estimator uses two reweighting steps. Let $(\hat{\mu}_1, \tilde{\Sigma}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, C_{FCH}) \leq \chi_{p,0.975}^2$, and let

$$\hat{\Sigma}_1 = \frac{\text{MED}(D_i^2(\hat{\mu}_1, \tilde{\Sigma}_1))}{\chi_{p,0.5}^2} \tilde{\Sigma}_1.$$

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

$$C_{RFCH} = \frac{\text{MED}(D_i^2(T_{RFCH}, \tilde{\Sigma}_2))}{\chi_{p,0.5}^2} \tilde{\Sigma}_2.$$

The RMVN estimator uses $(\hat{\mu}_1, \tilde{\Sigma}_1)$ and n_1 as above. Let $q_1 = \min\{0.5(0.975)n/n_1, 0.995\}$, and

$$\hat{\Sigma}_1 = \frac{\text{MED}(D_i^2(\hat{\mu}_1, \tilde{\Sigma}_1))}{\chi_{p,q_1}^2} \tilde{\Sigma}_1.$$

Then let $(T_{RMVN}, \tilde{\Sigma}_2)$ be the classical estimator applied to the n_2 cases with $D_i^2(\hat{\mu}_1, \hat{\Sigma}_1) \leq \chi_{p,0.975}^2$. Let $q_2 = \min\{0.5(0.975)n/n_2, 0.995\}$, and

$$C_{RMVN} = \frac{\text{MED}(D_i^2(T_{RMVN}, \tilde{\Sigma}_2))}{\chi_{p,q_2}^2} \tilde{\Sigma}_2.$$

Since there are several estimators under consideration, we will use the notation d_E where E stands for the estimator, e.g., RFCH. Then the RFCH and RMVN estimators are highly outlier resistant $\sqrt{n}$ consistent estimators of $(\mu, d_E \Sigma)$ when (E1) holds with $d_E = u_{0.5}/\chi_{p,0.5}^2$ and $d_E = 1$ for MVN data.

If the bulk of the data is $N_p(\mu, \Sigma)$, the RMVN estimator can give useful estimates of (μ, Σ) for certain types of outlier configurations where FCH and RFCH estimate $(\mu, d_E \Sigma)$ for $d_E > 1$. This claim will be illustrated in Section 3.1. Also, let $0 \leq \gamma < 0.5$ be the outlier proportion. If $\gamma = 0$, then $n_i/n \xrightarrow{P} 0.975$ and $q_i \xrightarrow{P} 0.5$. If $\gamma > 0$, suppose the outlier configuration is such that the $D_i^2(T_{FCH}, C_{FCH})$ are roughly χ_p^2 for the clean cases, and the outliers have larger D_i^2 than the clean cases. Then $MED(D_i^2) \approx \chi_{p,q}^2$ where $q = 0.5/(1 - \gamma)$. For example, if $n = 100$ and $\gamma = 0.4$, then there are 60 clean cases and the $q = 5/6$ quantile $\chi_{p,q}^2$ is being estimated instead of $\chi_{p,0.5}^2$. Now $n_i \approx n(1 - \gamma)0.975$, and q_i estimates q . Thus $C_{RMVN} \approx \Sigma$. Of course consistency cannot generally be claimed when outliers are present.

2.3 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multivariate statistical method to identify and quantify the association between two sets of variables. It focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. First, a pair of linear combinations is determined by maximizing the correlation. Next, a pair of linear combinations uncorrelated to previously selected pair is determined by maximizing the correlation, and so on. The pairs of combinations are called the *canonical variables* (*canonical variates*), and their correlations are called *canonical correlations*.

Denote the first set of variables by the p -dimensional variable $\mathbf{x}$ and the second set of variables by the q -dimensional variable $\mathbf{y}$.

$$\mathbf{x} = [X_1, X_2, \dots, X_p]' \quad \text{and} \quad \mathbf{y} = [Y_1, Y_2, \dots, Y_q]'$$

Without loss of generality, assume $p \leq q$. For the random vectors $\mathbf{x}$ and $\mathbf{y}$, let

$$E(\mathbf{x}) = \mu_1 \quad \text{and} \quad E(\mathbf{y}) = \mu_2,$$

$$\text{Cov}(\mathbf{x}) = \Sigma_{11} \quad \text{and} \quad \text{Cov}(\mathbf{y}) = \Sigma_{22},$$

$$\text{Cov}(\mathbf{x}, \mathbf{y}) = \Sigma_{12} = \Sigma_{21}'.$$

Considering $\mathbf{x}$ and $\mathbf{y}$ jointly in a random vector $\mathbf{W}$,

$$\mathbf{W}_{((p+q) \times 1)} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix},$$

with mean vector

$$\mu_{((p+q) \times 1)} = E(\mathbf{W}) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and covariance matrix

$$\Sigma_{((p+q) \times (p+q))} = E[(\mathbf{W} - \mu)(\mathbf{W} - \mu)'] = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \quad (6)$$

Then the canonical coefficients of the first pair of linear combination is determined by

$$(\alpha_1, \beta_1) = \arg \max_{\mathbf{a}, \mathbf{b}} \text{Corr}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}) \quad (7)$$

with the restriction $\text{Cov}(\mathbf{a}'\mathbf{x}) = 1$, $\text{Cov}(\mathbf{b}'\mathbf{y}) = 1$ and $\text{Cov}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}) = 0$. So the first pair of canonical variates is the pair of the linear combinations

$$U_1 = \alpha_1'\mathbf{x} \quad \text{and} \quad V_1 = \beta_1'\mathbf{y}$$

where $\text{Cov}(U_1) = 1$, $\text{Cov}(V_1) = 1$, and $\text{Cov}(U_1, V_1) = 0$. Higher order k th canonical vectors is then recursively defined by

$$(\alpha_k, \beta_k) = \arg \max_{\mathbf{a}, \mathbf{b}} \text{Corr}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}) \quad (8)$$

with the restriction $\text{Cov}(\mathbf{a}'\mathbf{x}) = 1$, $\text{Cov}(\mathbf{b}'\mathbf{y}) = 1$, $\text{Cov}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}) = 0$ and $(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y})$ is uncorrelated with all previous selected canonical variates (U_i, V_i) where $1 \leq i \leq k-1$. The canonical correlation ρ_k between the canonical variates of the k th pair is

$$\rho_k = \text{Corr}(U_k, V_k).$$

Johnson and Wichern (1998, Ch. 10) gives a simple solution to compute the canonical variates. The k th pair of canonical variates, $k = 1, 2, \dots, p$ can be computed as

$$U_k = \mathbf{e}_k' \Sigma_{11}^{-1/2} \mathbf{x} \quad V_k = \mathbf{f}_k' \Sigma_{22}^{-1/2} \mathbf{y} \quad (9)$$

and

$$\text{Corr}(U_k, V_k) = \rho_k$$

where $\rho_1^2 \geq \rho_2^2 \geq \dots \geq \rho_p^2$ are the eigenvalues of the matrix

$$\Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1/2}$$

with associated eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$. Moreover, $\rho_1 \geq \rho_2 \geq \dots \geq \rho_p$ are also the p largest eigenvalues of

$$\Sigma_{22}^{-1/2} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1/2}$$

with associated eigenvectors $\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_p$.

When the original variables to be studied by CCA have quite different measure scales or standard deviations, they usually will be standardized for better analysis and interpretation before computing the canonical variates. Let $\sigma_{ii} = \text{Cov}(X_i)$ and $\nu_{ii} = \text{Cov}(Y_i)$. Further let $\mathbf{V}_{11} = \text{diag}(\sigma_{11}, \sigma_{22}, \dots, \sigma_{pp})$ and $\mathbf{V}_{22} = \text{diag}(\nu_{11}, \nu_{22}, \dots, \nu_{qq})$. Then the standardized random vectors are

$$\mathbf{z}_x = \mathbf{V}_{11}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_x) \quad \text{and} \quad \mathbf{z}_y = \mathbf{V}_{22}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}_y).$$

Consequently the canonical variates standardized vectors, $\mathbf{z}_x$ and $\mathbf{z}_y$ have the form

$$U_k^* = (\boldsymbol{\alpha}_k^*)' \mathbf{z}_x = \mathbf{e}_k' \boldsymbol{\rho}_{11}^{-1/2} \mathbf{z}_x$$

and

$$V_k^* = (\boldsymbol{\beta}_k^*)' \mathbf{z}_y = \mathbf{f}_k' \boldsymbol{\rho}_{22}^{-1/2} \mathbf{z}_y$$

where $\text{Cov}(\mathbf{z}_x) = \boldsymbol{\rho}_{11}$, $\text{Cov}(\mathbf{z}_y) = \boldsymbol{\rho}_{22}$, $\text{Cov}(\mathbf{z}_x, \mathbf{z}_y) = \boldsymbol{\rho}_{12} = \boldsymbol{\rho}_{21}'$, and $\mathbf{e}_k$ and $\mathbf{f}_k$ are the eigenvectors of $\boldsymbol{\rho}_{11}^{-1/2} \boldsymbol{\rho}_{12} \boldsymbol{\rho}_{22}^{-1} \boldsymbol{\rho}_{21} \boldsymbol{\rho}_{11}^{-1/2}$ and $\boldsymbol{\rho}_{22}^{-1/2} \boldsymbol{\rho}_{21} \boldsymbol{\rho}_{11}^{-1} \boldsymbol{\rho}_{12} \boldsymbol{\rho}_{22}^{-1/2}$ respectively. The canonical correlations are given by

$$\text{Corr}(U_k^*, V_k^*) = \rho_k^*,$$

where $\rho_1^* \geq \rho_2^* \geq \dots \geq \rho_p^*$ are the eigenvalues of the matrices of both $\boldsymbol{\rho}_{11}^{-1/2} \boldsymbol{\rho}_{12} \boldsymbol{\rho}_{22}^{-1} \boldsymbol{\rho}_{21} \boldsymbol{\rho}_{11}^{-1/2}$ and $\boldsymbol{\rho}_{22}^{-1/2} \boldsymbol{\rho}_{21} \boldsymbol{\rho}_{11}^{-1} \boldsymbol{\rho}_{12} \boldsymbol{\rho}_{22}^{-1/2}$.

Note that in accordance with the definition of the canonical variate,

$$\begin{aligned} (\boldsymbol{\alpha}_k^*, \boldsymbol{\beta}_k^*) &= \arg \max_{\mathbf{a}^*, \mathbf{b}^*} \text{Corr}[(\mathbf{a}^*)' \mathbf{z}_x, (\mathbf{b}^*)' \mathbf{z}_y] \\ &= \arg \max_{\mathbf{a}^*, \mathbf{b}^*} \text{Corr}((\mathbf{a}' \mathbf{V}_{11}^{-1/2}) \mathbf{x}, (\mathbf{b}' \mathbf{V}_{22}^{-1/2}) \mathbf{y}) \\ &= \arg \max_{\mathbf{a}, \mathbf{b}} \text{Corr}(\mathbf{a}' \mathbf{x}, \mathbf{b}' \mathbf{y}) \\ &= (\mathbf{V}_{11}^{-1/2} \mathbf{a}, \mathbf{V}_{22}^{-1/2} \mathbf{b}). \end{aligned} \quad (10)$$

Therefore, unlike the principal component analysis, CCA has an equivariance property since the canonical correlations are unchanged by the standardization. That is, $\rho_k \equiv \rho_k^*$ for all $1 \leq k \leq p$.

Canonical variates are generally artificial and have no physical meaning. They are latent variables analogous to factors obtained in factor analysis. They often are looked as subject-matter variables. If the original variables are standardized to have zero means and unit variances, then the standardized canonical coefficients are interpreted in a similar manner to standardized regression coefficients. Being increased by one for a standardized variable is the same as being increased by one standard deviation for the corresponding original variable.

Let $\mathbf{A}_{(p \times p)} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_p]'$ and $\mathbf{B}_{(p \times p)} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_p]'$ so that the vectors of canonical variates are

$$\mathbf{U}_{(p \times 1)} = \mathbf{A} \mathbf{x} \quad \text{and} \quad \mathbf{V}_{(q \times 1)} = \mathbf{B} \mathbf{y}.$$

From (9), $A = E'\Sigma_{11}^{-1/2}$ and $B = F'\Sigma_{22}^{-1/2}$ where $E = [e_1, e_2, \dots, e_p]$ and $F = [f_1, f_2, \dots, f_q]$. So

$$\text{Cov}(U) = \text{Cov}(Ax) = A\Sigma_{11}A' = E'\Sigma_{11}^{-1/2}\Sigma_{11}\Sigma_{11}^{-1/2}E = I.$$

Likewise,

$$\text{Cov}(V) = \text{Cov}(By) = I.$$

Decompose Σ_{11} to get $\Sigma_{11} = P_1\Lambda_1P_1'$. It follows that

$$U = Ax = E'\Sigma_{11}^{-1/2}x = E'P_1\Lambda_1^{-1/2}P_1'x.$$

Hence, the canonical variates vector U can be geometrically interpreted as the three-step transformation as follows. A similar geometrical interpretation can be made to V .

- (i) A transformation from x to uncorrelated standardized principal components, $\Lambda_1^{-1/2}P_1'x$;
- (ii) an orthogonal rotation P_1 ;
- (iii) another orthogonal rotation E' .

The canonical coefficients are estimated by using sample covariance matrix instead of population covariance matrix. Denote the data matrix $X = [X_1, X_2, \dots, X_p]$ and $Y = [Y_1, Y_2, \dots, Y_q]$. Equation (9) becomes

$$\hat{U}_k = \hat{e}_k'\Sigma_{11}^{-1/2}X \quad \hat{V}_k = \hat{f}_k'\Sigma_{22}^{-1/2}Y \quad (11)$$

where $\hat{e}_k$, for $0 \leq k \leq p$, is an eigenvector of

$$\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$$

and $\hat{f}_k$, for $0 \leq k \leq p$, is an eigenvector of

$$\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}.$$

Eigenvalues $r_1, r_2, \dots, r_p$ of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ are the sample canonical correlations. Muirhead and Waternaux (1980) shows that if the population canonical correlation coefficients are distinct and the underlying population distribution has finite fourth order cumulant, then the limit joint distribution of $\sqrt{n}(r_i^2 - \rho_i^2)$, for $i = 1, \dots, p$, is p -variate normal. In particular, if the data are drawn from an elliptical distribution with kurtosis 3κ , then the limiting joint distribution of

$$\mu_i = \sqrt{n} \frac{r_i^2 - \rho_i^2}{2\rho_i(1 - \rho_i^2)}, \quad i = 1, \dots, p$$

is $N(0, (\kappa + 1)I_p)$. As a more special case, when the data are drawn from multivariate normal distribution ($\kappa = 0$), the u_i 's are asymptotically iid with a standard normal distribution.

However, these asymptotic results are nonrobust. The outliers have great distorting effect on the classical sample covariance matrix since the eigenvalues and eigenvectors are very sensitive to the presence of outliers. Replacing the classical sample covariance matrix by a robust dispersion estimator, such as RMVN, and then computing the eigenvalues and eigenvectors regularly from the robust dispersion estimator is an approach not only intuitive but also effective for a robust CCA. Later, a simulation will be implemented to compare the classical CCA and robust CCA based on Fast-MCD and RMVN dispersion estimators. The next section discusses the projection pursuit (PP) approach. The idea of the PP approach is to robustify the correlation measure in (7) rather than robustify the classical dispersion matrix.

2.4 Robust Canonical Correlation Analysis Using Projection Pursuit

One has learned that the PCA can be looked as a PP-technique since it searches for the directions that have maximum variances. The classical PCA PP-technique uses the variance function as a projection index and robust PCA uses a robust scale. A similar idea could be applied for canonical correlation analysis. CCA can also be seen as a PP-technique since it seeks for two directions a and b in which the correlation of two projections of the variables x and y , $\text{corr}(a'x, b'y)$, is maximized. The correlation measure in this case is the projection index. The robust PP-technique substitutes the classical correlation measure with a robust estimator of the correlation called robust projection index (RPI). Derivation from a robust covariance matrix of two univariate variables is a common

approach to obtain a RPI. RMVN, Fast-MCD, and M-estimator robust projection indices will be compared in a Monte Carlo study in Section 3.3. Muirhead and Waternaux (1980) provided a limit distribution for classical CCA when the underlying population distribution has finite fourth moment. However, so far there is still no asymptotic theory of RPP available since it is very difficult to work out the properties of the robust CCA estimator analytically. Only simulation studies are conducted to estimate those properties. Branco, Croux, Filzmoser, and Oliveira (2005) proposed an algorithm to perform projection pursuit CCA without the backup of any rigorous theories. The algorithm starts by estimating Σ using a robust estimator. Then Σ is partitioned as

$$\Sigma_{(p+q) \times (p+q)} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{matrix} (p \times p) & (p \times q) \\ (q \times p) & (q \times q) \end{matrix}.$$

Performing a spectral decomposition of Σ_{11} and Σ_{22} ,

$$\Sigma_{11} = \mathbf{A} \mathbf{M} \mathbf{A}' \quad \text{and} \quad \Sigma_{22} = \mathbf{B} \mathbf{N} \mathbf{B}',$$

where $\mathbf{M}$, $\mathbf{N}$ are diagonal and $\mathbf{A}$, $\mathbf{B}$ are orthogonal matrices. Transform the original data $\mathbf{x}$ and $\mathbf{y}$ into

$$(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{M}^{-1/2} \mathbf{A}' \mathbf{x}, \mathbf{N}^{-1/2} \mathbf{B}' \mathbf{y}).$$

Note that

$$\begin{aligned} \arg \max_{\mathbf{a}^*, \mathbf{b}^*} PI[(\mathbf{a}^*)' \mathbf{x}^*, (\mathbf{b}^*)' \mathbf{y}^*] &= \arg \max_{\mathbf{a}^*, \mathbf{b}^*} PI[(\mathbf{a}^*)' \mathbf{M}^{-1/2} \mathbf{A}' \mathbf{x}, (\mathbf{b}^*)' \mathbf{N}^{-1/2} \mathbf{B}' \mathbf{y}] \\ &= \arg \max_{\mathbf{a}^*, \mathbf{b}^*} PI[(\mathbf{A} \mathbf{M}^{-1/2} \mathbf{a}^*)' \mathbf{x}, (\mathbf{B} \mathbf{N}^{-1/2} \mathbf{b}^*)' \mathbf{y}] \\ &= \arg \max_{\mathbf{a}, \mathbf{b}} PI[\mathbf{a}' \mathbf{x}, \mathbf{b}' \mathbf{y}]. \end{aligned}$$

where PI is a robust projection index. So the robust CCA has the equivariance property, meaning new data $(\mathbf{x}^*, \mathbf{y}^*)$ have the same canonical correlation as the original data $(\mathbf{x}, \mathbf{y})$, and their canonical coefficients satisfy

$$\mathbf{a}_i = \mathbf{A} \mathbf{M}^{-1/2} \mathbf{a}_i^* \quad \text{and} \quad \mathbf{b}_i = \mathbf{B} \mathbf{N}^{-1/2} \mathbf{b}_i^*,$$

for $i = 1, \dots, p$. Note that for any $\mathbf{a}$ and $\mathbf{b}$,

$$\begin{aligned} \text{Var}(\mathbf{a}' \mathbf{x}^*) &= \mathbf{a}' \text{Var}(\mathbf{x}) \mathbf{a} = \mathbf{a}' \text{Var}(\mathbf{M}^{-1/2} \mathbf{A}' \mathbf{x}) \mathbf{a} \\ &= \mathbf{a}' (\mathbf{M}^{-1/2} \mathbf{A}) \text{Var}(\mathbf{x}) (\mathbf{A}' \mathbf{M}^{-1/2}) \mathbf{a} \\ &= \mathbf{a}' (\mathbf{M}^{-1/2} \mathbf{A}') (\mathbf{A} \mathbf{M} \mathbf{A}') \mathbf{A} \mathbf{M}^{-1/2} \mathbf{a} \\ &= \mathbf{a}' \mathbf{a}. \end{aligned}$$

Similarly, $\text{Var}(\mathbf{b}' \mathbf{y}^*) = \mathbf{b}' \mathbf{b}$. So to find the first canonical coefficients $(\mathbf{a}_1^*, \mathbf{b}_1^*)$, the projection index $PI(\mathbf{a}' \mathbf{x}^*, \mathbf{b}' \mathbf{y}^*)$ must be maximized subject to $\mathbf{a}' \mathbf{a} = 1$ and $\mathbf{b}' \mathbf{b} = 1$. One can write $\mathbf{a}$ and $\mathbf{b}$ in polar coordinates with norm 1 so that the constraint $\mathbf{a}' \mathbf{a} = 1$ and $\mathbf{b}' \mathbf{b} = 1$ can be satisfied automatically. See Branco, Croux, Filzmoser and Oliveira (2005) for more details. The projection index is then maximized, over the polar angle vectors $(\theta_1, \dots, \theta_{p-1})$, by a standard maximization routine, *mlminb* in R. Once two angle vectors are determined by *mlminb*, they will be converted back to $(\mathbf{a}_1^*, \mathbf{b}_1^*)$.

Now assume that the first $k-1$ pairs of canonical coefficients are already obtained. To get k th pair $(\mathbf{a}_k, \mathbf{b}_k)$, the projection index $PI(\mathbf{a}' \mathbf{x}^*, \mathbf{b}' \mathbf{y}^*)$ must be maximized subject to $\mathbf{a}' \mathbf{a} = 1$, $\mathbf{b}' \mathbf{b} = 1$, $\text{Cov}(\mathbf{a}_k \mathbf{x}^*, \mathbf{a}_i \mathbf{x}^*) = 0$, and $\text{Cov}(\mathbf{b}_k \mathbf{y}^*, \mathbf{b}_i \mathbf{y}^*) = 0$ for $i = 1, \dots, (k-1)$. Note that

$$\begin{aligned} \text{Cov}(\mathbf{a}_k' \mathbf{x}^*, \mathbf{a}_i' \mathbf{x}^*) &= \mathbf{a}_k' \text{Cov}(\mathbf{x}^*, \mathbf{x}^*) \mathbf{a}_i \\ &= \mathbf{a}_k' \mathbf{I} \mathbf{a}_i = \mathbf{a}_k' \mathbf{a}_i \end{aligned}$$

Likewise, $\text{Cov}(\mathbf{b}_k \mathbf{y}^*, \mathbf{b}_i \mathbf{y}^*) = \mathbf{b}_k' \mathbf{b}_i$. Hence $(\mathbf{a}_k, \mathbf{b}_k)$ can be obtained by maximizing the RPI in two subspaces that are orthogonal to $\mathbf{a}_1, \dots, \mathbf{a}_{k-1}$ and $\mathbf{b}_1, \dots, \mathbf{b}_{k-1}$ respectively. Using Gram-Schmidt process, one can construct two orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ such that

$$\mathbf{U} = [\mathbf{a}_1^*, \dots, \mathbf{a}_{k-1}^* | \hat{\mathbf{U}}] \quad \text{and} \quad \mathbf{V} = [\mathbf{b}_1^*, \dots, \mathbf{b}_{k-1}^* | \hat{\mathbf{V}}],$$

where $\hat{U}$ and $\hat{V}$ are orthogonal bases of the subspaces that are orthogonal to $\mathbf{a}_1, \dots, \mathbf{a}_{k-1}$ and $\mathbf{b}_1, \dots, \mathbf{b}_{k-1}$ respectively. Next project the original data to these two subspaces, one gets

$$(\mathbf{x}^{**}, \mathbf{y}^{**}) = (\hat{U}' \mathbf{x}^*, \hat{V}' \mathbf{y}^*).$$

Now one can obtain $(\mathbf{a}^{**}, \mathbf{b}^{**})$ with the data $(\mathbf{x}^{**}, \mathbf{y}^{**})$ by maximizing $PI(\mathbf{a}' \mathbf{x}^{**}, \mathbf{b}' \mathbf{y}^{**})$ subject to $\mathbf{a}' \mathbf{a} = 1$ and $\mathbf{b}' \mathbf{b} = 1$. After $(\mathbf{a}^{**}, \mathbf{b}^{**})$ is determined, it is transformed back to get $(\mathbf{a}_k^*, \mathbf{b}_k^*)$ by

$$\mathbf{a}_k^* = \hat{U} \mathbf{a}^{**} \quad \text{and} \quad \mathbf{b}_k^* = \hat{V} \mathbf{b}^{**}.$$

And then

$$\mathbf{a}_k = \mathbf{A} \mathbf{M}^{-1/2} \mathbf{a}_k^* \quad \text{and} \quad \mathbf{b}_k = \mathbf{B} \mathbf{N}^{-1/2} \mathbf{b}_k^*.$$

The k -th canonical correlation is estimated by $\rho_k = PI(\mathbf{a}_k' \mathbf{x}, \mathbf{b}_k' \mathbf{y})$ for $1 \leq k \leq p$. Once the k -th canonical covariate is obtained, a robust covariance matrix with dimension 2×2 is computed based on two univariate variables $\mathbf{a}_k' \mathbf{x}$ and $\mathbf{b}_k' \mathbf{y}$. The off-diagonal entry of this matrix is then taken to be the estimator of ρ_k .

One obvious advantage of projecting onto subspaces $(\hat{U}, \hat{V})$ is their lower dimensions. The maximization in a lower dimensional space can be much more computationally efficient. Another advantage is that the canonical coefficient $\mathbf{a}_k^*$ and $\mathbf{b}_k^*$ are orthogonal to all previously found $\mathbf{a}_i^*$ and $\mathbf{b}_i^*$ respectively so that the constraint of PI maximization is automatically satisfied.

3. Results

Examples are given and three simulation studies are done in this section. The first simulation shows that the RMVN estimator is estimating $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ when the bulk of the data comes from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution even when certain types of outliers are present. The second simulation compares the outlier resistance of five robust MLD estimators, while the third simulation compares several methods of CCA.

3.1 RMVN Estimator

Simulations suggested (T_{RMVN}, C_{RMVN}) gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a variety of outlier configurations. The *R/Splus* estimator `cov.mcd` is an implementation of the Rousseeuw and Van Driessen (1999) FMCD estimator which is also supposed to be a covariance matrix estimator for MVN data. Shown below are the averages, using 20 runs and $n = 1000$, of the dispersion matrices when the bulk of the data are iid $N_4(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \text{diag}(1, 2, 3, 4)$. The first pair of matrices used $\gamma = 0$. Here the FCH, RFCH and RMVN estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\Sigma}$, while C_{FMCD} seems to be approximately unbiased for $0.94\boldsymbol{\Sigma}$.

RMVN				FMCD			
0.9963	0.0137	0.0020	-0.0007	0.9309	0.0169	0.0112	0.0001
0.0137	2.0123	-0.0011	0.0291	0.0169	1.8845	-0.0034	0.0219
0.0020	-0.0011	2.9841	0.0032	0.0112	-0.0034	2.8026	0.0103
-0.0007	0.0291	0.0032	3.9942	0.0001	0.0219	0.0103	3.7520

Next the data had $\gamma = 0.4$ and the outliers had $\mathbf{x} \sim N_4((0, 0, 0, 15)^T, 0.0001\mathbf{I}_4)$, a near point mass at the major axis. FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN estimated $\boldsymbol{\Sigma}$. The FMCD estimator failed to estimate $d \boldsymbol{\Sigma}$. Note that $\chi^2_{4,5/6}/\chi^2_{4,0.5} = 1.9276$.

RMVN				FMCD			
0.9883	-0.0226	-0.0074	0.0214	0.2271	-0.0157	0.0021	0.0492
-0.0226	1.9642	-0.0216	-0.0018	-0.0157	0.4345	-0.0140	0.0130
-0.0074	-0.0216	3.0532	0.0072	0.0021	-0.0140	0.6732	0.1791
0.0214	-0.0018	0.0072	3.8699	0.0492	0.0130	0.1791	55.6480

Next the data had $\gamma = 0.4$ and the outliers had $\mathbf{x} \sim N_4((15, 15, 15, 15)^T, \boldsymbol{\Sigma})$, a mean shift with the same covariance matrix as the clean cases. Rocke and Woodruff (1996) suggest that outliers with mean shift are hard to detect. Again FCH and RFCH estimated $1.93\boldsymbol{\Sigma}$ while RMVN and FMCD estimated $\boldsymbol{\Sigma}$.

RMVN

1.0130	0.0075	0.0055	-0.0264
0.0075	1.9745	-0.0217	-0.0159
0.0055	-0.0217	2.8701	0.0042
-0.0264	-0.0159	0.0042	3.9760

FMCD

1.0241	0.0020	0.0026	-0.0249
0.0020	1.9995	-0.0337	-0.0167
0.0026	-0.0337	2.9310	0.0052
-0.0249	-0.0167	0.0052	4.0456

3.2 Outlier Resistance

A simple simulation for outlier resistance is to generate outliers and count the percentage of times the minimum distance of the outliers is larger than the maximum distance of the clean cases. Then the outliers can be separated from the clean cases with a horizontal line in the DD plot of classical distances versus robust distances. The simulation used 100 runs and γ was the percentage of outliers. The clean cases were MVN: $\mathbf{x} \sim N_p(\mathbf{0}, \text{diag}(1, 2, \dots, p))$. Outlier types were 1) $\mathbf{x} \sim N_p((0, \dots, 0, pm)^T, 0.0001\mathbf{I}_p)$, a near point mass at the major axis, and 2) the mean shift $\mathbf{x} \sim N_p(pm\mathbf{1}, \text{diag}(1, 2, \dots, p))$ where $\mathbf{1} = (1, \dots, 1)^T$. The near point mass and mean shift outlier configurations are often used in the literature.

Table 1 shows some results for the FCH, RMVN, the Maronna and Zamar (2002) OGK, FMCD and MB estimators. Smaller values of pm and larger values of γ suggest greater sensitivity to outliers. The inconsistent but HB MB estimator is useful for detecting outliers. The OGK and FMCD estimators can outperform the MB, FCH and RMVN estimators especially if p and γ are small. For fixed p , as γ approaches 0.5, the FCH, RMVN and MB estimators appear to have greater sensitivity. The following example illustrates the DD plot.

Example. Buxton (1920) gives various measurements on 87 men including *height*, *head length*, *nasal height*, *bigonal breadth* and *cephalic index*. Five *heights* were recorded to be about 19mm with the true heights recorded under head length. These cases are massive outliers. Figure 1 shows the DD plot with the identity line added as a visual aid. Lines corresponding to the 95th sample percentiles of the classical and robust RMVN distances are also shown.

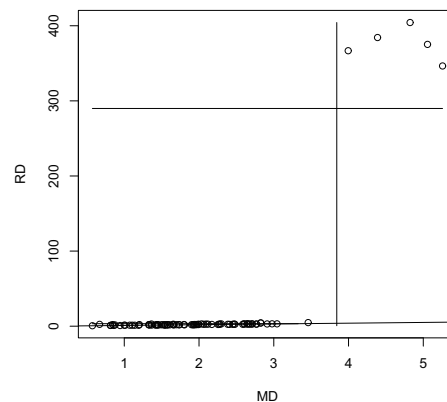


Figure 1. DD plot for Buxton data

Table 1. Percentage of times outliers were detected

p	γ	type	n	pm	FCH	RMVN	OGK	FMCD	MB
5	.25	1	100	10	35	36	0	0	63
5	.25	1	100	20	100	100	0	0	100
5	.49	1	100	20	100	99	0	0	100
5	.40	2	100	10	100	100	0	100	100
5	.47	2	100	10	98	98	0	1	100
20	.2	1	100	30	0	0	0	0	50
20	.2	1	100	50	100	100	0	0	100
20	.2	1	100	100	100	100	29	0	100
20	.2	1	100	4000	100	100	100	2	100
20	.2	1	100	10000	100	100	100	94	100
20	.05	2	100	5	83	91	98	82	86
20	.25	2	100	5	54	61	0	50	71
20	.4	2	100	10	50	50	0	0	100
20	.4	2	100	20	99	99	6	0	100
50	.4	1	200	80	88	84	0	0	88
50	.4	2	200	20	9	9	0	0	100
50	.4	2	200	40	100	100	100	0	100

Table 2. Estimation of Σ with $\gamma = 0.4$, $n = 35p$

p	type	n	pm	Q
5	1	175	16	0.153
5	2	175	6	0.213
10	1	350	21	0.326
10	2	350	6	0.326
15	1	525	26	0.856
15	2	525	7	0.675
20	1	700	33	0.798
20	2	700	8	0.792
25	1	875	39	1.014
25	2	875	10	1.867

Another simulation was done to check that the RMVN estimator estimates Σ for outlier configurations 1) and 2) used in Table 1 if $\gamma = 0.4$. On clean MVN data, $n \geq 20p$ gave good results for $2 \leq p \leq 100$. For the contaminated MVN data, the first $n\gamma$ cases were outliers, and the classical estimator S_c was computed on the clean cases. The diagonal elements of S_c and $\hat{\Sigma}_{RMVN}$ should both be estimating $(1, 2, \dots, p)^T$. The average diagonal elements of both matrices were computed for 20 runs, and the criterion Q was the sum of the absolute differences of the p average diagonal elements. Since $\gamma = 0.4$ and the initial subsets for the RMVN estimator are half sets, the simulations used $n = 35p$. The values of Q shown in Table 2 correspond to good estimation of the diagonal elements. Values of pm slightly smaller than the tabled values led to poor estimation of the diagonal elements.

3.3 Comparing Eight CCA Methods

Two simulation studies for Sections 2.3 and 2.4 are conducted to compare eight different CCA methods, based on:

- 1) the classical sample covariance matrix,
- 2) FMCD covariance matrix estimator,
- 3) M covariance matrix estimator,
- 4) RMVN covariance matrix estimator,
- 5) PP-C (using the classical correlation function as the PI),
- 6) PP-FMCD (using the FMCD correlation estimator as the PI),
- 7) PP-M (using the M correlation estimator as the PI),

8) PP-RMVN (using the RMVN correlation estimator as the PI).

Simulation 1

UCLA: Academic Technology Services (2011) provides a data analysis example of CCA at <http://www.ats.ucla.edu/stat/R/dae/canonical.htm>. The example uses a data file, *mmreg.csv*, available at <http://www.ats.ucla.edu/stat/R/dae/mmreg.csv>. The dataset consists of 600 observations on eight variables. They are *locus of control*, *self-concept*, *motivation*, *reading*, *writing*, *math*, *science*, and *female*. The first three variables are a group of psychological variables. The next four variables are a group of academic variables. The last variable *female* is a categorical indicator. The first simulation studies the canonical correlation between these two groups of variables. The *female* variable is not included in the simulation study since the FMCD is likely to be singular when some of the variables are categorical. See Olive (2004). In fact, two Fast-MCD algorithms, *cov.mcd* and *covMcd*, failed to generate a FMCD estimator when the female variable was included. The DD plot of the *mmreg* dataset from Figure 2 shows the data follows a multivariate normal distribution since all points tightly cluster about the identity line. With the absence of apparent outliers, it is reasonable to assume this dataset is “clean”. Hence, the classical canonical covariates and correlations obtained from this “clean” dataset will be used as benchmarks for a comparison of different CCA methods.

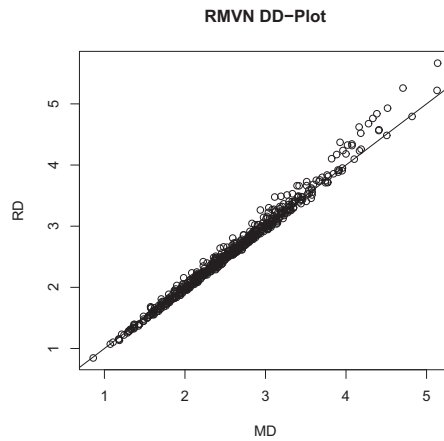


Figure 2. RMVN DD Plot for *mmreg* Data

Let (T, C) be the sample mean and covariance matrix of the *mmreg* dataset. The following different types of outliers are considered:

- 0) No outliers are added to original “clean” dataset.
- 1) 30% (in probability) of the data values are tripled.
- 2) 10% (in probability) of the data values are tripled.
- 3) 30% (in probability) of the observations are replaced by the data generated from a multivariate normal distribution, $N(T, 5C)$.
- 4) 10% (in probability) of the observations are replaced by the data generated from a multivariate normal distribution, $N(T, 5C)$.

Note that when some observations are replaced by outliers, their original values of the *motivation* variable are retained on purpose since it is categorical.

Denote the k -th canonical coefficients and correlation for the i -th replication by $\hat{a}_k^i$, $\hat{b}_k^i$ and $\hat{\rho}_k^i$ where $k = 1, \dots, p$ and $i = 1, \dots, m$. Then the final estimators of k -th canonical coefficients and correlation are computed by

$$\hat{a}_k = \frac{1}{m} \sum_{i=1}^m \hat{a}_k^i, \quad \hat{b}_k = \frac{1}{m} \sum_{i=1}^m \hat{b}_k^i, \quad \text{and} \quad \hat{\rho}_k = \frac{1}{m} \sum_{i=1}^m \hat{\rho}_k^i.$$

Denote the classical canonical coefficients and correlation computed from the “clean” *mmreg* dataset by a_k , b_k and ρ_k . In the first simulation study, a_k , b_k and ρ_k are used as benchmarks for a comparison of different CCA

methods. The correlation, such as $\text{corr}(\hat{\mathbf{a}}_k, \mathbf{a}_k)$, between a canonical covariate and its benchmark will be used as one robustness measure. The mean squared error (MSE) of $\hat{\rho}_k$, as another robustness measure, is defined by

$$\text{MSE}(\hat{\rho}_k) = \frac{1}{m} \sum_{i=1}^m (\tanh^{-1}(\hat{\rho}_k^i) - \tanh^{-1}(\rho_k))^2 \quad (12)$$

where $\tanh^{-1}$ is the inverse hyperbolic function known as the Fisher transformation in Statistics. The Fisher transformation turns the distribution of correlation coefficients toward a normal distribution. The MSE of $\mathbf{a}_k$ is defined by

$$\text{MSE}(\hat{\mathbf{a}}_k) = \frac{1}{m} \sum_{i=1}^m \cos^{-1} \left(\frac{|\hat{\mathbf{a}}_k^i \mathbf{a}_k^k|}{\|\hat{\mathbf{a}}_k^i\| \cdot \|\mathbf{a}_k\|} \right), \quad (13)$$

and the MSE of $\mathbf{b}_k$ is defined in a similar manner by

$$\text{MSE}(\hat{\mathbf{b}}_k) = \frac{1}{m} \sum_{i=1}^m \cos^{-1} \left(\frac{|\hat{\mathbf{b}}_k^i \mathbf{b}_k^k|}{\|\hat{\mathbf{b}}_k^i\| \cdot \|\mathbf{b}_k\|} \right). \quad (14)$$

See Branco, Croux, Filzmoser and Oliveira (2005).

Table 3. Robust CCA with Correlation Measure

outlier	method	ra1	ra2	ra3	rb1	rb2	rb3
0	1	1.00	1.00	1.00	1.00	1.00	1.00
0	2	1.00	-1.00	-0.99	0.99	1.00	-0.73
0	3	1.00	-1.00	-0.97	-0.99	-0.98	-0.55
0	4	1.00	-1.00	0.99	-0.98	-0.98	-0.16
0	5	1.00	-1.00	1.00	-1.00	1.00	-0.62
0	6	0.60	-0.46	0.43	-0.62	0.71	0.06
0	7	1.00	-1.00	0.96	-0.99	0.99	-0.51
0	8	0.96	-0.96	-0.86	-0.85	0.99	0.00
1	1	0.99	-0.99	0.78	-0.54	0.82	0.73
1	2	0.99	-0.99	0.96	0.81	-0.99	0.48
1	3	0.97	0.99	-0.93	0.99	0.98	-0.14
1	4	0.99	-1.00	-0.97	-0.96	0.99	0.28
1	5	0.95	-0.99	0.69	-0.93	0.90	0.00
1	6	0.27	0.69	0.52	-0.67	0.73	0.00
1	7	-0.18	0.96	-0.85	-0.06	-0.98	-0.42
1	8	0.98	0.37	0.56	-0.70	0.36	0.35
2	1	0.71	-1.00	0.52	-0.19	0.92	0.17
2	2	0.96	-1.00	0.97	0.96	0.99	0.99
2	3	0.99	-1.00	-0.99	0.98	0.99	0.35
2	4	1.00	-1.00	-0.99	-0.96	-0.97	0.41
2	5	-0.30	-0.81	0.66	-0.06	0.45	0.01
2	6	0.98	0.42	0.60	-0.34	-0.86	0.38
2	7	-0.98	-1.00	-0.88	0.91	0.99	0.07
2	8	0.92	1.00	-0.97	-0.54	-0.92	0.30
3	1	0.49	-0.95	-0.65	-0.40	0.97	0.28
3	2	0.74	0.01	-0.44	0.91	-0.57	-0.55
3	3	0.95	-0.97	-0.87	-0.89	0.93	-0.86
3	4	1.00	0.98	-0.52	-0.89	0.89	0.69
3	5	-0.81	-0.99	0.06	-0.20	0.90	-0.51
3	6	-0.51	0.67	-0.45	-0.33	0.55	-0.51
3	7	-1.00	-0.76	0.57	0.52	0.70	0.63
3	8	0.98	-0.50	-0.62	-0.86	-0.21	-0.74
4	1	0.89	1.00	1.00	0.96	0.71	0.47
4	2	1.00	-1.00	-0.96	1.00	1.00	0.59
4	3	1.00	-1.00	-0.93	0.99	0.96	0.61
4	4	1.00	1.00	1.00	0.87	0.97	0.41
4	5	1.00	-1.00	0.98	-0.99	0.87	-0.84
4	6	-0.30	0.33	-0.38	-1.00	-0.94	-0.21
4	7	-0.99	1.00	-1.00	0.84	-0.90	0.93
4	8	0.75	0.94	0.96	-0.55	-0.98	-0.35

Table 4. Robust CCA with MSE Measure -0.3cm

outlier	method	Mr1	Mr2	Mr3	Ma1	Ma2	Ma3	Mb1	Mb2	Mb3
0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	2	0.05	0.01	0.03	0.00	0.00	0.76	0.00	0.00	0.76
0	3	0.75	0.69	0.45	0.09	0.20	1.01	0.09	0.20	1.01
0	4	0.40	0.46	0.13	0.00	0.04	1.41	0.00	0.04	1.41
0	5	0.00	0.00	0.13	0.00	0.00	0.98	0.00	0.00	0.98
0	6	158.65	23.27	25.13	0.68	1.14	1.44	0.68	1.14	1.44
0	7	0.10	0.01	0.29	0.25	0.00	1.06	0.25	0.00	1.06
0	8	0.17	1.19	0.49	0.55	0.00	1.40	0.55	0.00	1.40
1	1	51.41	1.04	1.23	1.10	1.12	1.23	1.10	1.12	1.23
1	2	30.84	0.64	0.26	1.08	0.72	1.02	1.08	0.72	1.02
1	3	1.19	1.06	0.58	0.53	0.59	1.06	0.53	0.59	1.06
1	4	1.26	1.47	1.06	0.13	0.17	1.03	0.13	0.17	1.03
1	5	54.21	1.37	0.30	1.06	1.14	1.45	1.06	1.14	1.45
1	6	122.09	34.32	25.89	0.74	1.26	1.38	0.74	1.26	1.38
1	7	27.64	2.92	1.34	1.10	0.93	1.42	1.10	0.93	1.42
1	8	35.43	25.91	9.35	0.27	1.01	1.32	0.27	1.01	1.32
2	1	41.87	1.69	1.88	0.85	0.96	1.12	0.85	0.96	1.12
2	2	27.12	0.19	0.08	0.34	0.28	0.83	0.34	0.28	0.83
2	3	1.70	0.71	0.33	0.28	0.37	0.99	0.28	0.37	0.99
2	4	0.60	0.77	0.42	0.07	0.06	1.19	0.07	0.06	1.19
2	5	42.18	1.28	0.41	0.86	0.94	1.40	0.86	0.94	1.40
2	6	175.89	19.64	37.28	0.70	1.19	1.44	0.70	1.19	1.44
2	7	24.08	1.71	0.97	0.77	0.41	1.29	0.77	0.41	1.29
2	8	27.41	15.95	9.63	0.26	0.93	1.33	0.26	0.93	1.33
3	1	3.08	2.10	1.00	0.94	1.05	1.19	0.94	1.05	1.19
3	2	1.95	1.76	1.14	0.78	0.80	1.03	0.78	0.80	1.03
3	3	2.27	1.53	0.97	0.58	0.65	1.02	0.58	0.65	1.02
3	4	1.93	1.81	0.93	0.26	0.40	0.94	0.26	0.40	0.94
3	5	2.81	2.25	0.29	0.96	1.04	1.40	0.96	1.04	1.40
3	6	246.93	23.08	31.15	0.87	1.19	1.42	0.87	1.19	1.42
3	7	2.32	2.13	0.93	0.80	0.93	1.35	0.80	0.93	1.35
3	8	33.57	23.21	11.04	0.58	1.00	1.30	0.58	1.00	1.30
4	1	1.42	1.81	1.08	0.64	0.72	0.98	0.64	0.72	0.98
4	2	0.56	0.75	0.38	0.38	0.38	0.85	0.38	0.38	0.85
4	3	1.85	0.74	0.39	0.32	0.39	0.90	0.32	0.39	0.90
4	4	0.62	0.66	0.52	0.10	0.13	1.13	0.10	0.13	1.13
4	5	1.85	1.26	0.24	0.67	0.69	1.37	0.67	0.69	1.37
4	6	225.93	24.20	33.37	0.76	1.12	1.45	0.76	1.12	1.45
4	7	1.61	1.41	0.44	0.46	0.49	1.26	0.46	0.49	1.26
4	8	31.97	18.19	8.31	0.37	0.93	1.31	0.37	0.93	1.31

The results of the simulation, with the number of replications $m = 150$, are shown in Tables 3 and 4. In Table 3, the column with header “ra1” gives the value of $\text{corr}(\hat{\mathbf{a}}_1, \mathbf{a}_1)$. All other columns to the right are similar. Table 3 shows all CCA methods except PP-FMCD perform well on a clean dataset (*outlier* = 0) since $\text{corr}(\hat{\mathbf{a}}_k, \mathbf{a}_k)$ and $\text{corr}(\hat{\mathbf{b}}_k, \mathbf{b}_k)$ are quite close to 1 for $k = 1, 2$. When 30% the values are tripled, the PP-FMCD and PP-M estimators failed quite badly. RMVN works well both as PI and as robust dispersion estimator. In Table 4, the column with header “Mr1” gives the value $1000 * \text{MSE}(\hat{\rho}_1)$. The “Mr2” and “Mr3” columns are similar. The column with header “Ma1” gives the value $\text{MSE}(\hat{\mathbf{a}}_1)$. The rest of the columns to the right are similar. The PP-FMCD MSEs really stand out. It has larger MSEs than all other approaches for all different types of outliers.

Tables 3 and 4 are consistent regarding two aspects: (i) as a whole, the CCA methods using projection pursuit are not as good as the CCA methods based on robust dispersion estimators; (ii) PP-FMCD does not work well as a robust CCA technique.

By Theorem 7.1 in Olive (2012), the CCA estimator based on the RMVN estimator produces consistent estimators of the canonical correlations for a large class of elliptically contoured distributions. A possible cause of the poor performance of PP-FMCD is that PP-FMCD does not produce an adequate estimator of its population analog. The FMCD estimator has not been shown to be a consistent estimator of $(\mu, c\Sigma)$, and there are no large sample theory results for PP-FMCD.

The simulation program shows that the running time of the projection pursuit approach is at least 10 times longer than the approaches based on dispersion matrices. Among all RPP approaches, the PP-M is the most computationally inefficient.

Simulation 2

In the second CCA simulation study, the following sampling distributions are considered:

- 1) normal distribution, $N_{p+q}(0, \Sigma)$,
- 2) normal mixture, $.8N_{p+q}(0, \Sigma) + .2N_{p+q}(0, 8\Sigma)$,
- 3) normal mixture, $.95N_{p+q}(0, \Sigma) + .05N_{p+q}(0, 8\Sigma)$,
- 4) mixture distribution, $.8N_{p+q}(0, \Sigma) + .2\delta(\text{tr}(\Sigma)\mathbf{1}')$,
- 5) mixture distribution, $.95N_{p+q}(0, \Sigma) + .05\delta(\text{tr}(\Sigma)\mathbf{1}')$,
- 6) mixture distribution, $.8N_{p+q}(0, \Sigma) + .2\delta(\text{tr}(\Sigma) * [1, 0, \dots, 0]')$,
- 7) mixture distribution, $.95N_{p+q}(0, \Sigma) + .05\delta(\text{tr}(\Sigma) * [1, 0, \dots, 0]')$,

where $\text{tr}(\Sigma)$ represents the trace of the Σ and $\delta()$ represents a point mass distribution. To form the covariance matrix Σ , let $\Sigma_{11} = I_p$, $\Sigma_{22} = I_q$ and Σ_{12} be one of the following:

- 1) $\Sigma_{12} = \begin{bmatrix} .9 & 0 & 0 & 0 \\ 0 & .3 & 0 & 0 \end{bmatrix}$, $\Sigma_{11} = I_2$, $\Sigma_{22} = I_4$;
- 2) $\Sigma_{12} = \begin{bmatrix} .9 & 0 & 0 \\ 0 & .5 & 0 \\ 0 & 0 & .2 \end{bmatrix}$, $\Sigma_{11} = I_3$, $\Sigma_{22} = I_3$;
- 3) $\Sigma_{12} = \begin{bmatrix} .9 & 0 & 0 & 0 & 0 \\ 0 & .7 & 0 & 0 & 0 \\ 0 & 0 & .4 & 0 & 0 \\ 0 & 0 & 0 & .3 & 0 \\ 0 & 0 & 0 & 0 & .1 \end{bmatrix}$, $\Sigma_{11} = I_5$, $\Sigma_{22} = I_5$.

Σ_{11} and Σ_{22} are set to be identity matrices due to the equivariant property of CCA. The sample size of the simulation is $n = 1000$ and the number of replications is $m = 200$. The benchmarks in simulation 2 are the true values of ρ_k , a_k and b_k computed from the matrix Σ .

Table 5. Robust CCA Simulation 2, cov type=3

cov	sdt	mdt	ra1	ra2	rb1	rb2	Mr1	Mr2	Ma1	Ma2	Mb1	Mb2
3	1	1	1.00	1.00	1.00	1.00	3.21	1.35	0.32	0.22	0.32	0.22
3	1	2	1.00	1.00	1.00	1.00	2.96	1.25	0.30	0.22	0.30	0.22
3	1	3	1.00	1.00	1.00	0.99	2.92	0.48	0.28	0.27	0.28	0.27
3	1	4	1.00	1.00	1.00	0.99	2.24	0.63	0.26	0.25	0.26	0.25
3	1	5	1.00	1.00	1.00	1.00	3.21	1.35	0.32	0.22	0.32	0.22
3	1	6	0.83	0.81	0.84	0.80	1755.97	791.31	0.53	0.58	0.53	0.58
3	1	7	1.00	1.00	1.00	1.00	2.42	1.04	0.30	0.22	0.30	0.22
3	1	8	1.00	1.00	1.00	1.00	1.46	0.22	0.27	0.24	0.27	0.24
3	2	1	1.00	1.00	1.00	0.99	0.65	0.51	0.91	0.91	0.91	0.91
3	2	2	1.00	1.00	1.00	1.00	0.19	0.41	0.47	0.47	0.47	0.47
3	2	3	1.00	1.00	1.00	1.00	0.89	2.50	0.48	0.44	0.48	0.44
3	2	4	1.00	1.00	1.00	1.00	1.64	3.91	0.15	0.09	0.15	0.09
3	2	5	1.00	1.00	1.00	0.99	0.65	0.51	0.91	0.91	0.91	0.91
3	2	6	0.86	0.86	0.90	0.81	932.23	323.75	0.62	0.72	0.62	0.72
3	2	7	1.00	1.00	1.00	1.00	2.10	0.02	0.53	0.53	0.53	0.53
3	2	8	1.00	1.00	1.00	1.00	3.12	0.28	0.14	0.09	0.14	0.09
3	3	1	0.99	0.99	0.99	0.98	1.49	1.28	0.51	0.57	0.51	0.57
3	3	2	1.00	1.00	1.00	0.99	1.20	4.75	0.25	0.31	0.25	0.31
3	3	3	1.00	1.00	1.00	1.00	1.32	2.21	0.26	0.30	0.26	0.30
3	3	4	1.00	1.00	1.00	1.00	1.61	1.90	0.21	0.21	0.21	0.21
3	3	5	0.99	0.99	0.99	0.98	1.49	1.28	0.51	0.57	0.51	0.57
3	3	6	0.82	0.78	0.77	0.69	1235.39	1014.15	0.60	0.77	0.60	0.77
3	3	7	1.00	1.00	1.00	0.99	1.18	6.41	0.26	0.33	0.26	0.33
3	3	8	1.00	1.00	1.00	1.00	2.75	10.68	0.21	0.20	0.21	0.20
3	4	1	0.97	0.62	0.98	0.63	3154.60	124.85	1.43	0.90	1.43	0.90
3	4	2	0.97	0.62	0.98	0.62	3281.19	133.00	1.43	0.88	1.43	0.88
3	4	3	1.00	0.68	0.99	0.62	4684.50	223.34	1.44	0.81	1.44	0.81
3	4	4	1.00	0.99	1.00	1.00	1.50	0.23	0.27	0.00	0.27	0.00
3	4	5	0.97	0.62	0.98	0.63	3154.60	124.85	1.43	0.90	1.43	0.90
3	4	6	0.83	0.72	0.85	0.60	108.47	723.71	1.14	0.88	1.14	0.88
3	4	7	0.98	0.63	0.98	0.61	3252.78	133.14	1.42	0.90	1.42	0.90
3	4	8	1.00	0.99	1.00	1.00	0.60	0.08	0.27	0.00	0.27	0.00
3	5	1	0.97	0.61	0.97	0.59	1416.06	89.40	1.32	0.99	1.32	0.99
3	5	2	0.98	0.68	0.98	0.65	404.06	69.29	1.09	0.90	1.09	0.90
3	5	3	1.00	1.00	1.00	0.99	1.35	0.30	0.33	0.18	0.33	0.18
3	5	4	1.00	1.00	1.00	0.99	1.30	0.40	0.22	0.00	0.22	0.00
3	5	5	0.97	0.61	0.97	0.59	1416.06	89.40	1.32	0.99	1.32	0.99
3	5	6	1.00	0.86	1.00	0.87	4.72	425.08	0.33	0.45	0.33	0.45
3	5	7	1.00	0.62	1.00	0.60	62.54	96.29	1.17	0.98	1.17	0.98
3	5	8	1.00	1.00	1.00	1.00	2.30	0.79	0.22	0.00	0.22	0.00
3	6	1	0.30	0.46	0.31	0.46	384.04	159.29	1.53	1.43	1.53	1.43
3	6	2	0.30	0.45	0.31	0.45	389.42	156.69	1.52	1.43	1.52	1.43
3	6	3	1.00	1.00	1.00	1.00	1.49	0.34	0.43	0.34	0.43	0.34
3	6	4	1.00	1.00	1.00	1.00	1.61	0.68	0.20	0.00	0.20	0.00
3	6	5	0.30	0.46	0.31	0.46	384.04	159.29	1.53	1.43	1.53	1.43
3	6	6	0.87	0.80	0.83	0.80	564.07	362.96	0.62	0.66	0.62	0.66
3	6	7	0.30	0.45	0.29	0.43	375.30	152.88	1.54	1.44	1.54	1.44
3	6	8	1.00	0.95	1.00	0.98	1.23	6.55	0.19	0.15	0.19	0.15
3	7	1	0.23	0.35	0.29	0.43	395.83	231.61	1.55	1.48	1.55	1.48
3	7	2	0.24	0.39	0.32	0.49	412.43	223.62	1.54	1.45	1.54	1.45
3	7	3	1.00	0.99	1.00	0.99	0.55	2.30	0.13	0.31	0.13	0.31
3	7	4	1.00	1.00	1.00	0.99	1.11	1.83	0.00	0.02	0.00	0.02
3	7	5	0.23	0.35	0.29	0.43	395.83	231.61	1.55	1.48	1.55	1.48
3	7	6	0.75	0.77	0.79	0.77	2121.97	552.10	0.51	0.66	0.51	0.66
3	7	7	0.23	0.38	0.30	0.44	397.27	217.56	1.54	1.50	1.54	1.50
3	7	8	1.00	1.00	1.00	1.00	0.03	1.53	0.00	0.00	0.00	0.00

The result of simulation 2 when Σ is formed by the third choice above is presented in Table 5. The “cov” column indicates the choice of Σ , the “std” column indicates the type of sampling distribution, and the “mdt” column indicates the CCA methods. Although $p = q = 5$ in this case, only the results of first two canonical covariates are listed due to the limit of the space. Table 5 shows that the $MSE(\rho_k)$ of classical CCA (as well as classical PP) increases rapidly when the point mass outliers are introduced. For the normal mixture sampling distribution, only PP-MCD does not work well. For the mixture distribution $.8N_{p+q}(0, \Sigma) + .2\delta(tr(\Sigma)\mathbf{1}')$, only RMVN and PP-RMVN CCA perform well. The result of the mixture distribution $.8N_{p+q}(0, \Sigma) + .2\delta(tr(\Sigma) * [1, 0, \dots, 0]')$ is quite similar. In general, it is observed that RMVN and PP-RMVN have the best performance when the underlying distribution is multivariate normal. Between them, the CCA based on RMVN approach should be adopted since it has the computational efficiency advantage.

6. Discussion

Robust outlier resistant estimators of MLD should be i) $\sqrt{n}$ consistent for a large class of distributions, ii) easy to compute, iii) effective at detecting certain types of outliers and iv) outlier resistant. Although Hawkins and Olive (2002) showed that almost all of the literature focuses either on i) and iv) or on ii) and iii), Olive and Hawkins (2010) shows that it is simple to construct estimators satisfying i)-iv) provided that $n > 20p$ and $p \leq 40$. These results represent both a computational and theoretical breakthrough in the field of robust MLD.

The new FCH, RFCH and RMVN estimators use information from both location and dispersion criteria and are more effective at screening attractors than estimators such as MBA and FMCD that only use the MCD dispersion criterion. The new estimators are roughly two orders of magnitude faster than FMCD.

The collection of easily computed “robust estimators” for MLD that have not been shown to be both HB and consistent is enormous, but without theory the methods should be classified as outlier diagnostics rather than robust statistics.

Examine the estimator on many “benchmark data sets.” FCH was examined on 30 such data sets. Outlier performance was competitive with estimators such as FMCD. For any given estimator, it is easy to find outlier configurations where the estimator fails. For the modified wood data of Rousseeuw (1984), MB detected the planted outliers but FCH used DGK. For another data set, 2 clean cases had larger MB distances than 4 of 5 planted outliers that FMCD can detect. For small p , elemental methods can be used as outlier diagnostics.

Acknowledgements

The authors thank Douglas M. Hawkins, the editor and two referees for their comments that improved this article.

References

- Alkenani, A., & Yu, K. (2012). A comparative study for robust canonical correlation methods. *Journal of Statistical Computation and Simulation*, to appear.
- Branco, J. A., Croux, C., Filzmoser, P., & Oliveira, M. R. (2005). Robust canonical correlation: a comparative study. *Computational Statistics*, 20, 203-229. <http://dx.doi.org/10.1007/BF02789700>
- Butler, R. W., Davies, P. L., & Jhun, M., (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21, 1385-1400. <http://dx.doi.org/10.1214/aos/1176349264>
- Buxton, L. H. D. (1920). The anthropology of Cyprus. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 50, 183-235. <http://dx.doi.org/10.2307/2843379>
- Cator, E. A., & Lopuhaä, H. P. (2010). Asymptotic expansion of the minimum covariance determinant estimators. *Journal of Multivariate Analysis*, 101, 2372-2388. <http://dx.doi.org/10.1016/j.jmva.2010.06.009>
- Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76, 354-362. <http://dx.doi.org/10.1080/01621459.1981.10477654>
- Hawkins, D. M., & Olive, D. J. (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics & Data Analysis*, 30, 1-11. [http://dx.doi.org/10.1016/S0167-9473\(98\)00082-6](http://dx.doi.org/10.1016/S0167-9473(98)00082-6)
- Hawkins, D. M., & Olive, D. J. (2002). Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm (with discussion). *Journal of the American Statistical Association*, 97, 136-159. <http://dx.doi.org/10.1198/016214502753479293>

- Johnson, M. E. (1987). *Multivariate statistical simulation*. New York, NY: John Wiley & Sons.
- Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Lopuhaä, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *The Annals of Statistics*, 27, 1638-1666. <http://dx.doi.org/10.1214/aos/1017939145>
- Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 50, 295-304. <http://dx.doi.org/10.1198/004017008000000190>
- Muirhead, R. J., & Waternaux, C. M. (1980). Asymptotic distribution in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika*, 67, 31-43. <http://dx.doi.org/10.1093/biomet/67.1.31>
- Olive, D. J. (2004). A resistant estimator of multivariate location and dispersion. *Computational Statistics and Data Analysis*, 46, 99-102. [http://dx.doi.org/10.1016/S0167-9473\(03\)00119-1](http://dx.doi.org/10.1016/S0167-9473(03)00119-1)
- Olive, D. J. (2008). *Applied robust statistics*. Online text retrieved from <http://www.math.siu.edu/olive/ol-bookp.htm>
- Olive, D. J. (2012). *Robust multivariate analysis*. Unpublished manuscript retrieved from <http://www.math.siu.edu/olive/multbk.htm>
- Olive, D. J., & Hawkins, D. M. (2010). Robust multivariate location and dispersion. Preprint at <http://www.math.siu.edu/olive/preprints.htm>
- Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91, 1047-1061. <http://dx.doi.org/10.1080/01621459.1996.10476975>
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-880. <http://dx.doi.org/10.1080/01621459.1984.10477105>
- Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212-223. <http://dx.doi.org/10.1080/00401706.1999.10485670>
- UCLA: Academic Technology Services. (2011). R data analysis examples, canonical correlation analysis. <http://www.ats.ucla.edu/stat/R/dae/canonical.htm>
- Zhang, J. (2011). *Applications of a robust dispersion estimator*, Ph. D. Thesis, Southern Illinois University. <http://www.math.siu.edu/olive/szhang.pdf>