A Projection Pursuit Method for Detecting Outliers and Finding Clusters in Multivariate Data

Daniel Peña
Dept. Estadística y Econometría, Universidad Carlos III de Madrid
C/ Madrid, 126
28903 Getafe (Madrid), Spain
dpena@est-econ.uc3m.es

Francisco J. Prieto
Dept. Estadística y Econometría, Universidad Carlos III de Madrid
C/ Madrid, 126
28903 Getafe (Madrid), Spain
fjp@est-econ.uc3m.es

1. Introduction

A recent and fast procedure for finding outliers in multivariate samples developed by Peña and Prieto (2001a) is based on the analysis of the projections of the sample points onto a certain set of $2p$ directions, where p is the dimension of the sample space. These directions are obtained by maximizing and minimizing the kurtosis coefficient of the projections. In this paper we briefly summarize this procedure and illustrate how it can be combined with the Stahel-Donoho estimate to obtain a very powerful and flexible method for outlier detection. This extension can also be applied to the method for cluster analysis developed by Peña and Prieto(2001b). A description of the procedure and the use of these directions is presented in Section 2 and Section 3 describes the combination of the systematic directions with random directions.

2. Outliers and Kurtosis

We want to find interesting directions such that the projection of the sample points onto these directions can reveal the presence of outliers. It can be shown that in univariate samples outliers produce an increase in the kurtosis of the observed data and therefore it seems sensible to project the data onto the directions which maximize the kurtosis of the projected points. Peña and Prieto (2001a) have shown that a small number of outliers will produce heavy tails and a larger kurtosis coefficient, but if we increase the amount of outliers we can start introducing bimodality and the kurtosis coefficient may decrease. This property suggests considering as the objective function to maximize and minimize the kurtosis coefficient of the projections.

The implementation of the outlier identification procedure requires the computation of $2p$ S-orthogonal directions. An efficient implementation of this procedure is described in the following steps:

1. To simplify the computations associated with the solution of the kurtosis minimization and maximization problems, the data is rescaled and centered using $y_i = S^{-1/2}(x_i - \bar{x})$, $i = 1, \ldots, n$, where $\bar{x}$ and S are the mean and covariance matrix of the original data.

2. Compute the p orthogonal directions and projections that maximize the kurtosis coefficient. Define $y_i^{(1)} = y_i$ and an iteration index $j = 1$. Solve the problem

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

   Once a unit vector d_j has been obtained from (1), the data is transformed so that (i) it remains standardized to justify the simplifications introduced in (1); (ii) the optimal direction generated from the modified data is orthogonal to the preceding ones; and, (iii) to reduce the computational effort, the transformed data should lie in a lower-dimension subspace. An orthogonal transformation of the data such that the direction defined by d_j is transformed into the first coordinate axis will have these properties. The projected data are then obtained as the remaining coordinates $(2, \ldots, p-j+1)$ of the transformed data. The corresponding transformation matrix is given by $Q_j = I - v_j v_j'/v_j' d_j$, where $v_j = d_j - e_1$ and e_1 denotes the first unit vector. Note that Q_j is orthogonal and $Q_j d_j = e_j$. We then compute the transformed 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 $z_i^{(j)}$, the first component of $u_i^{(j)}$, satisfies $z_i^{(j)} = d_j' y_i^{(j)}$, and $y_i^{(j+1)}$ corresponds to the remaining $p - j$ components of $u_i^{(j)}$. Set $j = j + 1$, and go back to step 2 if $j < p$. Otherwise, let $z_i^{(p)} = y_i^{(p)}$.

3. After completing the preceding procedure, compute another set of p orthogonal directions and projections , replacing the maximization by minimization in problem (1).

4. From the projections onto the set of $2p$ computed directions, obtain the outlyingness measure $r_i = \max_k \left(|d_k' x_i - \text{median}(d_k' x_l)| \,/ \text{MAD}(d_k' x_l)\right)$. Eliminate from the sample those observations having values $r_i > c_1$, and repeat step 1. Terminate the procedure either after all remaining observations have outlyingness measures smaller than c_1, or before the number of observations left in the sample becomes less than half the original number of observations.

The sample mean m and covariance matrix S are computed for the remaining observations, and the Mahalanobis distances for the whole sample are determined using these values. Those observations having distances such that $\gamma d_i > \chi^2_{p,v}$, $v = \alpha/n$, where α is the desired level of significance, are labeled as outliers. The value $\gamma < 1$ is a factor that attempts to correct the bias introduced in S given that some of the observations have been removed prior to its computation. The value of γ that we have used in our tests has been obtained from simulation experiments for an uncontaminated multivariate normal sample.

3. Incorporating random directions

Peña and Prieto (2001a) made a Monte Carlo study to compare the procedure presented in the previous section to (i) the FASTMCD algorithm proposed by Rousseeuw and Van Driessen (1999) for the implementation of the Minimum Covariance Determinant (MCD) and (ii) a version of the Stahel-Donoho (SD) algorithm, corresponding to the implementation described in Maronna and Yohai (1995). The proposed method seems to perform better than FASTMCD for concentrated contaminations, while its behavior is worse for small sample sizes and when the shape of the contamination is similar to that of the original data. It can be shown that this case is the most difficult one for the kurtosis algorithm, as the objective function is nearly constant for all directions, and for finite samples it tends to present many local minimizers, particularly along directions that are nearly orthogonal to the outliers. Nevertheless, this worst-case behavior disappears when the variability in the data is different that the variability in the cluster of outliers or the sample size increases. Regarding the SD algorithm, the proposed method behaves better for large space dimensions and large contamination levels, showing that it is advantageous to study the data on a small set of reasonably chosen projection directions, particularly in those situations when a random choice of directions would appear to be inefficient.

These results suggest that if we have a large set of random uniformly distributed outliers in high dimension, a method that computes a very large set of random directions will be more powerful than another one that computes a small number of specific directions. On the other hand, when the outliers appear along specific directions, a method that searches for these directions should be very useful. The results emphasize the advantages of combining random and specific directions in the search for multivariate outliers. In particular, the incorporation of the kurtosis directions in the standard Stahel-Donoho procedure could improve it in many cases, with small additional computational effort. Table 1 presents results from a simulation experiment that compares the success rates in the identification of single clusters of outliers for different algorithms, including using only the $2p$ directions obtained from the optimization of the kurtosis coefficient (Kurtosis), and a procedure that combines these directions with 1000 directions obtained from the Stahel-Donoho procedure with resampling (Kurtosis+SD). Only those cases with a significant number of failures for one of the algorithms are shown in the

table.

Table 1: Success rates for the detection of outliers forming one cluster

			$\delta = 10$				$\delta = 100$			
p	α	$\sqrt{\lambda}$	FASTMCD	SD	Kurtosis	Kurtosis+SD	FASTMCD	SD	Kurtosis	Kurtosis+SD
5	0.3	0.1	0	100	100	100	100	100	100	100
		1	100	100	95	100	100	100	94	100
	0.4	0.1	0	0	100	100	0	100	100	100
10	0.2	0.1	0	100	100	100	100	100	100	100
		1	100	100	87	100	100	100	80	100
	0.3	0.1	0	100	96	98	0	100	98	100
		1	100	100	31	95	100	100	26	100
	0.4	0.1	0	0	100	100	0	0	99	100
		1	74	0	98	96	67	0	97	100
		5	100	53	100	100	100	73	100	100
20	0.1	0.1	86	100	100	100	100	100	100	100
		1	100	100	51	100	100	100	47	100
	0.2	0.1	0	72	89	91	0	100	92	100
		1	98	61	0	33	100	100	2	96
		5	100	67	100	100	100	100	100	100
	0.3	0.1	0	0	4	1	0	0	3	19
		1	19	0	0	2	20	0	0	16
		5	100	0	100	100	100	0	100	100
	0.4	0.1	0	0	0	1	0	0	2	3
		1	0	0	11	10	1	0	15	16
		5	100	0	100	100	100	0	100	100

REFERENCES

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.

Peña, D., and Prieto, F. J. (2001a), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," Technometrics, August 2001.(to be published)

Peña, D., and Prieto, F. J. (2001b), "Cluster identification using projections," Working paper, Universidad Carlos III de Madrid.

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

RESUME

Dans cet article nous décrivons une méthode efficace pour la détection d'observations atypiques dans un échantillon de données multivariées. La méthode est basée sur l'analyse de la projection des observations sur un certain ensemble de directions obtenues en maximisant et minimisant le kurtosis de la projection. Nous commentons en outre, comment il est possible d'améliorer cette méthode en combinant ces directions avec les directions aléatoires générées par l'algorithme de Stahel-Donoho. Cette combinaison permet d'obtenir une méthode de détection des données atypiques souple et puissante. Enfin, nous présentons quelques résultats de simulations qui illustrent la méthode.