Convergence Rates in Multivariate Robust Outlier Identification

Ursula GATHER* Claudia BECKER*

Abstract

In investigations on the behaviour of robust estimators, typically their consistency and their asymptotic normality are studied as a necessity. Their rates of convergence, however, are often given less weight. We show here that the rate of convergence of a multivariate robust estimator to its true value plays an important role when using the estimator in procedures for identifying outliers in multivariate data.

AMS 1991 Subject Classifications: Primary 62F35; secondary 62H99, 62F12.
Key words: Outlier identification; Convergence rates.

1 Introduction

In the past few years, there has been a growing interest in methods for identifying outliers in multivariate data sets (see Barnett and Lewis [1]). Especially in the light of the existence of mere data-floods to be analysed in high dimensional online monitoring situations nowadays, the development of such methods has become a real need.

Many researchers in this field suggest the use of robust estimators in outlier identification procedures to avoid masking and swamping (Hampel [17], Rousseeuw [19, 20]). When investigating properties of these estimators, their convergence to some true underlying model parameter, i.e. consistency, is often studied first, neglecting the order of this convergence or just relying on $\sqrt{N}$-convergence. But similar to the importance of convergence rates in limit theorems for distribution functions (Cramèr [9], Butzer, Nessel [8], Theorem of Butzer,

*This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

Hahn, Westphal in Gänssler, Stute [14], p. 159, Butzer, Gather [6, 7], and many others), the convergence rate in the weak or strong convergence of estimators should not be disregarded either, especially when investigating outlier identification rules. We will show here, that in the case of so-called outlier identifiers, as introduced for the multivariate setting by Gather and Becker [16], following a univariate approach of Davies and Gather [12], the use of robust estimators with a sufficient rate of convergence is highly recommendable.

In Section 2, we introduce the concept of outlier identifiers and summarize some results concerning estimators, which are used in such identification procedures. The central part of this paper is Section 3, which deals with relations between the convergence rate of an estimator and properties of an identifier that is based on this estimator.

2 Identification of outliers in multivariate data

We concentrate here on methods for identifying outliers in data supposed to come from a multivariate normal distribution, that is, the model distribution under consideration is $N(\underline{\mu}, \Sigma)$, with $\underline{\mu} \in I\!R$, $\Sigma \in I\!R^{p \times p}$ positive definite. Following Gather and Becker [16], an α outlier with respect to this distribution is defined as an element of the α outlier region

$$\operatorname{out}(\alpha, \underline{\mu}, \Sigma) := \{\underline{x} \in I\!R^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) > \chi^2_{p;1-\alpha}\}$$

for some given $\alpha \in (0, 1)$.

In the same way, we can consider an α_N outlier region $\operatorname{out}(\alpha_N, \underline{\mu}, \Sigma)$ when dealing with a sample of size N. In this case, for $\alpha \in (0, 1)$ we have $\alpha_N = 1 - (1 - \alpha)^{1/N}$, where

$$\mathrm{P}_{N(\underline{\mu},\Sigma)}(\underline{X} \in \operatorname{out}(\alpha, \underline{\mu}, \Sigma)) = \alpha$$

and

$$\mathrm{P}_{N(\underline{\mu},\Sigma)}(\underline{X}_i \notin \operatorname{out}(\alpha_N, \underline{\mu}, \Sigma), i = 1, \ldots, N) = 1 - \alpha.$$

Usually, the outlier region will be unknown and it is a statistical task to identify all α_N outliers in a sample $\underset{\sim}{x}_N = (\underline{x}_1, \ldots, \underline{x}_N)$ which is possibly corrupted by 'bad' observations itself. One possibility to do this is to estimate the unknown outlier region and identify all observations of $\underset{\sim}{x}_N$ lying in the estimated region as outliers. Such an empirical outlier region may be defined by

$$\underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N) := \{\underline{x} \in I\!R^p : (\underline{x} - \underline{m})^T S^{-1}(\underline{x} - \underline{m}) \geq c\},$$

where $\underline{m} = \underline{m}(\underset{\sim}{x}_N) \in I\!R^p$ and $S = S(\underset{\sim}{x}_N) \in I\!R^{p\times p}$, positive definite and symmetric, are estimators for $\underline{\mu}$ and Σ, respectively. The constant $c = c(p, N, \alpha_N) \in I\!R$ is used for normalization purposes, for example according to

$$\mathrm{P}_{N(\underline{\mu},\Sigma)}(\underline{X}_i \notin \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha \tag{1}$$

with $\alpha_N = 1 - (1-\alpha)^{1/N}$ and $\alpha \in (0, 1)$.

The set $\underline{\mathrm{OR}}$ itself is also referred to as an α_N outlier identifier.

Investigations of the properties of such identifiers show that the use of very robust estimators $\underline{m}$ and S with high finite-sample breakdown points (cf. Donoho and Huber [13]) helps to avoid unfavourable effects of such rules such as masking and swamping (see e.g. Rosner [18], Simonoff [22, 23], Hampel [17], Rousseeuw [19, 20], Bendre, Kale [5], Barnett, Lewis [1], Becker [2], Becker, Gather [3]). Estimators of this kind are for example the MCD estimators of Rousseeuw and Leroy [21] and the S-estimators introduced by Davies [10]. At the same time, using estimators with this property yields finite limits for the so-called maximum asymptotic bias of outlier identifiers. More precisely, the estimators $\underline{m}$ and S must have a bounded maximum asymptotic bias themselves. We do not give the exact definition of the maximum asymptotic bias here (cf. Becker, Gather [4] for details). Roughly spoken, this bias indicates, how far away the respective estimator ($\underline{m}$, S, $\underline{\mathrm{OR}}$) can lie from the true ($\underline{\mu}$, Σ, out), when there is a certain amount of maliciously placed outliers in a sample. Therefore, a bounded maximum asymptotic bias is highly desirable.

However, the above mentioned condition of a bounded maximum asymptotic bias for the estimators is only necessary, but not sufficient. We get an identifier with bounded maximum asymptotic bias though, if we use estimators with this property and additionally guarantee a certain maximum growth rate for the normalizing constant.

Let $\underline{\mathrm{OR}}$ *be an outlier identifier as defined above with corresponding normalizing constant* $c(p, N, \alpha_N)$*. If the constant* c *fulfills the condition* $c(p, N, \alpha_N) = \mathcal{O}(\chi^2_{p;1-\alpha_N})(N \to \infty)$*, then* *the use of estimators $\underline{m}$ and S with bounded maximum asymptotic bias in $\underline{\mathrm{OR}}$ implies that the bias of the identifier $\underline{\mathrm{OR}}$ is bounded, too.* (Proof: Becker, Gather [4]).

The growth of $c(p, N, \alpha_N)$ is related to the rate of convergence of the estimators $\underline{m}$ and S used in $\underline{\mathrm{OR}}$.

3 Growth of the normalizing constant

The relationship between the growth rate of c and the properties of $\underline{m}$ and S is not immediately obvious. First, notice that if $\underline{m}$ and S are consistent estimators for $\underline{\mu}$ and Σ, then with $\underline{X}_1, \ldots, \underline{X}_N$ i.i.d. $\sim N(\underline{\mu}, \Sigma)$, it follows that $Y_i := (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m})$ are asymptotically χ_p^2 distributed. Now, for the normalizing condition (1) we have

$$\begin{aligned} & \mathrm{P}_{N(\underline{\mu},\Sigma)}(\underline{X}_i \in \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha \\ \Leftrightarrow \quad & \mathrm{P}_{N(\underline{\mu},\Sigma)}(\max_{i=1,\ldots,N} (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m}) < c(p, N, \alpha_N)) = 1 - \alpha. \end{aligned}$$

Therefore, the constant $c(p, N, \alpha_N)$ equals the $(1-\alpha)$-quantile of the distribution of $\max\limits_{i=1,\ldots,N} (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m})$. This distribution can be determined asymptotically.

Theorem 3.1 *Let denote $Y_i := (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m})$, $i = 1, \ldots, N$, with $\underline{X}_i$ as above. Then*

$$\lim_{N\to\infty} P\left(N f_{\chi_p^2}(\chi^2_{p;1-1/N})(\max(Y_1, \ldots, Y_N) - \chi^2_{p;1-1/N}) < y \right) = \exp(-\exp(-y)).$$

Here, $f_{\chi_p^2}$ denotes the Lebesgue density of the χ^2 distribution with p degrees of freedom.

Proof: Using a result of Galambos [15], p. 102, it can be shown that the χ^2 distribution lies in the maximum domain of attraction of the double exponential. Consider a distribution function F with Lebesgue density f and let $\omega(F) \leq \infty$ be the right endpoint of the support of F. Further, let there exist some $x_1 \in \mathbb{R}$ such that $\forall x : x_1 \leq x < \omega(F)$ the derivative $f'(x)$ exists and it holds that $f(x) \neq 0$. If

$$\lim_{x\to\omega(F)} \frac{d}{dx} \left(\frac{1 - F(x)}{f(x)} \right) = 0,$$

then F lies in the maximum domain of attraction of the double exponential. For the χ^2 distribution we only have to check the above limit, because all other conditions are obviously fulfilled. Now, $f(x) = f_{\chi^2_p}(x) = 2^{-p/2}\Gamma(p/2)^{-1}x^{p/2-1}e^{-x/2}$, $x > 0$, thus

$$\frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = -\frac{f'(x)(1-F(x))}{f^2(x)} - 1$$

and

$$\frac{f'(x)(1-F(x))}{f^2(x)} = \underbrace{\left(\left(\frac{p}{2}-1\right)\frac{1}{x} - \frac{1}{2}\right)}_{\to -1/2 (x\to\infty)}\left(\frac{1-F(x)}{f(x)}\right).$$

With the rules of l'Hospital we get

$$\lim_{x\to\infty}\frac{1-F(x)}{f(x)} = -\lim_{x\to\infty}\frac{f(x)}{f'(x)} = -\lim_{x\to\infty}\frac{1}{(\frac{p}{2}-1)\frac{1}{x}-\frac{1}{2}} = 2.$$

Therefore, we can conclude that

$$\lim_{x\to\infty}\frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = -\left(-\frac{1}{2}\right)2 - 1 = 0.$$

Thus, the χ^2 distribution lies in the maximum domain of attraction of the double exponential. That means, there exist sequences a_N, b_N ($b_N > 0$), such that

$$\lim_{N\to\infty} \mathrm{P}\left(\frac{\max(Y_1,\ldots,Y_N) - a_N}{b_N} < y\right) = \exp(-\exp(-y)).$$

These sequences a_N, b_N can be chosen according to (Galambos [15], p. 54, 105)

$$a_N = \inf\{x : 1-F(x) \le \frac{1}{N}\},\ b_N = \frac{1-F(a_N)}{f(a_N)}.$$

Hence we get:

$$\begin{aligned} a_N &= \inf\{x : 1-F(x) \le \frac{1}{N}\} = \inf\{x : F(x) \ge 1-\frac{1}{N}\} = F^{-1}(1-\frac{1}{N}) \\ &= \chi^2_{p;1-1/N}, \end{aligned}$$

where F^{-1} denotes the inverse of the χ^2_p distribution function, and

$$b_N = \frac{1-F(\chi^2_{p;1-1/N})}{f(\chi^2_{p;1-1/N})} = \frac{\frac{1}{N}}{f(\chi^2_{p;1-1/N})} = \frac{1}{Nf(\chi^2_{p;1-1/N})}.$$

Here, F, f denote the distribution function and Lebesgue density of the χ^2_p distribution, respectively.

Therefore, it follows that

$$\lim_{N\to\infty} \mathrm{P}\left(N f_{\chi^2_p}(\chi^2_{p;1-1/N})(\max(Y_1,\ldots,Y_N) - \chi^2_{p;1-1/N}) < y \right) = \exp(-\exp(-y))$$

or, for large N,

$$\mathrm{P}\left(\max(Y_1,\ldots,Y_N) < y\right) \simeq \exp(-\exp(-N f_{\chi^2_p}(\chi^2_{p;1-1/N})(y - \chi^2_{p;1-1/N}))).$$

$\square$

Corollary 3.1 *Let $\underline{\mathrm{OR}}$ be an outlier identifier, based on $\sqrt{N}$ consistent estimators of location and covariance and normalized according to (1). Then the normalizing constant $c(p, N, \alpha_N)$ can, for large sample sizes N, be approximated by*

$$c(p, N, \alpha_N) \simeq \chi^2_{p;1-1/N} - \frac{\ln(-\ln(1-\alpha))}{N f_{\chi^2_p}(\chi^2_{p;1-1/N})},$$

where $\alpha_N = 1 - (1-\alpha)^{1/N}$.

With this approximation, we can calculate the growth rate of c for identifiers, which are based on $\sqrt{N}$ consistent estimators.

Theorem 3.2 *Consider an outlier identifier $\underline{\mathrm{OR}}$ with $\underline{m}$, S as above. Then the growth of the normalizing constant $c(p, N, \alpha_N)$ is given by*

$$c(p, N, \alpha_N) = \mathcal{O}(\chi^2_{p;1-\alpha_N})\ (N \to \infty),$$

if condition (1) is used for normalization.

Proof: According to Corollary 3.1 we write (for large N)

$$c(p, N, \alpha_N) \simeq \chi^2_{p;1-1/N} - \frac{\ln(-\ln(1-\alpha))}{N f_{\chi^2_p}(\chi^2_{p;1-1/N})},$$

where $\alpha_N = 1-(1-\alpha)^{1/N}$. With this, we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}} \\
= &\lim_{N\to\infty} \frac{\chi^2_{p;1-1/N}}{\chi^2_{p;(1-\alpha)^{1/N}}} + \frac{-\ln(-\ln(1-\alpha))}{\chi^2_{p;(1-\alpha)^{1/N}} N f_{\chi^2_p}(\chi^2_{p;1-1/N})} \quad .
\end{aligned}$$

Application of the rules of l'Hospital to the second summand yields

$$\lim_{N\to\infty} N f_{\chi^2_p}(\chi^2_{p;1-1/N}) = \lim_{N\to\infty} \frac{1}{2} - \frac{p-2}{2\chi^2_{p;1-1/N}} = \frac{1}{2},$$

thus

$$\lim_{N\to\infty} \frac{1}{N f_{\chi^2_p}(\chi^2_{p;1-1/N})} = 2$$

and

$$\lim_{N\to\infty} \frac{-\ln(-\ln(1-\alpha))}{\chi^2_{p;(1-\alpha)^{1/N}} N f_{\chi^2_p}(\chi^2_{p;1-1/N})} = 0.$$

By similar arguments we calculate the limit of the first summand:

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\chi^2_{p;1-1/N}}{\chi^2_{p;(1-\alpha)^{1/N}}} = \lim_{N\to\infty} \frac{F^{-1}_{\chi^2_p}(1-1/N)}{F^{-1}_{\chi^2_p}((1-\alpha)^{1/N})} \\
= &\lim_{N\to\infty} \frac{f_{\chi^2_p}(\chi^2_{p;(1-\alpha)^{1/N}})}{-(1-\alpha)^{1/N}\ln(1-\alpha) f_{\chi^2_p}(\chi^2_{p;1-1/N})} \\
= &\lim_{N\to\infty} \frac{\dfrac{p-2}{2\chi^2_{p;(1-\alpha)^{1/N}}} - \dfrac{1}{2}}{-\ln(1-\alpha) f_{\chi^2_p}(\chi^2_{p;1-1/N}) + \dfrac{p-2}{2\chi^2_{p;1-1/N}} - \dfrac{1}{2}} \\
= &\ 1.
\end{aligned}$$

Therefore, we have

$$\lim_{N\to\infty} \frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}} = 1,$$

such that $c(p,N,\alpha_N) = \mathcal{O}(\chi^2_{p;1-\alpha_N})$ $(N\to\infty)$.

$\square$

These results show that the use of $\sqrt{N}$ consistent estimators in outlier identifiers must be strongly recommended.

4 Conclusion

Often, very robust estimators are proposed when having to deal with corrupted samples or for the purpose of outlier identification. Sometimes we find the consistency of such estimators proved, whereas the rate of convergence is not considered. For example, MVE estimators, which are highly robust, only possess a converge rate of $N^{1/3}$ (Davies [11]) and can therefore not be recommended as a choice for the use in outlier identifiers as defined above. On the other hand, MCD and S-estimators show the desired property of $\sqrt{N}$ consistency. Our investigations show that the rate of convergence of robust estimators plays an important role in multivariate outlier identification. Hence, calculating the convergence rates of such estimators is indeed worth the trouble.

References

[1] Barnett, V., Lewis, T., *Outliers in Statistical Data* (3rd ed.), Wiley, New York, 1994.

[2] Becker, C., Bruchpunkt und Bias zur Beurteilung multivariater Ausreißeridentifizierung, Dissertation, Fachbereich Statistik, Universität Dortmund, 1996.

[3] Becker, C., Gather, U., The Masking Breakdown Point of Multivariate Outlier Identification Rules, *Preprint*, 1997.

[4] Becker, C., Gather, U., The Maximum Asymptotic Bias of Outlier Identifiers, *Preprint*, 1997.

[5] Bendre, S., Kale, B.K., Masking Effect on Tests for Outliers in Normal Samples, *Biometrika*, **74**(1987), 891–896.

[6] Butzer, P.L., Gather, U., Asymptotic Expansions for Central Limit Theorems for General Linear Stochastic Processes. I: General Theorems of Convergence, *Mathematical Methods in the Applied Sciences*, **1**(1979), 241–264.

[7] Butzer, P.L., Gather, U., Asymptotic Expansions for Central Limit Theorems for General Linear Stochastic Processes. II: Models of the General Random Noise and Pulse Train-Processes, *Mathematical Methods in the Applied Sciences*, **1**(1979), 346–353.

[8] Butzer, P.L., Nessel, R.J., *Fourier Analysis and Approximation*, Birkhäuser, Basel, 1971.

[9] Cramèr, H., *Random Variables and Probability Distributions*, Cambridge University Press, 1937.

[10] Davies, P.L., Asymptotic Behaviour of S–Estimates of Multivariate Location Parameters and Dispersion Matrices, *The Annals of Statistics*, **15**(1987), 1269–1292.

[11] Davies, P.L., The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator, *The Annals of Statistics*, **20**(1992), 1828–1843.

[12] Davies, P.L., Gather, U., The Identification of Multiple Outliers, *Journal of the American Statistical Association*, **88**(1993), 782–792.

[13] Donoho, D.L., Huber, P.J., The Notion of Breakdown Point, in: Bickel, P.J., Doksum, K.A., Hodges, J.L. (eds.), *A Festschrift for Erich Lehmann*, Wadsworth, Belmont, CA, 1983, pp. 157–184.

[14] Gänssler, P., Stute, W., *Wahrscheinlichkeitstheorie*, Springer, Berlin, 1977.

[15] Galambos, J., *The Asymptotic Theory of Extreme Order Statistics*, 2^{nd} ed., Krieger Publishing Company, Malabar, Florida, 1987.

[16] Gather, U., Becker, C., Outlier Identification and Robust Methods, in: Maddala, G.S., Rao, C.R. (eds.), *Handbook of Statistics, Vol. 15: Robust Inference*, Elsevier, Amsterdam, 1997, pp. 123–143.

[17] Hampel, F.R., The Breakdown Points of the Mean Combined with Some Rejection Rules, *Technometrics*, **27**(1985), 95–107.

[18] Rosner, B., On the Detection of Many Outliers, *Technometrics*, **17**(1975), 221–227.

[19] Rousseeuw, P.J., Least Median of Squares Regression, *Journal of the American Statistical Association*, **79**(1984), 871–880.

[20] Rousseeuw, P.J., Multivariate Estimation with High Breakdown Point, in: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (eds.), *Mathematical Statistics and Applications*, **8**, Reidel, Dordrecht, 1985, pp. 283–297.

[21] Rousseeuw, P.J., Leroy, A.M., *Robust Regression and Outlier Detection*, Wiley, New York, 1987.

[22] Simonoff, J.S., A Comparison of Robust Methods and Detection of Outliers Techniques When Estimating a Location Parameter, *Communications in Statistics – Theory and Methods*, **13**(1984), 813–842.

[23] Simonoff, J.S., Outlier Detection and Robust Estimation of Scale, *Journal of Statistical Computation and Simulation*, **27**(1987), 79–92.

Ursula GATHER
Lehrstuhl Mathematische Statistik
und Industrielle Anwendungen
Fachbereich Statistik
Universität Dortmund
D-44221 Dortmund
Germany
e-mail: gather@omega.statistik.uni-dortmund.de

Claudia BECKER
Fachbereich Statistik
Universität Dortmund
D-44221 Dortmund
Germany
e-mail: becker@omega.statistik.uni-dortmund.de