Detecting multivariate outliers using projection pursuit with particle swarm optimization

Anne Ruiz-Gazen[1], Souad Larabi Marie-Sainte[2], and Alain Berro[2]

[1] Toulouse School of Economics (Gremaq et IMT),
21, allée de Brienne, 31000 Toulouse, France
ruiz@cict.fr
[2] IRIT, 21, allée de Brienne, 31000 Toulouse, France
larabi@irit.fr, berro@irit.fr

Abstract. Detecting outliers in the context of multivariate data is known as an important but difficult task and there already exist several detection methods. Most of the proposed methods are based either on the Mahalanobis distance of the observations to the center of the distribution or on a projection pursuit (PP) approach. In the present paper we focus on the one-dimensional PP approach which may be of particular interest when the data are not elliptically symmetric. We give a survey of the statistical literature on PP for multivariate outliers detection and investigate the pros and cons of the different methods. We also propose the use of a recent heuristic optimization algorithm called Tribes for multivariate outliers detection in the projection pursuit context.

Keywords: Heuristic algorithms, Multivariate outliers detection, Particle Swarm optimization, Projection Pursuit, Tribes algorithm

1 Introduction

The definition of outliers as a small number of observations that differ from the remainder of the data is commonly accepted in the statistical literature (Barnett and Lewis (1994), Hadi et al. (2009)). Most of the detection methods in continuous multivariate data are based either on the Mahalanobis distance or on Projection Pursuit. In the first approach, an observation is declared an outlier if its Mahalanobis distance is larger than a given cut-off value. Because the classical non-robust Mahalanobis distances suffer from masking, Rousseeuw and Van Zomeren (1990) propose to use robust location and scatter estimators. Moreover, reliable methods for defining cut-off points have been recently proposed (Cerioli et al. (2009)). The PP approach consists in looking for low dimensional linear projections that are susceptible to reveal outlying observations. In the following, we focus on this second approach which does not assume that the non-outlying part of the data set originates from a particular distribution (like elliptically symmetric distributions for the first approach). In general, exploratory PP gives insight about a multivariate

continuous data set by finding and proposing to the analyst high revealing low-dimensional projections. A projection pursuit method is based on two ingredients: a projection index which measures the interestingness of a given projection and a strategy for searching the optima of this index. In the second section, we give a survey of the different projection indices that are aimed at detecting multivariate outliers. PP is computationally intensive and the choice of the strategy of "pursuit" together with the optimization algorithm are also important. In the third section, we present the existing "pursuit" strategies and propose a new strategy that relies on a optimization algorithm that can find several local minima in a reasonable time. We also investigate the pros and cons of the different strategies. In the fourth section, we present the Tribes algorithm which is a recent heuristic optimization algorithm (Clerc (2005), Cooren et al. (2009)). Heuristic optimization methods are attractive on the one hand, because they don't rely on strong regularity assumptions about the index and on the other hand, because they offer an efficient way to explore the whole space of solutions. But they usually imply the choice of some parameters. Tribes belongs to the family of Particle Swarm optimization (PSO) methods which are biologically-inspired optimization algorithms based on a cooperation strategy. Its main advantage relies on the fact that it is a parameter-free algorithm. We give some generalities concerning PSO and Tribes and propose to use it for the detection of outliers in an exploratory PP context. In the last section, we present the java interface we are currently developing for exploratory PP and give some perspectives.

2 Projection indices for detecting outliers

As said above a PP method assigns a numerical value (defined via an index) to low dimensional projections of the data. The index is then optimized to yield projections that reveal interesting structure. In the following, we review several one-dimensional indices that can be useful for the detection of outliers. We use the following notations: the data set is a n (observations) by p (variables) matrix X and X_i denotes the vector in R^P associated with the ith observation. For one-dimension exploratory PP, a real-valued index function $I(a)$ is defined for all projection vectors $a \in R^p$ such that $a'a = 1$ (where a' denotes the transpose of a). This function I is such that interesting views correspond to local optima of the function.

The most well-known projection index is the variance which leads to Principal Component Analysis (PCA). As detailed in Jolliffe (2002, section 10.1), observations that inflate variances will be detectable on the first principal components while outliers with respect to the correlation structure of the data may be detected on the last principal components. PCA is generally the first step in multivariate continuous data analysis but it is not specifically designed for the detection of outliers and further exploration with other PP indices are of interest. Moreover, in order to avoid masking as previously

mentioned for Mahalanobis distances, it is advisable to consider as a projection index a robust variance estimator rather than the usual variance (Li and Chen (1985)). Such a method, called robust PP-based PCA, may detect outliers which inflate the variance (without the possible masking of the non-robust PCA) but is not aimed at detecting other types of outliers.

The definition of an "interesting" projection has been discussed in the founding papers on PP (Friedman and Tukey (1974), Huber (1985), Jones and Sibson (1987), and Friedman (1987)). Several arguments (see Friedman (1987) for details) have led to the conclusion that gaussianity is uninteresting. Consequently, as noted by Huber (1985), any measure of departure from normality can be viewed as a measure of interestingness and thus as a PP index. The objective of measuring departures from normality is more general than looking for projections that reveal outlying observations. However, several indices are very sensitive to departure from normality in the tails of the distribution which means that they will reveal outliers in priority. We will focus on such indices. In particular, the Friedman and Tukey (1974) and Friedman (1987) indices are known to be quite sensitive to the presence of outliers (see Friedman and Tukey (1974) and Hall (1989)). A detailed presentation of these indices can be found in Caussinus and Ruiz-Gazen (2009) and Berro et al. (2009).

As mentioned by Huber (1985, p. 446) and further studied by Peña and Prieto (2001), the kurtosis of the projected data is an index well adapted for detecting outliers. While heavy tailed distributions lead to high values of the kurtosis, bimodality leads to low values of the kurtosis. Thus, Peña and Prieto (2001) propose to detect outliers by looking at projections that minimize or maximize the kurtosis.

Recently, the Friedman index (Achard et al. (2004)) and the kurtosis index (Malpica et al. (2008)) have been used successfully for detecting anomalies in hyperspectral imagery. We also mention the index proposed in Juan and Prieto (2001) which is well suited for concentrated contamination patterns but which does not seem appropriate in other situations as detailed in Smetek and Bauer (2008) also in the field of hyperspectral imagery.

Another well-known projection index which is dedicated to the research of outliers is the measure of outlyingness defined independently by Stahel (1981) and Donoho (1982). For each observation $i = 1, \ldots, n$, we look for a projection that maximizes

$$I_i(a) = \frac{|a'X_i - \mathrm{med}_j(a'Xj)|}{\mathrm{mad}_j(a'X_j)}$$

where the "med" (resp. the "mad") corresponds to the median (resp. the median absolute deviation) of the projected data. The main difference between this index and the ones previously introduced is that the search of an optimal projection has to be done for each observation while the previous proposals consist in looking for the most interesting projections without refering to any

particular observation. The Stahel-Donoho index is generally used as a first step in order to define weights of highly robust location and scatter estimators. But it may be used also in the exploratory PP context when the number of observations is small.

Finally, Caussinus and Ruiz-Gazen (1990, 2003), and Ruiz-Gazen (1993) proposed a generalization of PCA designed for the detection and the vizualization of outliers. The methodology is based on the spectral decomposition of a scatter estimator relative to another scatter estimator and has been recently revisited in a more general framework by Tyler et al. (2008). Contrary to usual and robust PP-based PCA, Generalized PCA (GPCA) cannot be defined as a problem of optimizing a function $I(a)$ of a projection vector a. Even if it is detailed as a projection pursuit method in Caussinus and Ruiz-Gazen (2009), there is no projection index associated with GPCA. Moreover, like PCA (and unlike robust PP-based PCA), the projections obtained by GPCA rely on spectral decomposition and do not need any pursuit. In the following we do not consider PCA and GPCA any further and focus on possible strategies for pursuit in the ususal exploratory PP context.

3 Different "pursuit" strategies

The structure of complex data sets in more than two dimensions is usually observable in many one-dimensional projections. So, as already stated in Friedman and Tukey (1974), PP should find as many potentially informative projections as possible. Consequently, the first strategy proposed by Friedman and Tukey (1974) and Jones and Sibson (1987) consists in using local optimization methods with several starting points. Useful suggested initial directions are the original coordinate axes, the principal axes but also some random starting points. This strategy is also the one followed by Cook et al. (1995) in their grand tour proposal but with the difference that the initial directions are chosen by the viewer in an interactive way. To our opinion, looking at rotating clouds as in Cook et al. (2007) may be tedious for the data-analyst.

A second strategy is proposed in Friedman (1987) and most of the literature on PP focus on this second strategy. The procedure repeatedly invokes a global optimization method, each time removing from the data the solutions previously found. Several global optimization methods have been considered in the literature (e.g. Friedman (1987), Sun (1993), Peña and Prieto (2001)). For continuously differentiable indices, such as the Friedman index with a smooth kernel or the kurtosis index, the global optimization procedure usually involves a local optimization step based on steepest ascent or quasi-Newton. Concerning the "structure removal", the simplest idea is to consider orthogonal projections as in PCA. This methods used in Peña and Prieto (2001) is easy to implement and greatly accelerates the procedure. However, as noticed in Huber (1985) and Friedman (1987), it may miss inter-

esting oblique projections. Friedman (1987) proposed a more sophisticated "structure removal" procedure but it is not easy to implement and, as noticed in Nason (1992), the way it may affect the later application of PP is unclear.

We propose to go back to the first strategy and offer to the data-analyst several views of the data based on numerous starting directions and an efficient local optimization algorithm. The reasons we advocate for such a choice are the following:

(i) the aim of PP is to explore several local optima and global optimization methods that consider non-global local extrema as a nuisance are time consuming and not adapted,
(ii) the structure removal may miss some interesting projections or/and is also time consuming,
(iii) by using numerous starting directions and examining the plot of the index values, we can detect whether an extremum is found by accident (because of sampling fluctuations) or discovered several times.

The drawback of this strategy, as noticed in Friedman's discussion of Jones and Sibson (1987), is that it leads to numerous views of the data that are not imediately interpretable. One does not know the extent to which a new view reflects a similar or a different structure compared with the previous views. As detailed in the perspectives, in order to circumvent the problem, we propose several simple tools to analyse and compare the different views.

Concerning the Stahel-Donoho index, Stahel (1981) and Maronna and Yohai (1995) suggest to calculate the maximum over a finite set of vectors. The vectors are taken at random and there is no local optimization step. This idea of taking a finite set of projection directions is also used to derive algorithms for robust PP-based PCA. The Croux and Ruiz-Gazen (2005) algorithm uses the directions of the observations as projection vectors. Because the index is a (robust) measure of dispersion, directions that are pointing where the data are, lead to interesting results, at least when n is larger than p (see Croux et al. (2007) for further improvement). However, this algorithm is not relevant for other types of indices.

In order to be able to deal with unsmooth indices such as the Stahel-Donoho index and explore in the most efficient way the whole space of solutions, we propose to use a recent Particle Swarm optimization algorithm called Tribes.

4 Tribes: a parameter-free Particle Swarm optimization algorithm

Tribes is a recent heuristic optimization algorithm (Clerc (2005), Cooren et al. (2009)) which belongs to the family of Particular Swarm optimization (PSO). As explained in Gilli and Winker (2008) in a statistical context, heuristics optimization methods can tackle optimization problems that are not tractable

with classical optimization tools. Moreover, such algorithms usually mimic some behavior found in nature. In the case of PSO, the algorithm mimics the behavior of a swarm of insects or a school of fish that is, the collective learning of individuals when they are in groups. There are two families of heuristic optimization methods: the trajectory methods (e.g. simulated annealing or Tabu search) which consider one single solution at a time and population based methods (e.g. genetic algorithms) which update a whole set of solutions simultaneously. For the second family of methods to which belongs PSO, the exploration of the whole search space is sometimes more efficient and this property is of importance given our objectives in the context of exploratory PP. Particle Swarm Optimization was introduced by Eberhart and Kennedy (1995) (see also Kennedy and Eberhart (2001)). The solution vectors of the population are called particles and the algorithm consists in updating the position of the particles of the swarm from one generation to another by adding an increment called velocity. More precisely, a particle is defined by a current position (which corresponds to a projection vector) and a velocity of moving in the search space. At each generation, the particle calculates the value of the function (index value). If this value is the best found so far, the particle memorizes the current position as the best position. The best value is called pbest. The particle looks also in its neighborhood the best value found. This value is called lbest. Then the particle changes its velocity toward its pbest and lbest positions in a stochastic way. Finally, she updates its position (which means that the projection vector is updated).

Recently, researchers have used PSO for solving various optimization problems (e.g. Gilli and Schumann (2009) for robust regression). But like other heuristics methods, PSO depends heavily on the selection of its parameter values which may be difficult to tune. In our case, the parameters depend notably on the number of observations and the number of variables. As described in Cooren et al. (2009), Tribes is a new adaptive PSO algorithm that avoids manual tuning by defining adaptation rules which aim at automatically changing the particles behaviors as well as the topology of the swarm. In particular, the strategies of moving are chosen according to the performances of the particles. A precise description of the Tribes algorithm is given in Larabi et al. (2009) for exploratory PP.

In Berro et al. (2009), we propose to use Genetic algorithm and standard PSO for exploratory PP but Tribes is clearly more adapted to the research of local optima. This feature is considered a drawback in a global optimization strategy ; but according to our strategy (see section 3), it is a clear advantage.

5 Perspectives

We are currently developing a java interface in order to propose to the data-analyst an efficient exploratory tool based on the PP strategy we have detailed

in the third section and on the heuristics algorithms as detailed in the fourth section.

In Berro et al. (2009), we stress the importance of using numerous indices and looking at as many views as possible. Among the implemented indices, several ones are adapted to the detection of outliers such as the Friedman-Tukey, the Friedman and the kurtosis indices. The user can center and sphere the data, a preliminary process which may ease the discovery of interesting projections (see for instance Cook et al. (1995)). Following the strategy detailed in the third section, we divide the exploratory process in two stages: the first stage consists in running several times the Tribes algorithm and obtain several projections. This research of several local optima may be time consuming especially if the number of observations or the number of variables or the number of runs are large. But the statistician does not need to be in front of the computer during this first step! Moreover, because the different runs are independent, one could use parallel computing. During this research process, the potentially interesting projections obtained by optimization of a projection index are stored in an output file. At the second stage of the procedure, the statistician has many one-dimensional views of the data at his disposal and he can begin the analysis of the potential structure. Note that at this stage, there is no more need of computing power. The user can display either histograms or kernel density estimators of the univariate distributions of the projected data (see Figure 1 for an illustration of the interface on a simulated data set). These histograms or density estimators can be examined and outliers can be easily detected by vizualisation. Comparison of the different projections (similarities and differences) is more tricky and we propose several simple tools to help the user in this process. On Figure 1, some of the tools can be vizualised. First, the projections are ordered according to the decreasing values of the projection index and the values of the index are plotted so that the different local minima are easily detected (see the plot at the top right of Figure 1). Note that the data analysed on Figure 1 are simulated data with a majority of observations following a standardized gaussian distribution and a few points following a mean-shifted gaussian distribution in eight dimensions. For this artificial example, we know that there is only one interesting projection and if we exclude a small number of runs (see the right part of the index plot), all the runs have led to almost the same value of index (see Berro et al. (2009) for more details). By repeating the local search many times, we avoid considering spurious projections (due to sampling fluctuations) since interesting projections are usually recovered several times and associated with larger index values. But similar values of the index does not correspond necessarily to similar projection vectors. We add a plot of the cosines of the angles between any chosen projection vector and the other projection directions. This plot is very helpful in order to measure how far two projection directions are. Note that the different projections are simply obtained by mouse-clicking on the index or on the cosine plot and a selection

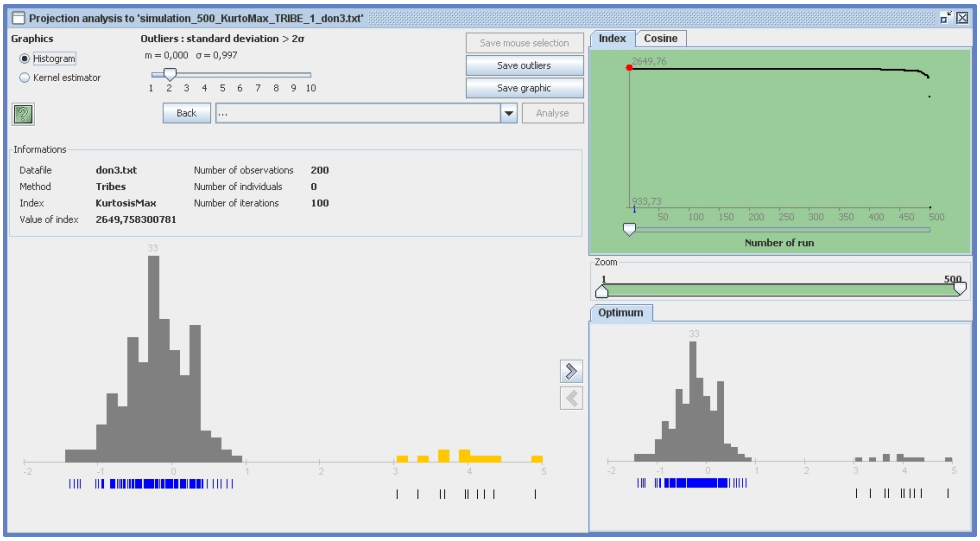

Fig. 1. A screenshot of the Java interface currently in development.

of the most interesting projections can be stored on the right bottom panel of the window (see Figure 1). In a general exploratory PP context, the analysis of many projections may be tricky and need some more dedicated tools that we are currently developing. But in the context of outliers detection, once defined an automatic rule to flag one-dimensional outlying observations, it is easy to save the outlying observations in a file together with the number of times they have been discovered on the different projections. As can be seen on Figure 1, the present version of the interface offers the possibility to declare as outliers, observations with an absolute distance to the mean larger than a certain number of times the standard deviation. The choice of the number of standard deviations is based on the vizualization of the histograms and can be changed interactively (on Figure 1, the choice is two standard deviations and the observations in yellow on the right of the histograms are identified as outliers). The interface will be soon available and will offer all the described possibilities.

Among the perspectives, we also plan to implement the Tribes algorithm for the Stahel-Donoho index in an exploratory PP context. Finally, in the context of outliers detection, we would like to compare our proposal with other existing detection methods on several data sets.

Acknowledgements

We thank Maurice Clerc, Salvador Flores, Marcel Mongeau and David Tyler for fruitful discussions.

References

ACHARD, V., LANDREVIE, A. and FORT, J.-C. (2004): Anomalies detection in hyperspectral imagery using projection pursuit algorithm In: L. Bruzzone (Ed): *Image and Signal Processing for Remote Sensing X*. Proceedings of the SPIE, Vol. 5573, 193–202.

BARNETT, V. and LEWIS, T. (1994): *Outliers in statistical data*, third edition. Wiley.

BERRO, A., LARABI MARIE-SAINTE, S. and RUIZ-GAZEN, A. (2009): Genetic and Particle Swarm Optimization for Exploratory Projection Pursuit. Submited.

CAUSSINUS, H., FEKRI, M., HAKAM, S. and RUIZ-GAZEN, A. (2003): A monitoring display of Multivariate Outliers. *Computational Statististics and Data Analysis 44, 237–252*

CAUSSINUS, H. and RUIZ-GAZEN, A. (1990): Interesting projections of multidimensional data by means of generalized principal component analysis, COMPSTAT 90, Physica-Verlag, 121–126.

CAUSSINUS, H. and RUIZ-GAZEN, A. (2009): Exploratory projection pursuit. In: G. Govaert: *Data Analysis (Digital Signal and Image Processing series)*. Wiley, 67–89.

CERIOLI, A., RIANI, M. and ATKINSON A. C. (2009): Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing 19, 341–353*.

CLERC, M. (2005): *L'optimization par essaims particulaires*. Lavoisier.

COOK, D. , BUJA. A. and CABRERA, J. (1993): Projection Pursuit Indices Based on Orthogonal Function Expansions. *Journal of Computational and Graphical Statistics 2, 225–250*.

COOK, D. and SWAYNE, D. F. (2007): *Interactive and Dynamic Graphics for Data Analysis*. Springer Verlag, New York.

COOREN, Y., CLERC, M. SIARRY, P. (2009): Performance evaluation of TRIBES, an adaptive particle swarm optimization algorithm. *Swarm Intelligence 3, 149–178*.

CROUX C. and RUIZ-GAZEN, A. (2005): High Breakdown Estimators for Principal Components: the Projection-Pursuit Approach Revisited. *Journal of Multivariate Analysis, 95, 206-226*.

CROUX, C., FILZMOSER, P. and OLIVEIRA, M. R. (2007): Algorithms for projection-pursuit robust principal components analysis. *Chemometrics and Intelligent Laboratory Systems, 87, 218-225*.

DONOHO, D. L. (1982): Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Harvard University.

EBERHART, R. C. and KENNEDY, J. (1995): A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micromachine and Human Science. Nagoya, Japan, 39–43.

FRIEDMAN, J. H. (1987): Exploratory projection pursuit. *Journal of the American Statistical Association, 82, 249–266*.

FRIEDMAN J. H. and TUKEY J. W. (1974): A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, Ser. C, 23, 881–889*.

GILLI, M. and SCHUMANN, E. (2009): Robust regression with optimization heuristics. Comisef Working paper series, WPS-011.

GILLI, M. and WINKER, P. (2008): Review of heuristic optimization methods in econometrics. Comisef working papers series WPS-OO1.

HADI, A. S., RAHMATULLAH IMON, A. H. M. and WERNER, M. (2009): Detection of outliers. *Wiley Interdisciplinary Reviews: computational statistics, 1, 57-70.*

HALL, P. (1989): On polynomial-based projection indexes for exploratory projection pursuit. *The Annals of Statistics, 17, 589–605.*

HUBER, P. J. (1985): Projection pursuit. *The Annals of Statistics, 13, 435–475.*

JOLLIFFE, I. T. (2002): *Principal Component Analysis*, second edition. Springer.

JONES, M. C. and SIBSON, R. (1987): What is projection pursuit? *Journal of the Royal Statistical Society, 150, 1–37.*

JUAN, J. and PRIETO, F. J. (2001): Using angles to identify concentrated multivariate outliers. *Technometrics 43, 311–322*

KENNEDY, J. and EBERHART, R. C. (with Yuhui Shi) (2001): *Swarm Intelligence*. Morgan Kaufmann.

LARABI MARIE-SAINTE, S., RUIZ-GAZEN, A. and BERRO, A. (2009): Tribes: une méthode d'optimization efficace pour révéler des optima locaux d'un indice de projection. Preprint.

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.*

MALPIKA, J. A., REJAS, J. G. and ALONSO, M. C. (2008): A projection pursuit algorithm for anomaly detection in hyperspectral imagery. *Pattern recognition, 41, 3313–3327*

MARONNA, R. A. and YOHAI, V. J. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association, 90 (429), 330–341.*

NASON, G. P. (1992): *Design and choice of projections indices*. Ph.D. dissertation, University of Bath.

PEÑA, D. and PRIETO, F. (2001): Multivariate outlier detection and robust covariance matrix estimation. *Technometrics, 43, 286–310*

ROUSSEEUW, P. J. and VAN ZOMEREN, B. H. (1990): Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85, 633–639.*

RUIZ-GAZEN, A. (1993): Estimation robuste d'une matrice de dispersion et projections rvlatrices. Ph.D. Dissertation. Universit Paul Sabatier. Toulouse.

SMETEK, T. E. and BAUER, K. W. (2008): A Comparison of Multivariate Outlier Detection Methods for Finding Hyperspectral Anomalies. *Military Operations Research, 13, 19–44.*

STAHEL, W. A. (1981): Breakdown of covariance estimators. Research report 31. Fachgruppe für Statistik, E.T.H. Zürich.

SUN, J. (1991): Significance levels in exploratory projection pursuit. *Biometrika, 78(4), 759–769.*

TYLER, D. E., CRITCHLEY F., DÜMBGEN L. and OJA, H. (2009): Invariant co-ordinate selection. *Journal of the Royal Statistical Society. Series B, 71(3), 549–592.*