Robust methods for complex data structures

Ursula Gather[1], Claudia Becker[2], and Sonja Kuhnt[3]

[1] Department of Statistics, University of Dortmund,
D-44221 Dortmund, Germany
(e-mail: gather@statistik.uni-dortmund.de)
[2] (e-mail: cbecker@statistik.uni-dortmund.de)
[3] (e-mail: kuhnt@statistik.uni-dortmund.de)

Abstract. It is well known that the results of classical statistical methods may be affected by model deviations as for example the occurrence of outliers. The more complex a data structure or statistical procedure, the more complicated might be the mechanism of how outliers influence the analysis. The impact of spurious observations becomes less transparent with growing complexity of models and methods. Detailed sensitivity analyses and the development of suitable robustness concepts are therefore necessary.

1 Introduction

The statistical analysis of a dataset can be affected in various ways by deviations from the assumed model. This leads to a need for robust statistical procedures. Developing such procedures becomes harder and even sometimes impossible with increasing complexity of the data structure. Here, we concentrate on the case of model deviations given by outliers in a dataset. For example, when estimating the expectation of a univariate normal sample, we can understand the effects of outliers just by visualizing what happens. This becomes clearly more difficult with increasing dimension of the data. The easy looking task of identifying all outliers in a normal sample becomes rather complicated when dealing with 10-dimensional data, say. If, additionally, the data are of a more complex structure, like the one given by a contingency table, it is even more difficult to determine the influence of outliers. This is also the case if not only the data structure is complex, but the statistical procedure is complex, too. As an example, think of a compound procedure which is carried out in two steps. Here, an outlier may affect the first step of the procedure and the result of the first step affects the analysis in the second step. At the same time, the same outlier may also directly affect the second step of the procedure.

Such examples show that we need detailed analyses of the influence outliers may have when dealing with complex data structures and complex statistical procedures. For this purpose, we need criteria describing the impact of outliers on the analysis. A criterion measuring the influence of outliers or extreme values on an estimator is the finite-sample breakdown point (Donoho and Huber (1983)). Estimators with high breakdown points do not produce arbitrarily bad results in the presence of a certain amount of extreme outliers and are robust in this respect. Taking the finite-sample breakdown point as a starting and worst case criterion, we can think of transferring its concept also to more complex situations. In the following sections, we discuss the development of robust methods with respect to breakdown criteria for some selected complex data structures. Section 2 focuses on multivariate simultaneous outlier identification rules in the case of a multivariate normal null model. In Section 3, we deal with structured data given by contingency tables. Finally, we discuss robust methods and possible breakdown concepts for a compound procedure in the context of dimension reduction.

2 Outliers in multivariate data

Consider a sample of size N from a p-dimensional normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Based on the idea of Davies and Gather (1993) we define the α_N outlier region of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the set

$$\mathbf{out}(\alpha_N, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) > \chi^2_{p;1-\alpha_N}\}.$$

We take $\alpha_N = 1 - (1-\alpha)^{1/N}$ for some given value of $\alpha \in (0,1)$, adjusting the size of the outlier region to the sample size N (see Gather and Becker (1997), Becker and Gather (1999) for details). Now assume that we wish to detect all observations in a sample $\underline{X}_N = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ which are outliers in the sense that they lie in the (unknown) outlier region $\mathbf{out}(\alpha_N, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. To identify such outliers, we define and use a so-called α_N outlier identifier (which estimates the outlier region):

$$\mathbf{OR}(\underline{X}_N, \alpha_N) = \{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \mathbf{m})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{m}) \geq c\},$$

where $\mathbf{S} = \mathbf{S}(\underline{X}_N) \in PDS(p)$ estimates $\boldsymbol{\Sigma}$ ($PDS(p) = \{\mathbf{S} \in \mathbb{R}^{p \times p} : \mathbf{S}$ positive definite and symmetric$\}$), $\mathbf{m} = \mathbf{m}(\underline{X}_N) \in \mathbb{R}^p$ estimates $\boldsymbol{\mu}$, and $c = c(p, N, \alpha_N) \in \mathbb{R}, c \geq 0$ denotes a normalizing constant. Any $\mathbf{x} \in \mathbf{OR}(\underline{X}_N, \alpha_N)$ is identified as an α_N outlier with respect to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. As a normalizing condition to determine c, we use

$$\mathrm{P}(\mathbf{X}_i \in \mathbb{R}^p \backslash \mathbf{OR}(\underline{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha,$$

where in this case $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are i.i.d. according to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This means we claim that with probability $1-\alpha$ in a sample of size N from the p-variate normal, no observation will be identified as an outlier.

Breakdown of such an outlier identification procedure can for example occur if non-outliers are declared as arbitrarily large outliers (swamping breakdown) or if arbitrarily large outliers are not identified at all (masking breakdown). Thus, the idea of the finite-sample breakdown point of an estimator can be transferred to the situation of outlier identification by defining masking and swamping breakdown points of identification rules. For the masking case,

$$\varepsilon^M(\mathbf{OR}) \;=\; \frac{k^M}{n+k^M},$$

denotes the so-called masking breakdown point of **OR**. The interpretation is that in a sample of size N, a number k^M of observations can be placed as outliers in such a way that they cause a masking breakdown of the identification procedure **OR**, where $N = n + k^M$, and that k^M is the minimum number of outliers to achieve this.

As shown in Becker and Gather (1999), for every affine equivariant identifier **OR**, we have

$$\frac{k}{N} \leq \varepsilon^M(\mathbf{OR}) \leq \frac{K}{N},$$

where $\varepsilon^*(\mathbf{m}) = k_1/N$ and $\varepsilon^*(\mathbf{S}) = k_2/N$ denote the finite-sample breakdown points of $\mathbf{m}$ and S with $k_i < N/2$, $i = 1, 2$, and $k = \min\{k_1, k_2\}$, $K = \max\{k_1, k_2\}$. Thus, the masking breakdown point of **OR** and the finite-sample breakdown point of the estimators involved in the identification procedure are strongly related. Hence robust multivariate simultaneous outlier identification procedures as introduced above should be based on robust estimators of location and scatter with high breakdown points to avoid masking breakdown of the identifier. An example of such an identification procedure would be the MCD identifier based on the robust MCD estimators (Rousseeuw (1985), Rousseeuw and van Driessen (1999), Becker and Gather (1999)).

From the findings presented above we see that the concept of the finite-sample breakdown point can be immediately applied to the context of outlier identification procedures in multivariate unstructured samples. In structured situations, like for example given by contingency tables, the situation appears to be slightly different.

3 Robust methods for the analysis of contingency tables

Contingency tables represent samples from distributions of categorical (i.e. finite discrete) vectors. Let such a vector consist of T variables with possible outcomes $\{1, \ldots, i_t\}$, $t \in \{1, \ldots, T\}$. For each possible combination $\mathbf{i} \in \mathcal{I} = \otimes_{t=1}^{T}\{1, \ldots, i_t\}$ of outcomes, a cell with entry $x_{\mathbf{i}}$ means that combination $\mathbf{i}$ occurs $x_{\mathbf{i}}$ times in the sample. The contingency table then consists of all frequencies $x_{\mathbf{i}}, \mathbf{I} \in \mathcal{I}$, which can also be seen as a vector $\mathbf{x} = (x_{\mathbf{i}})_{\mathbf{i}\in\mathcal{I}}$. Assuming a loglinear Poisson model (see Bishop, Fienberg and Holland (1975)) the corresponding random counts $X_{\mathbf{i}}$ are independent Poisson variables with expected values $m_{\mathbf{i}}$, $\mathbf{m} = (m_{\mathbf{i}})_{\mathbf{i}\in\mathcal{I}} \in \mathcal{M}$ with $\mathcal{M} \subseteq I\!R^{|\mathcal{I}|}$ as the model parameter space. Here, an α_I-outlier region of the respective Poisson distribution

$Poi(m_{\mathbf{i}})$ can be defined for each cell $\mathbf{i} \in \mathcal{I}$ of the table:

$$out(\alpha_I, Poi(m_{\mathbf{i}})) = \{x \in I\!N : poi(x, m_{\mathbf{i}}) < \delta(\alpha_I)\},$$

(where $I\!N$ denotes the set of all positive integers and $poi(\cdot, m_{\mathbf{i}})$ the probability density function of the $Poi(m_{\mathbf{i}})$ distribution), with

$$\delta(\alpha_I) = \sup\{\delta > 0 : \sum_{x \in I\!N} poi(x, m_{\mathbf{i}}) \mathbf{1}_{[0,\delta]}(poi(x, m_{\mathbf{i}})) \leq \alpha_I\}.$$

The adjustment $\alpha_I = 1 - (1-\alpha)^{1/I}$ is made according to the number of cells in the table ($|\mathcal{I}| = I$). A cell count $x_{\mathbf{i}}$ is then considered as an outlier, if it lies in the outlier region $out(\alpha_I, Poi(m_{\mathbf{i}}))$, which is again unknown and must be estimated from the data.

Let $\widehat{\mathbf{m}}(\mathbf{X})$ be an estimator for the expectation vector $\mathbf{m}$ of $\mathbf{X}$, with $\mathbf{m} = (m_{\mathbf{i}})_{\mathbf{i} \in \mathcal{I}}$ and $\mathbf{X} = (X_{\mathbf{i}})_{\mathbf{i} \in \mathcal{I}}$. A simultaneous outlier identifier can then be defined by

$$\mathbf{OI}(x_{\mathbf{i}}; \mathbf{x}, \alpha_I) = \mathbf{1}_{out(c(\alpha_I), Poi(\widehat{m}_{\mathbf{i}}(\mathbf{x})))}(x_{\mathbf{i}}) \qquad \mathbf{i} \in \mathcal{I}, \mathbf{x} \in I\!N^I,$$

with $c(\alpha_I)$ a given constant. The interpretation is that if $\mathbf{OI}(x_{\mathbf{i}}; \mathbf{x}, \alpha_I) = 1$ the count $x_{\mathbf{i}}$ is identified as an outlier. The constant $c(\alpha_I)$ should be chosen in a way that the identifier fulfills the condition

$$P_{\otimes_{\mathbf{i} \in \mathcal{I}} Poi(m_{\mathbf{i}})}(\mathbf{OI}(x_{\mathbf{i}}; \mathbf{x}, \alpha_I) = 0 \quad \forall\, \mathbf{i} \in \mathcal{I}) \geq 1 - \alpha \qquad \forall\, \mathbf{m} \in \mathcal{M}.$$

The concepts of breakdown for estimators and outlier identifiers presented in section 2 can be adjusted to the situation of contingency tables. However, for the finite-sample breakdown point, we have to note here, that $\mathbf{x}$ takes values in $I\!N^I$, that the number of observations is fixed by I, and that $m_{\mathbf{i}} \in\,]0, \infty[$, $\mathbf{i} \in \mathcal{I}$. We therefore define

$$\varepsilon^*(\mathbf{x}, \widehat{\mathbf{m}}) = \frac{k^*}{I},$$

where

$$k^* = \min\{k : \sup_{\mathbf{i} \in \mathcal{I}, \mathbf{x}^k} \widehat{m}_{\mathbf{i}}(\mathbf{x}^k) = \infty \text{ or } \inf_{\mathbf{i} \in \mathcal{I}, \mathbf{x}^k} \widehat{m}_{\mathbf{i}}(\mathbf{x}^k) = 0, k \in \{1, \ldots, I\}\},$$

with $\mathbf{x}^k$ constructed by replacing within $\mathbf{x}$ a subset of k components by arbitrary positive integers.

In the same way, the masking breakdown point has to be defined by the minimum number k^M of cell counts, which have to be replaced by outliers to cause masking breakdown:

$$\varepsilon^M(\mathbf{x}, \mathbf{OI}) = \frac{k^M}{I}.$$

It can be shown that

$$\varepsilon^M(\mathbf{x}, \mathbf{OI}) \geq \varepsilon^*(\mathbf{x}, \widehat{\mathbf{m}}).$$

An outlier identifier **OI** with a high masking breakdown point can therefore be obtained by using an estimator with a high finite-sample breakdown point.

The considered loglinear Poisson models describe complex associations between the categorical variables. The variables $X_{\mathbf{i}}$ are independent, but not identically distributed, and replacing one subset of the cell counts may be more effective in disturbing an estimator/identifier than replacing another subset. In this sense the above discussed breakdown concepts take the most pessimistic viewpoint and alternative concepts considering the pattern of the outliers in the table may be more appropriate. Some ideas in this direction will be presented. For example a stochastic breakdown function following Donoho and Huber (1983) can be defined. This function gives for each possible number of replaced observations the probability that they are arranged in a pattern causing breakdown.

From these investigations for the data structure given by a contingency table, it can be seen that a modification of well-known robustness concepts like the breakdown concept to such situations seems appropriate and possible. The following section deals with the case where also the procedure under consideration is of a more complex type.

4 Robustness for compound procedures: the example of sliced inverse regression

Statistical methods for analysing highdimensional and complex data structures often combine several procedures. This is the case for hybrid methods from different disciplines (like e.g. from computer science and statistics) as well as for statistical procedures composed of several parts. Developing robustness concepts for such compound procedures means investigating the performance of rather complex methods under model deviations. Concepts like the breakdown point cannot be transferred into this framework immediately.

As an example, consider the method Sliced Inverse Regression (SIR) for reducing the dimensionality in regression problems (Li (1991)). Here, we are in the regression setting $Y = g(\mathbf{X}) + \varepsilon$, where $Y, \mathbf{X}$ are $\mathbb{R}$- and $\mathbb{R}^d$-valued random variables, respectively, and d may be rather large. The idea for dimension reduction is that there exists some lowdimensional space $\mathcal{B} = \text{span}[\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$, $K \ll d$, and some link funktion f such that to investigate the relation between Y and $\mathbf{X}$ it suffices to consider $\mathcal{B}$ because we may write $Y = f(\boldsymbol{\beta}_1^T\mathbf{X}, \ldots, \boldsymbol{\beta}_K^T\mathbf{X}, \varepsilon)$. The aim of SIR is to estimate $\mathcal{B}$, and this is done in a five-step procedure consisting of standardization, "slicing", estimating slice locations, estimating the covariance of the slice locations, and performing a principal component analysis.

Thus, a first approach to study the breakdown of SIR could be to consider the breakdown of the estimators used in each step of the procedure and to define the breakdown point of the whole procedure as the minimum of the breakdown points of all estimators. But it turns out that this approach is not appropriate. For example, the first step of SIR consists in standardizing the observations, where a location and a covariance estimator are used. Detailed analysis of the performance of SIR shows that the location estimate in this step does not influence the outcome of the procedure. It should therefore not be taken into account when defining the breakdown point of the procedure. Altogether more refined concepts are needed to describe the breakdown and robustness of SIR and the same is true for other statistical compound procedures as well.

References

BECKER, C. and GATHER, U. (1999): The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, *94*, 947–955.

BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975): *Discrete multivariate analysis*. The MIT Press Cambridge, Massachusetts.

DAVIES, P.L. and GATHER, U. (1993): The Identification of Multiple Outliers. Invited paper with discussion and rejoinder. *Journal of the American Statistical Association*, *88*, 782–801.

DONOHO, D.L. and HUBER, P.J. (1983): The Notion of Breakdown Point. In: P.J. Bickel, K.A. Doksum and J.L. Hodges Jr. (Eds.): *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont, CA, 157–184.

GATHER, U. and BECKER, C. (1997): Outlier Identification and Robust Methods. In: G.S. Maddala and C.R. Rao (Eds.): *Handbook of Statistics, Vol. 15: Robust Inference*. Elsevier, Amsterdam, 123–143.

LI, K.-C. (1991): Sliced Inverse Regression for Dimension Reduction (with discussion). *Journal of the American Statistical Association*, *86*, 316–342.

ROUSSEEUW, P.J. (1985): Multivariate Estimation with High Breakdown Point. In: W. Grossmann, G. Pflug, I. Vincze and W. Wertz (Eds.): *Mathematical Statistics and Applications*. Reidel, Dordrecht, 283–297.

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