Outlier Resistant Estimators for Canonical Correlation Analysis

P. Filzmoser[1], C. Dehon[2] and C. Croux[2]

[1] Department of Statistics, Probability Theory, and Actuarial Mathematics, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria
[2] ECARES, Université Libre de Bruxelles, CP139, Av. F.D. Roosevelt 50, B-1050 Bruxelles, Belgium

Abstract. Canonical correlation analysis studies associations between two sets of random variables. Its standard computation is based on sample covariance matrices, which are however very sensitive to outlying observations. In this note we introduce, discuss and compare different ways for performing a robust canonical correlation analysis. Two methods are based on robust estimators of covariance matrices, the others on projection-pursuit techniques.

Keywords. Canonical correlation, minimum covariance determinant estimator, projection-pursuit, robustness.

1 Introduction

The aim of Canonical Correlation Analysis (CCA) is to identify and quantify the relations between a p-dimensional random variable $\boldsymbol{X}$ and a q-dimensional random variable $\boldsymbol{Y}$. Herefore we look for linear combinations $a^t\boldsymbol{X}$ and $b^t\boldsymbol{Y}$ of the original variables having maximal correlation. Expressed in mathematical terms, CCA seeks for vectors $\alpha \in I\!R^p$ and $\beta \in I\!R^q$ such that

$$(\alpha, \beta) = \underset{a,b}{\operatorname{argmax}} |\operatorname{Corr}(a^t\boldsymbol{X}, b^t\boldsymbol{Y})|. \tag{1}$$

The resulting univariate variables $U = \alpha^t\boldsymbol{X}$ and $V = \alpha^t\boldsymbol{X}$ are then called the *canonical variates* and can be used for dimension reduction and graphical display. Note that the vectors α and β are only determined up to a constant factor by definition (1). The first canonical correlation ρ is defined as the absolute value of the correlation between the two canonical variates, which equals the maximum attained in (1).

Higher order canonical variates and correlations are defined as in (1), but now under the additional restriction that a canonical variate of order k, with $1 < k \leq \min(p, q)$, should be uncorrelated with all canonical variates of lower order. Due to space limitations, we restrict attention to a first order canonical analysis.

The above CCA problem (1) has a fairly simple solution (see e.g. Johnson and Wichern, 1998, Chapter 10). Denote by Σ the population covariance matrix of the random variable $\boldsymbol{Z} = (\boldsymbol{X}^t, \boldsymbol{Y}^t)^t$. We decompose Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

The vectors α and β are now the eigenvectors corresponding to the largest eigenvalues of the matrices

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \quad \text{and} \quad \Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \tag{2}$$

Both of the above matrices have the same positive eigenvalues and the largest one equals the squared first canonical correlation.

For estimating the unknowns α, β, and ρ one typically computes the sample covariance matrix $\hat{\Sigma}$ from a sample $z_1, \ldots, z_n$, with $z_i = (x_i^t, y_i^t)^t \in I\!R^p \times I\!R^q$. Computing eigenvectors/values of the empirical counterparts of the matrices in (2) results then immediately in estimates of the canonical variates and correlations. Since the classical estimator of a covariance matrix is very vulnerable with respect to outlying observations, also the eigenvalues and -vectors based on $\hat{\Sigma}$ will be very sensitive, as was shown in the context of CCA by Romanazzi (1992). In Section 2 of this paper an approach based on robust estimators of the covariance matrix is outlined and illustrated with a real data example. Two other approaches to robust CCA, which are in the spirit of projection-pursuit, are proposed and discussed in Section 3. They are compared by means of a modest stability study in Section 4. A more comprehensive study of the different approaches is part of current research of the authors.

2 Robust CCA based on Robust Covariance Matrices

2.1 Using the Minimum Covariance Determinant Estimator

The obvious way for robustifying CCA is to estimate Σ robustly and to compute eigenvectors/values from the estimated version of (2) in the usual way. Theoretical results for this approach have been obtained by Croux and Dehon (2000). As robust covariance estimators one could use M-estimators as in Karnel (1991), but it is known that these estimators have poor robustness properties in higher dimensions. A more appropriate choice is the *Minimum Covariance Determinant* (MCD) estimator of Rousseeuw (1985). The MCD estimator is obtained by looking for that subset of size h of the data which has the smallest value of the determinant of the empirical covariance matrix computed from it. Maximal robustness is obtained for $h \approx n/2$. The resulting estimator is then nothing else but the covariance matrix computed over that optimal subset. An efficient algorithm for computing the MCD estimator has been proposed by Rousseeuw and Van Driessen (1999).

Robust covariance matrix estimators can routinely be used in multivariate statistics. Filzmoser (1999) applied them for robust factor analysis of geostatistical data.

To illustrate the usefulness of a robust CCA, we applied the MCD-based approach to the "Diabetes data" (Andrews and Herzberg 1985, data set 36, page 215). For a group of $n = 76$ normal persons, the variables Glucose intolerance (X_1), Insulin Response to Oral Glucose (X_2), Insulin Resistance (X_3), Relative Weight (Y_1) and Fasting Plasma Glucose (Y_2) were measured. It is of medical interest to establish a relation between the X and the Y variables. The classical estimators for the eigenvectors are $\hat{a}^{cl} = (-0.32, 0.47, -1.04)^t$ and $\hat{b}^{cl} = (0.98, 0.06)^t$, compared to $\hat{a}^{rob} = (-0.11, 0.46, -1.06)^t$ and $\hat{b}^{rob} = (0.83, 0.36)^t$ for the robust estimates. From these estimates, a scatter plot of the scores of x_i and y_i $(1 \leq i \leq n)$ on the first canonical variates for X

and Y is constructed in Figure 1. Corresponding regression fits are indicated and the estimates for ρ are given by $\hat{\rho}^{cl} = 0.50$ and $\hat{\rho}^{rob} = 0.71$. From this, we conclude that the robust estimate found a relation between the X and Y variables which is well followed by a huge majority of the data. The classical approach tries to find an association between X and Y which is valid for all data points, outliers included, and thereby leading to a much weaker first canonical correlation. It was also verified that the robust first order canonical variates takes 97% of the total correlation between X and Y into account (defined as the first eigenvalue divided by the sum of all eigenvalues of the estimated matrices in (2)), while this was only 80% for the classical approach.

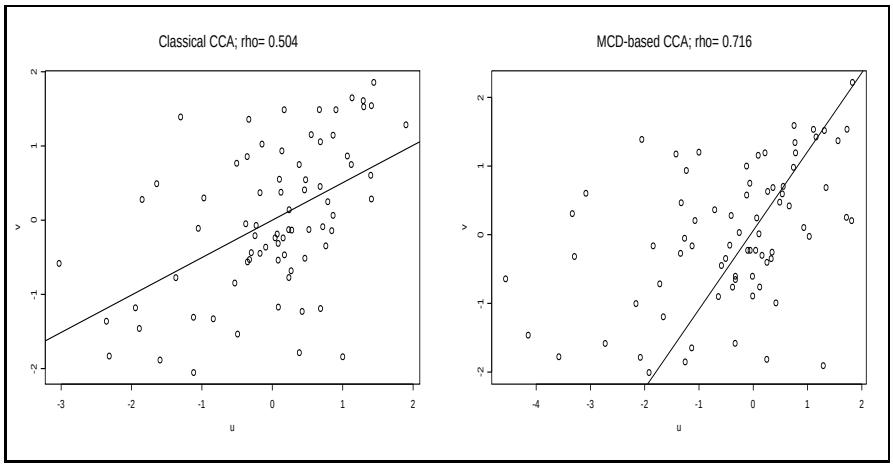

Fig. 1. Scatter plot of the first canonical variate b^tY versus a^tX for the classical estimator (left) and the MCD-based estimator (right).

2.2 Using the Signs of the Observations

The MCD-estimator requires that the number of observations is twice as much as the number of variables present in the data set. But in practical applications (e.g. in psychology, signal processing, etc.) it occurs often that one has a lot of variables but only a limited number of observations.

As an alternative, sign-based covariance matrices (Visuri, Koivunen, and Oja, 2000) can be used. For computing the sign of an observation x_i (and similarly for y_i), we first need to compute the spatial median μ_X of the data cloud $x_1, \ldots, x_n$. The spatial median is defined as

$$\mu_X = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} ||x_i - \mu||,$$

with $||\cdot||$ the Euclidean norm. The sign of x_i is then the projection of x_i on a unit sphere centered at μ_X:

$$S(x_i) = \frac{x_i - \mu_X}{||x_i - \mu_X||}.$$

Note that the signs are vector valued. By projecting observations on this unit sphere the influence of outliers is heavily reduced, leading to an outlier resistant procedure. Afterwards, ordinary covariance matrices are computed from these signs. Since fast iterative algorithms to compute μ_X and μ_Y exist and the signs can be computed in $O(n)$ time, the resulting procedure will be extremely fast. However, the statistical efficiency of the method can be quite low, since a lot of information is lost by only taking the direction of the data points into account.

3 Robust CCA based on Projection-Pursuit

Projection-Pursuit (PP) techniques for CCA start from the initial definition (1) of CCA. We are looking for two directions a and b which maximize the projection-pursuit index $|\mathrm{Corr}(a^t\boldsymbol{X}, b^t\boldsymbol{Y})|$. Taking as an estimate for the population correlation an ordinary correlation coefficient, yields of course the classical non robust approach. The idea is therefore to work with a robust projection-pursuit index.

The methods outlined below have the feature that they allow to compute just the first few canonical variates, without using an estimate of Σ. For high dimensions, where $p+q$ is huge, this is an important advantage.

3.1 Using the Spearman Correlation Index

Since the correlation index in (1) is between two univariate variables, a simple Spearman rank correlation can be used to measure the correlation between a^tX and a^tY. By working with ranks, the influence of outliers will be mitigated.

In practice, it is not obvious how to find the vectors α and β maximizing (the absolute value of) the Spearman correlation index. A simple and fairly good approximation is obtained by restricting the search to the finite set $\{(a_i, b_j)|1 \leq i, j \leq n\}$, where a_i is the normed vector $x_i - \mu_X$, μ_X a robust location measure of the X-population (e.g. the spatial median defined in Section 2.2 or the coordinate-wise median) and b_j the normed $y_j - \mu_Y$.

Although the non-parametric nature of the Spearman correlation is very appealing, the computational complexity of $O(n^3 \log n)$ may become prohibitive for bigger sample sizes.

3.2 Using Robust Alternating Regressions

Application of the alternating regressing technique to CCA was already proposed by Wold (1966). Its use is motivated by the observation that, for a given α,

$$\beta = \underset{b}{\mathrm{argmax}}\, |\mathrm{Corr}(\alpha^t\boldsymbol{X}, b^t\boldsymbol{Y})|.$$

But then it follows from standard results on multiple regression, that β is proportional to the regression coefficient b in the model

$$\alpha^tX = b^tY + \gamma_1 + \varepsilon_1. \tag{3}$$

In the same way, for a given β, the optimal α equals (up to a scalar term) the parameter a in the regression model

$$b^tY = a^tX + \gamma_2 + \varepsilon_2. \tag{4}$$

Start now with an initial value α^0. (This can for example be obtained by performing a robust principal components analysis on the data matrix formed

by $x_1, \ldots, x_n$.) According to (3), we get a first β^1 by regressing the univariate $X^t\alpha^0$ on the Y variables. Afterwards, using (4), an updated α^1 is obtained by regressing $Y^t\beta^1$ on X. This procedure is then iterated until convergence. The estimated regression coefficients are normalized in each step, and computing a bivariate (robust) correlation coefficient between the estimated canonical variates yields an estimator of ρ.

To be outlier resistant, the regression estimators in the above alternating regression scheme need to be robust. Since they are computed several times, a fast, but reliable estimator should be chosen. We propose to use a weighted L_1-estimator, as was motivated by Croux and Filzmoser (1998) in an application of alternating regressions to two-way tables.

4 Stability Experiment

In this section all proposed methods of Section 2 are compared by a small statistical experiment. We generated a data set $Z = \{z_1, \ldots, z_n\}$ with $z_i = (x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^q$ from a specified $N(0, \Sigma)$ distribution, and with $n = 30$, $p = 2$, and $q = 3$. Since we are interested in the behavior of the estimators when outliers are present, we added randomly generated noise from $N(0, 50I_p)$ to the first 5 observations. This means that 16% of all observations are contaminated.

A tool for assessing the robustness of an estimator is the *empirical influence function* (EIF). Here we compute the EIF of every observation z_i on the estimator $\hat{\rho}$ of the the first canonical correlation coefficient. By definition

$$\mathrm{EIF}(z_i, \hat{\rho}) = \hat{\rho}(Z \setminus \{z_i\}) - \hat{\rho}(Z),$$

so it measures the effect of deleting the observations z_i on the estimator. In the spirit of robustness, we do not want single points to have a too high influence on the estimator. Figure 2a plots $\mathrm{EIF}(z_i, \hat{\rho})$ versus the index of each observation for the classical estimator, the MCD-based estimator and the Sign-based estimator. The non robustness of the classical estimator appears by the extremely high value for the fifth point. The MCD-based estimator seems to be more robust for this example than the Sign-based method: the empirical influence function is indeed flatter. Figure 2b compares the EIF of the MCD with the two projection pursuit procedures. The MCD-based method remains the most robust, closely followed by the Spearman and the robust alternating regression method. Note the different scale of the vertical axis in Figure 2a and 2b.

As a final conclusion, making a choice between the available robust procedures is quite difficult at the present state. More theory needs to be developed and more practical experience is necessary. On the basis of several experiments we performed, it looks as if the MCD-based procedure is performing quite well in cases were the number of observations is high enough. Otherwise, the robust alternating regression method is an alternative.

References

Andrews, D.F. and Herzberg, A.M. (1985). Data: a collection of problems from many fields for the student and research worker. New York: Springer-Verlag.

Croux, C. and Dehon, C. (2000). Robust Canonical Correlations using High Breakdown Scatter Matrices. Preprint, Université Libre de Bruxelles.

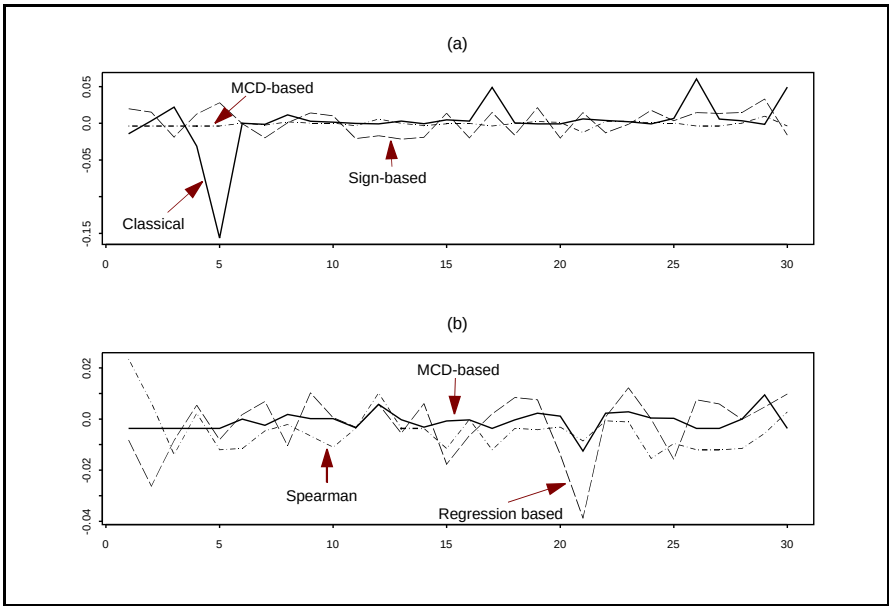

Fig. 2. Empirical influence functions for (a) the classical method, the MCD-based method and the Sign-based method; (b) the Spearman correlation based method, the robust alternating regression estimator and the MCD-based method.

Croux, C. and Filzmoser, P. (1998). A Robust Biplot Representation of Two-way Tables. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*, 355-361. Berlin: Springer-Verlag.

Filzmoser, P. (1999). Robust Principal Component and Factor Analysis in the Geostatistical Treatment of Environmental Data. *Environmetrics*, **10**, 363-375.

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis: Fourth Edition.* New Jersey: Prentice Hall.

Karnel, G. (1991). Robust Canonical Correlation and Correspondence Analysis. *The Frontiers of Statistical Scientific Theory & Industrial Applications*, 335-354.

Romanazzi, M. (1992). Influence in Canonical Correlation Analysis. *Psychometrika*, **57**, 237-259.

Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point. In: W. Grossmann et al. (Eds.): *Mathematical Statistics and Applications, Vol. B*, 283-297. Dordrecht: Reidel.

Rousseeuw, P.J. and Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**, 212-223.

Visuri, S., Koivunen, V., and Oja, H. (2000). Sign and Rank Correlation Matrices. *Journal of Statistical Planning and Inference*. To appear.

Wold, H. (1966). Nonlinear Estimation by Iterative Least Squares Procedures. In: F.N. David (Ed.): *A Festschrift for J. Neyman*, 411-444. New York: Wiley and Sons.