Robust Principal Components using Projection-Pursuit

Christophe Croux
ECARES, Université Libre de Bruxelles
CP 114, Av. F.D. Roosevelt 50
B-1050 Brussels, Belgium
ccroux@ulb.ac.be

1. Projection-Pursuit based robust multivariate statistics

While robust regression and estimation of location and scatter are well explored areas of statistical research, there still remains work to do for the robustification of other techniques in multivariate statistics. Techniques like principal components analysis (PCA), canonical correlation analysis (CCA), and discriminant analysis (DA) are all based on estimation of the population covariance matrix. An obvious way to obtain robust versions of them is therefore to use a resistant estimator of the population covariance matrix Σ. When using an affine equivariant estimator of Σ, having "regular" asymptotic behavior, it is not too difficult to derive influence functions for the estimators of the eigenvectors in a PCA, the canonical vectors in a CCA, or the parameters of the linear discriminant function based on this robust, preferably high breakdown, affine equivariant estimator of Σ (see Croux and Haesbroeck 2000, Croux and Dehon 2001a, 2001b). They appear to be bounded as soon as the estimator of Σ has a redescending influence function.

There are however 2 drawbacks related to this approach: the first one is that affine equivariant high breakdown estimators require that the number of observations n exceeds the number of variables p. There are existing fields of applications, like spectroscopy and functional data analysis, where this condition is not met. The second drawback is that it these robust estimators of scatter are only consistent estimators of Σ when the underlying distribution has a certain degree of symmetry, like elliptical symmetry or belonging to a location-scale family, as in Visuri et al (2000). It is therefore worth to consider an approach based on the initial definitions of PCA, CCA and DA, without passing by the estimation of Σ. By looking in the classical textbooks on multivariate statistics (e.g. Johnson and Wichern 1998), one observes that many multivariate statistical techniques are defined by maximizing a certain projection index. It has been suggested by Huber (1985) that robust versions can be obtained by working with a more robust projection index, thereby obtaining a *Projection-Pursuit* (PP) based method.

In principal component analysis, for example, the first principal component associated with a random vector $X \in I\!R^p$ is defined by maximizing

$$\mathrm{Var}(a^t X), \tag{1}$$

over all unit norm vectors a. In discriminant analysis, we have a second variable Y of the same dimension as X, and we try to find the direction giving optimal discrimination

$$\frac{|\mathrm{Ave}(a^tX) - \mathrm{Ave}(a^tY)|}{\sqrt{p_1\mathrm{Var}(a^tX)) + p_2\mathrm{Var}(a^tY)}}, \tag{2}$$

for suitable constants $p_1 + p_2 = 1$. For canonical correlation analysis, we have Y of dimension q, and try to maximize

$$\mathrm{Corr}(\alpha^tX, \beta^tY). \tag{3}$$

Note that these averages, variances, and correlations are all computed from univariate random variables, so there is no need for multivariate estimators of covariance. By using robust location and scale estimators, or robust bivariate correlation measures in the definitions above, we will obtain robust PP-based methods (see, among others, Li and Chen 1985, for PCA, Posse 1992, for DA, and Oliveira and Branco 2000, for CCA).

While intuitively very attractive, the PP-based methods also have some drawbacks. It has been shown (Croux and Ruiz-Gazen 2000, Croux and Filzmoser 2001, Pires and Branco 2001) that the corresponding influence functions are unbounded, thereby showing a lack of local robustness. The global robustness properties, however, seem to be much better (e.g. Boente and Orellana 2000). Another drawback of the method is that it is not obvious how to compute the PP-based estimators, since they are defined as solution of a maximization problem under constraints.

2. A fast algorithm for PP-based Principal Components Analysis

In principal component analysis based on Projection-Pursuit we need to maximize

$$S(a^tX), \tag{4}$$

over all unit norm vectors a, for S a specified robust measure of spread. Typically we will take for S a high breakdown estimator of scale. Solving the above maximisation problem is only trivial for S equal to the (square root of the) variance. Some time ago, an approximative algorithm has been proposed by Croux and Ruiz-Gazen (1996) which is fast and accurate for the first few components and p not being too large in comparison with the smaple size. This algorithm, let us call it the BASIC algorithm, has been applied by Filzmoser (1999), Boente and Orellana (2000), and Gather et al (1998) in different contexts. For high-dimensional data sets, and typically for the case $p > n$, the algorithm has been adapted by Verboven et al (2000). A problem with their algorithm is that it will systematically underestimate the eigenvalues, and even give rise to a breakdown of the procedure (in the sense that the estimates for the higher order eigenvalues break down to zero). Moreover, it will give sub-optimal solutions: the direction a found by this approximative algorithm explains much less dispersion than it should

be. Recently, a better improvement of the BASIC algorithm was developed (and MATLAB-code can be obtained from the author upon request). This new algorithm performs local improvement steps around initial estimates, and is inspired upon the classical algorithm for finding the largest eigenvalue of a symmetric matrix. We show that this new algorithm is very fast, and finds good approximations of the solution of the above maximisation procedure. Moreover, it is applicable to high dimensional data sets.

REFERENCES

Boente, G., and L. Orellana, L. (2001). A Robust Approach to Common Principal Components. To appear in *Statistics in Genetics and Environmental Sciences.* L. T. Fernholtz, S. Morgenthaler, and W. Stahel, editors. Birkhauser Verlag AG, Basel, Swizterland.

Croux, C. and Dehon, C. (2001a). Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. To appear in *La Revue de Statistique Appliquée.*

Croux, C. and Dehon, C. (2001b). The Most Robust Estimator for Linear Discriminant Analysis. To appear in *The Canadian Journal of Statistics*, September issue.

C. Croux, C. and Haesbroeck, G. (2000). Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika*, 71, 161–190.

Croux, C. and Ruiz-Gazen, A. (1996). A Fast Algorithm for Robust Principal Components based on Projection Pursuit. in *Compstat: Proceedings in Computational Statistics*, ed. A. Prat, Heidelberg: Physica-Verlag. 211–217.

Croux, C. and Ruiz-Gazen, A. (2000). "High Breakdown Estimators for Principal Components: the Projection-Pursuit Approach Revisited," *under revision,* http://homepages.ulb.ac.be/~ccroux.

Croux, C. and Filzmoser, P. (2001). "A Projection-Pursuit based Measure of Association between two Multivariate Variables. In preparation.

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis, fourth edition* Prentice Hall International Editions.

Gather, U., Hilker, T., and Becker, C. (2001). A Robustified Version of Sliced Inverse Regression. To appear in *Statistics in Genetics and Environmental Sciences.* L. T. Fernholtz, S. Morgenthaler, and W. Stahel, editors. Birkhauser Verlag AG, Basel, Swizterland.

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

Li, G. and Chen, Z. (1985). Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: primary theory and Monte Carlo. *Journal of the American Statistical Association,* 80, 759–766.

M. R. de Oliveira, M.R. and Branco, J.A. (2000). Projection Pursuit Approach to Robust Canonical Correlation Analysis. In *Compstat: Proceedings in Computational Statistics*, ed. J.G. Bethlehem, and P.G.M. van de Heijden, Heidelberg: Physica-Verlag. 211–217.

Pires, A.M. and Branco, J.A.] (2001). Projection-Pursuit Approach for Robust Linear Discriminant Analysis. Preprint, Instituto Superior Tecnico, Dept. of Math.

Posse, C. (1992). Projection Pursuit Discriminant Analysis for two Groups. *Communications in Statistics - Theory and Methods*, 21, 1-19.

Verboven, S., Rousseeuw, P.J., and Hubert, M. (2000). An Improved Algorithm for Robust PCA. In *Compstat: Proceedings in Computational Statistics*, ed. J.G. Bethlehem, and P.G.M. van de Heijden, Heidelberg: Physica-Verlag. 211–217.

S. Visuri, S., Koivunen, V., Möttönnen, J., Ollila, E. and Oja, H. (2001). Affine Equivariant Multivariate Rank Methods. To appear in *Journal of Statistical Planning and Inference.*

RESUME

Plusieurs méthodes existent pour robustifier des techniques d'analyse multivariée. La plupart des méthodes sont basées sur l'estimation robuste de la matrice de variance-covariance, mais c'est également possible d'utiliser le principe de "projection-pursuit." Nous comparons les différentes approches, et plus particulierement pour l'analyse en composants robustes, nous présentons un algorithm rapide.