A Projection Method for Robust Estimation and Clustering in Large Data Sets

Daniel Peña[1] and Francisco J. Prieto[2]

[1] Departamento de Estadística,
Universidad Carlos III de Madrid, Spain
daniel.pena@uc3m.es
[2] Departamento de Estadística,
Universidad Carlos III de Madrid, Spain
franciscojavier.prieto@uc3m.es

Abstract. A projection method for robust estimation of shape and location in multivariate data and cluster analysis is presented. The key idea of the procedure is to search for heterogeneity in univariate projections on directions that are obtained both randomly, using a modification of the Stahel-Donoho procedure, and by maximizing and minimizing the kurtosis coefficient of the projected data, as proposed by Peña and Prieto (2005). We show in a Monte Carlo study that the resulting procedure works well for robust estimation. Also, it preserves the good theoretical properties of the Stahel-Donoho method.

1 Introduction

As a few outliers in multivariate data may distort arbitrarily the sample mean and the sample covariance matrix, the robust estimation of location and shape is a crucial problem in multivariate statistics. See for instance Atkinson et al. (2004) and the references therein. For high-dimensional large data sets a useful way to avoid the curse of dimensionality in data mining applications is to search for outliers in univariate projections of the data. Two procedures that use this approach are the Stahel-Donoho procedure, that searches for univariate outliers in projections on random directions, and the method proposed by Peña and Prieto (2001, b), that searches for outliers in projections obtained by maximizing and minimizing the kurtosis coefficient of the projected data. The first procedure has good theoretical properties, but fails for concentrated contamination and requires prohibitive computer times for large dimension problems. The second procedure works very well for concentrated contamination and it can be applied in large dimension problems, but its theoretical properties are unknown. As both procedures are based on projections, it seems sensible to explore if a combination of both could avoid their particular limitations and this has been proposed by Peña and Prieto (2005). They show that the combination of random and specific directions leads to a affine equivariant procedure which inherits the good theoretical properties of the Stahel-Donoho method and it is fast to compute so that it can be applied for large data sets.

The procedure can also be applied for cluster analysis by generalizing the approach presented in Peña and Prieto (2001,a). Then, instead of just searching for directions which are extremes of the kurtosis coefficient, we add random directions to obtain a better exploration of the space of the data.

This article summarizes the method proposed by Peña and Prieto (2005) and includes two contributions: we present new results on the relative performance of the procedure with several groups of outliers, and we discuss the application of the procedure for cluster analysis. The article is organized as follows. Section 2 summarizes the main ideas of the procedure for generating directions and presents the algorithm combining random and specific directions. Section 3 discuss the extensions of these ideas for clustering . Section 4 illustrates the performance of the proposed method as an outlier detection tool for robust estimation.

2 Finding Interesting Directions

Suppose we have a sample $(x_1, \ldots, x_n)$ of a p-dimensional vector random variable X. We are interested in searching for heterogeneity by projecting the data onto a set of directions d_j, $j = 1, \ldots, J$. The key step of the method is obtaining the directions d_j. The Stahel-Donoho procedure is based on generating these directions randomly: a random sample of size p is chosen, a hyperplane is fitted to this sample and the direction d_j orthogonal to this hyperplane is chosen. Note that if we have a set of outliers and the data is standardized to have variance equal to one, the direction orthogonal to the fitted plane is, a priori, a good one to search for outliers.

A procedure for obtaining specific directions that can reveal the presence of heterogeneity was proposed by Peña and Prieto (2001b). They showed that the projection of the data on the direction of the outliers will lead to (1) a distribution with large univariate kurtosis coefficient if the level of contamination is small and (2) a distribution with small univariate kurtosis coefficient if the level of contamination is large. In fact, if the data come from a mixture of two distributions $(1-\alpha)f_1(X) + \alpha f_2(X)$, with $.5 < \alpha < 0$ and f_i, $i = 1, 2$, is an elliptical distribution with mean μ_i and covariance matrix V_i, the directions that maximize or minimize the kurtosis coefficient of the projected data are of the form of the admissible linear classification rules. In particular, if the distributions were normal with the same covariance matrix and the proportion of contamination is not large, $0 < \alpha < 0.21$, the direction obtained by maximizing the kurtosis coefficient is the Fisher linear discriminant function whereas when the proportion of contamination is large, $0.21 < \alpha < .5$, the direction which minimizes the kurtosis coefficient is again the Fisher linear discriminant function. Thus, the extreme directions of the kurtosis coefficient seem to provide a powerful tool for searching for groups of masked outliers. Peña and Prieto (2001b) proposed an iterative procedure based on the projection on a set of $2p$ orthogonal directions obtained as extremes for the kurtosis

of the projected data. Note that the first set of p directions are closely related to the independent components of the data, which are defined as a set of p variables obtained by linear transformations of the original data such that the new variables are as independent as possible. It can be shown that the independent components can be obtained by maximizing the absolute value or the square of the kurtosis coefficient and, as this coefficient cannot be smaller than one, these directions will be the same as the one obtained by maximizing the kurtosis coefficient. The performance of these directions for outlier detection was found to be very good for concentrated contamination but, as it can be expected from the previous results, it was not so good when the proportion of contamination is close to .3 and the contaminating distribution has the same variance as the original distribution. This behavior of the algorithm is explained because then the values of the kurtosis for the projected data are not expected to be either very large or very small.

Thus it seems that we may have a very powerful procedure by combining the specific directions obtained as extremes of the kurtosis with some random directions. However, as we are interested in a procedure that works in large data sets and it is well known (and it will be discussed in the next section) that the Stahel-Donoho procedure requires a huge number of directions to work as the sample size increases, the random directions are not generated by random sampling, but by using some stratified sampling scheme that is found to be more useful in large dimensions. The univariate projections onto these directions are then analyzed as previously described in a similar manner to the Stahel-Donoho algorithm. See Peña and Prieto (2005) for the justification of the method.

We assume that first the original data are scaled and centered. Let $\bar{x}$ denote the mean and S the covariance matrix of the original data, the points are transformed using $y_i = S^{-1/2}(x_i - \bar{x}), \quad i = 1, \ldots, n.$

Stage I: Analysis based on directions computed from finding extreme values of the kurtosis coefficient. Compute n_1 orthogonal directions and projections maximizing the kurtosis coefficient ($1 \leq n_1 \leq p$) and n_2 directions minimizing this coefficient ($1 \leq n_2 \leq p$).

1. Set $y_i^{(1)} = y_i$ and the iteration index $j = 1$.

2. The direction that maximizes the coefficient of kurtosis is obtained as the solution of the problem

$$\begin{aligned} d_j = \arg\max_d \ & \frac{1}{n} \sum_{i=1}^{n} \left(d' y_i^{(j)} \right)^4 \\ \text{s.t.} \quad & d'd = 1. \end{aligned} \tag{1}$$

3. The sample points are projected onto a lower dimension subspace, orthogonal to the direction d_j. Define

$$v_j = d_j - e_1, \quad Q_j = \begin{cases} I - \dfrac{v_j v_j'}{v_j' d_j} & \text{if } v_j' d_j \neq 0 \\ I & \text{otherwise,} \end{cases}$$

where e_1 denotes the first unit vector. The resulting matrix Q_j is orthogonal, and we compute the new values

$$u_i^{(j)} \equiv \begin{pmatrix} z_i^{(j)} \\ y_i^{(j+1)} \end{pmatrix} = Q_j y_i^{(j)}, \quad i = 1, \ldots, n,$$

where $z_i^{(j)}$ is the first component of $u_i^{(j)}$, which satisfies $z_i^{(j)} = d_j' y_i^{(j)}$ (the univariate projection values), and $y_i^{(j+1)}$ corresponds to the remaining $p - j$ components of $u_i^{(j)}$.
We set $j = j + 1$, and if $j < n_1$ we go back to step 1(b). Otherwise, we let $z_i^{(p)} = y_i^{(p)}$.
4. The same process is applied to the computation of the directions d_j (and projections $z_i^{(j)}$), for $j = n_1 + 1, \ldots, n_1 + n_2$, minimizing the kurtosis coefficient.
5. For finding outliers, as in the Stahel-Donoho approach the normalized univariate distances r_i^j are computed as

$$r_i^j = \frac{1}{\beta_p} \frac{|z_i^{(j)} - \text{median}_i(z_i^{(j)})|}{\text{MAD}_i(z_i^{(j)})}, \tag{2}$$

for each direction $j = 1, \ldots, n_1 + n_2$, where β_p is a predefined reference value.

Stage II: Analysis based on directions obtained from a stratified sampling procedure as follows:

1. In iteration l, two observations are chosen randomly from the sample and the direction $\hat{d}_l$ defined by these two observations is computed. The observations are then projected onto this direction, to obtain the values $\hat{z}_i^l = \hat{d}_l^T y_i$. Then the sample is partitioned into K groups of size n/K, where K is a prespecified number, based on the ordered values of the projections $\hat{z}_i^l$, so that group k, $1 \leq k \leq K$, contains those observations i satisfying
$$\hat{z}^l_{(\lfloor (k-1)n/K \rfloor + 1)} \leq \hat{z}_i^l \leq \hat{z}^l_{(\lfloor kn/K \rfloor)}.$$
2. From each group k, $1 \leq k \leq K$, a subsample of p observations is chosen without replacement. The direction orthogonal to these observations, $\tilde{d}_{kl}$,

is computed, as well as the corresponding projections $\tilde{z}_i^{kl} = \tilde{d}_{kl}^T y_i$ for all observations i. These projections are used to obtain the corresponding normalized univariate distances r_i^j,

$$r_i^j = \frac{1}{\bar{\beta}_p} \frac{|\tilde{z}_i^{kl} - \text{median}_i(\tilde{z}_i^{kl})|}{\text{MAD}_i(\tilde{z}_i^{kl})}, \tag{3}$$

where $j = 2p + \lfloor (k-1)n/K \rfloor + l$, and $\bar{\beta}_p$ is a prespecified reference value.

3. This procedure is repeated a number of times L, until $l = L$.

Stage III: For each observation i its corresponding normalized outlyingness measure r_i is obtained from the univariate distances r_i^j defined in (2) and (3), as

$$r_i = \max_{1 \le j \le 2p + \lfloor Ln/K \rfloor} r_i^j.$$

Those observations having values $r_i > 1$ are labeled as outliers and removed from the sample, if their number is smaller than $n - \lfloor (n+p+1)/2 \rfloor$. Otherwise, only those $n - \lfloor (n+p+1)/2 \rfloor$ observations having the largest values of r_i are labeled as outliers.

The values of the parameters needed in the procedure are explained in Peña and Prieto (2005).

3 Application to Clustering

The directions obtained in the previous section can be used for finding clusters by identifying holes in the distribution of the projected data; we use the sample spacings or first-order gaps between the ordered statistics of the projections. If the univariate observations come from a unimodal distribution, there will be large gaps near the extremes of the distribution and small gaps near the center. However, this pattern will change if there are clusters in the data. For example, with two clusters of similar size we expect a large gap separating the clusters, lying towards the center of the observations. Thus, once the univariate projections are computed for each one of the $n_{1+}n_2$ projection directions, the problem is reduced to finding clusters in unidimensional samples, where these clusters are defined by regions of high probability density. We consider that a set of observations can be split into two clusters when we find a sufficiently large first-order gap in the sample. Let $z_{ki} = \boldsymbol{x}_i' \boldsymbol{d}_k$ for $k = 1, \ldots, n_1 + n_2$, and let $z_{k(i)}$ be the order statistics of this univariate sample. The first-order gaps or spacings of the sample, w_{ki}, are defined as the successive differences between two consecutive order statistics

$$w_{ki} = z_{k(i+1)} - z_{k(i)}, \qquad i = 1, ..., n-1$$

As the expected value of the gap w_i is the difference between the expected values of two consecutive order statistics, it will be in general a function of i

and the distribution of the observations. For a unimodal symmetric distribution Peña and Prieto (2001a) showed that, under reasonable assumptions, the largest gaps in the sample are expected to appear at the extremes, w_1 and w_{n-1}, while the smallest ones should be those corresponding to the center of the distribution. Therefore, if the projection of the data onto $\boldsymbol{d}_k$ produces a unimodal distribution we would expect the plot of w_{ki} with respect to k to decrease until a minimum is reached (at the mode of the distribution) and then to increase again. The presence of a bimodal distribution in the projection would be shown by a new decreasing of the gaps after some point. A sufficiently large value in these gaps would provide indication of the presence of groups in the data. The cut-off for the gaps can be determined by Monte Carlo. In summary, the algorithm will be as follows:

1. For each one of the directions d_k compute the univariate projections of the original observations $u_{ki} = x_i' d_k$.
2. Standardize these observations, $z_{ki} = (u_{ki} - m_k)/s_k$, where $m_k = \sum_i u_{ki}/n$ and $s_k = \sum_i (u_{ki} - m_k)^2/(n-1)$.
3. Sort the projections z_{ki} for each value of k, to obtain the order statistics $z_{k(i)}$ and transform then using the inverse of the standard normal distribution function $\bar{z}_{ki} = \Phi^{-1}(z_{k(i)})$
4. Compute the gaps between consecutive values, $w_{ki} = \bar{z}_{k,i+1} - \bar{z}_{ki}$.
5. Search for the presence of significant gaps in w_{ki}. These large gaps will be indications of the presence of more than one cluster. In particular, we introduce a threshold $\kappa = \nu(c)$, where $\nu(c) = 1 - (1-c)^{1/n}$ denotes the c-th percentile of the distribution of the spacings, define $i_{0k} = 0$ and

$$r = \inf_j \{n > j > i_{0k} : w_{kj} > \kappa\}.$$

   If $r < \infty$, the presence of several possible clusters has been detected. Otherwise, go to the next projection direction.
6. Label all observations l with $\bar{z}_{kl} \leq \bar{z}_{kr}$ as belonging to clusters different to those having $\bar{z}_{kl} > \bar{z}_{kr}$. Let $i_{0k} = r$ and repeat the procedure.

4 Simulation results

We present in Table 1 the percentage of successes in a simulation experiment where we have compared: (1) An efficient algorithm for the implementation of the Minimum Covariance Determinant (MCD) procedure, the FASTMCD algorithm as proposed by Rousseeuw and van Driessen (1999). (2) An implementation of the Stahel-Donoho algorithm, as described in Maronna and Yohai (1995). (3) A computationally efficient algorithm recently proposed by Maronna and Zamar (2002), based on the analysis of the principal components of an adjusted covariance matrix computed from information on pairs of observations. Two iterations of the algorithm have been carried out, as suggested by the authors. (4) An algorithm based on the directions computed

from the minimization and maximization of the kurtosis coefficient, as described in Peña and Prieto (2001b). (5) A stratified Stahel-Donoho sampling procedure, corresponding to the second part of the RASP algorithm described in Section 2. (6) An implementation of the RASP(1) algorithm described in Section 2. (7) An implementation of RASP(p), that is, the same algorithm as before but using now the full $2p$ directions maximizing and minimizing the kurtosis coefficient. The data for the experiment in Table 1 was generated from a standard normal multivariate distribution in dimensions 5, 10 and 20, contaminated with a proportion of outliers in a single cluster (from 10% to 40%), obtained from a second normal distribution with different covariance matrices. A total of 100 replications were carried out for each case and each algorithm.

Table 1. Overall success rates for the detection of outliers forming one cluster

FASTMCD	SD	MZ	kurtosis	mod-SD	RASP(1)	RASP(p)
74.9	90.1	70.2	88.0	94.9	97.5	98.0

In a second computational experiment, we have generated samples composed of one main cluster obtained from a standard normal distribution and two or four additional clusters. The success rates for algorithms FASTMCD, SD, MZ and RASP(p) are presented in Table 2. The number of directions generated in algorithm SD was chosen to have comparable running times for both SD and RASP(p). Note again the improvement obtained when using RASP(p) over the alternative algorithms.

Table 2. Overall success rates for the detection of clusters

FASTMCD	SD	MZ	RASP(p)
88.5	97.5	86.2	100.0

Table 3. Average running times for the algorithms

FASTMCD	SD	MZ	RASP(p)
233.8	7.0	17.8	7.9

Finally, to illustrate the computational efficiency of the different algorithms, Table 3 presents the average running times for the analysis of sets of 100 replications corresponding to the preceding algorithms, given in seconds

on a Pentium M 1.6 GHz. Those for both SD and RASP(p) are significantly lower than for the other algorithms.

References

1.ATKINSON, A.C., RIANI, M. and CERIOLI, A. (2004): *Exploring Multivariate Data with the Forward Search*. Springer, Berlin.

MARONNA, R. A. and YOHAI, V. J. (1995): The Behavior of the Stahel-Donoho Robust Multivariate Estimator. *Journal of the American Statistical Association, 90, 330–341.*

MARONNA, R. A. and ZAMAR, R. (2002): Robust estimates of location and dispersion for high dimensional data sets. *Technometrics, 44, 307–317.*

PEÑA, D. and PRIETO, F. J. (2001a): Cluster Identification Using Projections. *Journal of the American Statistical Association, 96, 1433–1445.*

PEÑA, D. and PRIETO, F. J. (2001b): Robust Covariance Matrix Estimation and Multivariate Outlier Detection. *Technometrics, 43, 286–310.*

PEÑA, D. and PRIETO, F. J. (2005): Combining Random and Specific Directions for Robust Estimation for High-Dimensional Multivariate Data. *Manuscript.*

ROUSSEEUW, P. J. and VAN DRIESSEN, K. (1999): A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics, 41, 212–223.*