Disclosure risk for high dimensional business microdata

Flavio Foschi (foschi@istat.it)

Istat, Development of Information Systems and Corporate Products, Information Management and Quality Assessment Directorate

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality

Tarragona, Spain, 26-28 october 2011

Overview

- Some preliminary remarks
- Disclosure risk and continuous microdata
- The need of robust estimators
- Robust finite Gaussian Mixtures
- Robustness via Trimmed Likelihood Estimators
- Corrections for consistency and small sample bias
- Multiple hypothesis test issues
- A simulation experiment
- A first application to ESA survey
- Conclusions

Some preliminary remarks

- This talk presents a work in progress about the disclosure risk assessment for business microdata.

- Due to the experimental stage of analyses, a very simple framework is considered.

- To our aims, the Enterprises' System of Accounts (ESA) survey by the Italian National Institute of Statistics is suitable for a real data application:

    - since ESA data are collected by a census on companies (having at least 100 workers), survey weights issues are ignored.

    - as partial and total non response in ESA survey are coped with integration by balance-sheets for stock companies and analytical imputations for the remaining ones, missing value problems are also ignored.

Disclosure risk and quantitative microdata

- Several definitions of disclosure risk were proposed because of different informative gains for the intruder.

- The kind of the informative gain is often related to the scale of data (here intended as scenario variables).

- Neglecting, for simplicity, any issue related to the cognitive purpose of the intruder, obvious but crucial facts are:

  - for data on a nominal scale different labels imply different meanings; this circumstance is independent from the statistical model suitable to approximate the data generating process.

  - for quantitative data any judgment depends on the statistical model. Different statistical models easily give different answer to the same crucial question: the existence of a significant diversity between given quantities.

Disclosure risk and quantitative microdata

Hence, at least two consequences follow:

- dealing with data on a quantitative scale, the strategy to assess the disclosure risk cannot be checked out of the assumed data model. This holds true even if census data are available: in addition "all units are unique (rare) w.r.t. a small set of quantitative variables" (AA.VV., 2010).

- the need of a reliable model as well the impossibility to verify the risk assessment strategy out of the assumed model, imply in turn the need of robustness against deviations from the assumptions, so that estimates can preserve their consistency even if the underling model is only approximately true (the famous quote of George E.P. Box seems suitable: "All models are wrong, but some are useful").

The need of robust estimators

- According to the meaning pointed out by Peter J. Huber the term “robustness” is intended as *insensitivity to small deviations from the assumptions.* Hence distributional robustness concerns deviations from the assumed model,

- the term “small deviations” regards gross errors in a small fraction of the observations,

- robust procedures are superior to approaches like “first clean data, then use classical estimators”: especially in multivariate settings, false rejections and false retentions largely affect the distribution of “clean” data selected by a two step strategy,

- the term “superior”, used above, means that consistent estimates of the parameters featuring the “idealized model” are achieved even if the latter is only approximately true.

The need of robust estimators

Consider

- an estimator T,
- a sample $\mathbf{X}$ of n observations on p variables,
- the set of all possible corrupted samples $\mathbf{X}'$ obtained substituting m original data points,
- the quantity

$$b(m) = \sup_{\mathbf{X}'} \left\| T\left(\mathbf{X}'\right) - T\left(\mathbf{X}\right) \right\|$$

Then, the finite sample breakdown fraction of T at the sample $\mathbf{X}$ is

$$\varepsilon_n\left(T, \mathbf{X}\right) = \min\left\{ \frac{m}{n} : b\left(m\right) = \infty \right\}$$

Hence the higher $\varepsilon_n(T, \mathbf{X})$, the higher the robustness of T at the sample $\mathbf{X}$. A breakdown fraction greater than 0.5 seems meaningless: the concept of data contamination should not be referred to the majority of sample units.

Robust finite Gaussian mixtures

- Robust methods, as intended here, work with parametric models, but the latter are no longer supposed to be exactly true.

- As business microdata are featured by (estimated) skewness and kurtosis values not consistent w.r.t. the assumption of multivariate normal distribution, a suitable parametric model is necessary.

- Finite Gaussian mixtures can approximate a wide range of distributional shapes and allow a simple extension of robustness findings about normal models.

$$L(\mathbf{x}_i;\theta)=\sum_{g=1}^{G}\varphi_g N\left(\mathbf{x}_i;\mu_g,\Sigma_g\right),\quad \forall g\ \varphi_{\mathrm{g}}\geq 0,\quad \sum_{g=1}^{G}\varphi_g=1$$

- Robust finite Gaussian mixtures are encompassed by the class of Trimmed Likelihood Estimators

Robustness via Trimmed Likelihood Estimators

- Let ℓ be a log-likelihood function and $\nu(i)$ a permutation of indices such that:

$$\ell\left(\mathbf{x}_{\nu(i)};\theta\right) \geq \ell\left(\mathbf{x}_{\nu(i+1)};\theta\right), \quad \mathbf{x}_{\nu(i)} \in \mathbb{R}^p, \quad i=1,\cdots,n$$

- Given $h < n$, the maximum trimmed log-likelihood estimator of θ is:

$$TL_h = \arg\max_{\theta} \sum_{i=1}^{h} \ell\left(\mathbf{x}_{\nu(i)};\theta\right)$$

- An iterative process gives an approximated solution:

1. at the rth iteration define $Q^{(r)} = \sum_{i=1}^{h} \ell\left(\mathbf{x}_{\nu(i)};\theta^{(r)}\right)$
2. sort $\ell\left(\mathbf{x}_{\nu(i)};\theta^{(r)}\right)$ in descending order and select the first h indices
3. compute $\theta^{(r+1)} = \arg\max_{\theta} Q^{(r+1)}$ and return to step 1

Robustness via Trimmed Likelihood Estimators

- Steps 1, 2, 3 constitute on the whole a Concentration step (C-step):
  - given θ, a C-step select the subset of h observations having the larger log-likelihoods,
  - given that subset, a new estimate of θ is computed,
  - each C-step achieves a non decreasing value of the objective function Q.

- Considering a p-variate finite Gaussian mixture of G components:
  - parameters are $\theta=\{\mu_1,\ldots,\mu_G,\Sigma_1,\ldots,\Sigma_G,\varphi_1,\ldots,\varphi_G\}$,
  - by using an information criterion, i.e. the BIC, each C-step can select the best model between candidates featured by different values of G and parameterizations of $\Sigma_{g,}$
  - for $\{\mu_1,\ldots,\mu_G,\Sigma_1,\ldots,\Sigma_G\}$ a breakpoint not less than $(n-h)/n$ is achieved when

$$0.5\left[n+G\left(p+1\right)\right]<h<\left[n-G\left(p+1\right)\right]$$

Corrections for consistency and small sample bias

Assign each of the $n–h$ observations to the mixture component which minimizes the squared Mahalanobis distance; notice that assignation is only used to estimate the total number of units belonging to each component. For g=1,…,G:

- let a_g be the number of units (from the $n–h$ ones) assigned to the g^{th} mixture component

- keep the the number of units h_g allocated to the g^{th} component by the trimmed likelihood maximization as well the estimated total number of its units $n_g = h_g+a_g$,

- the (approximate) correction which makes the estimate of Σ_g consistent to the normal model (Tallis, 1963) is:

$$c_1\left(h_g\right)=\left(h_g/n_g\right)\Big/P\left\{\chi^2_{p+2}<\chi^2_{p,\alpha=h_g/n_g}\right\}$$

Corrections for consistency and small sample bias

The correction of $\hat{\Sigma}_g$ for small sample bias is based on the following facts:

- multivariate estimators of location and shape are affine equivariant; If $\mathbf{A}$ is a non singular $p{\cdot}p$ matrix, $\boldsymbol{c}$ a $p{\cdot}1$ vector and $\mathbf{X}$=($\mathbf{x}_1,\ldots,\mathbf{x}_\mathrm{n}$):

$$\hat{\mu}\left(\mathbf{XA}+\mathbf{1}_n\mathbf{c}^t\right)=\hat{\mu}\left(\mathbf{X}\right)\mathbf{A}+\mathbf{1}_n\mathbf{c}^t \qquad \hat{\Sigma}\left(\mathbf{XA}+\mathbf{1}_n\mathbf{c}^t\right)=\mathbf{A}^t\hat{\Sigma}\mathbf{A}$$

- Mahalanobis distances are invariant to affine transformations and simulating data from p-variate standard normal, should be $|\hat{\Sigma}_g|^{1/p}= 1$

- hence, by k Monte Carlo simulations the correction factor would be

$$c_2\left(h_g\right)=\left(k^{-1}\sum_{J=1}^{k}\left|\hat{\Sigma}_g^{(j)}\right|^{1/p}\right)^{-1}$$

Usually, formulas to approximate the correction factor at any n and p are used (Pison, Van Aelst, Willems, 2002) .

Multiple hypothesis test issues

For g=1,…,G, define the robust squared Mahalanobis distances as

$$D_{g,i}{}^{2} \equiv c_1^{-1}\left(h_g\right)c_2^{-1}\left(h_g\right)\left(\mathbf{x}_i - \hat{\mu}_g\right)^t \hat{\Sigma}_g^{-1}\left(\mathbf{x}_i - \hat{\mu}_g\right)$$

Well known distributional results give:

- for each of the h units involved in robust estimations (Wilks, 1963)

$$D_{g,i}{}^{2} \sim \frac{\left(n_g - 1\right)^2}{n_g} Beta\left(\frac{p}{2}, \frac{n_g - p - 1}{2}\right), \quad i \in g \cap \{1, \cdots, h\}$$

- for each of the trimmed n–h data points (Chew, 1966)

$$D_{g,i}{}^{2} \sim \frac{\left(n_g^2 - 1\right)p}{\left(n_g - p\right)n_g} F\left(p, n_g - p\right), \quad i \in g \cap \{h+1, \cdots, n\}$$

Multiple hypothesis test issues

- Since extreme observations are approximately independent from location and scale estimates, the intersection between multiple tests sharing the same estimates should be negligible.

- with probability $1-\alpha$, no observation lies in the critical region if

$$H_0 : \bigcap_{i=1}^{n_g} \left\{ \mathbf{x}_{g,\,i} \sim N\left(\mu_g, \Sigma_g\right) \right\} \qquad \alpha_{n_g} = 1-(1-\alpha)^{1/n_g}$$

- to improve the power of the test, if the null is rejected at level α_{n_g} , then each observation is tested at level α:

$$H_{0,\,i} : \mathbf{x}_{g,\,i} \sim N\left(\mu_g, \Sigma_g\right)$$

- Observations whose maximum p-value over g=1,…,G, falls in the critical region of that test are labelled as atypical.

Multiple hypothesis test issues

- Outcomes related to the g^{th} mixture component can be:

	H_0 not rejected	H_0 rejected	Total
H_0 True	$n_{g,0/0}$	$n_{g,1/0}$	$n_{g,0}$
H_0 False	$n_{g,0/1}$	$n_{g,1/1}$	$n_{g,1}$
Total	n_g-R	R	n_g

- $n_{g,1/0}$ is the I error type, indicated as “swamping” (*S*), and is controlled by the level of the test*.*

- $n_{g,0/1}$ is the II error type representing the amount of “masking” (*M*) and depends on the power of the test,

- the proposed strategy allows an increase of swamping to alleviate the amount of masking if the absence of contamination is confuted.

A simulation experiment

- Clean data are mixtures of two spherical normal:
  - n=90 or 180, p=6, $\mu_g=\{0\cdot\mathbf{1}p,\ 2c\cdot\mathbf{1}p\}$ with $c=(\chi^2_{p,\,0.99}/p)^{1/2}$, $\varphi_g=\{5/9,\ 4/9\}$
- Added contaminated data points are 20% of the smallest clean component, according to three kind of contamination: separate, radial, diffuse:

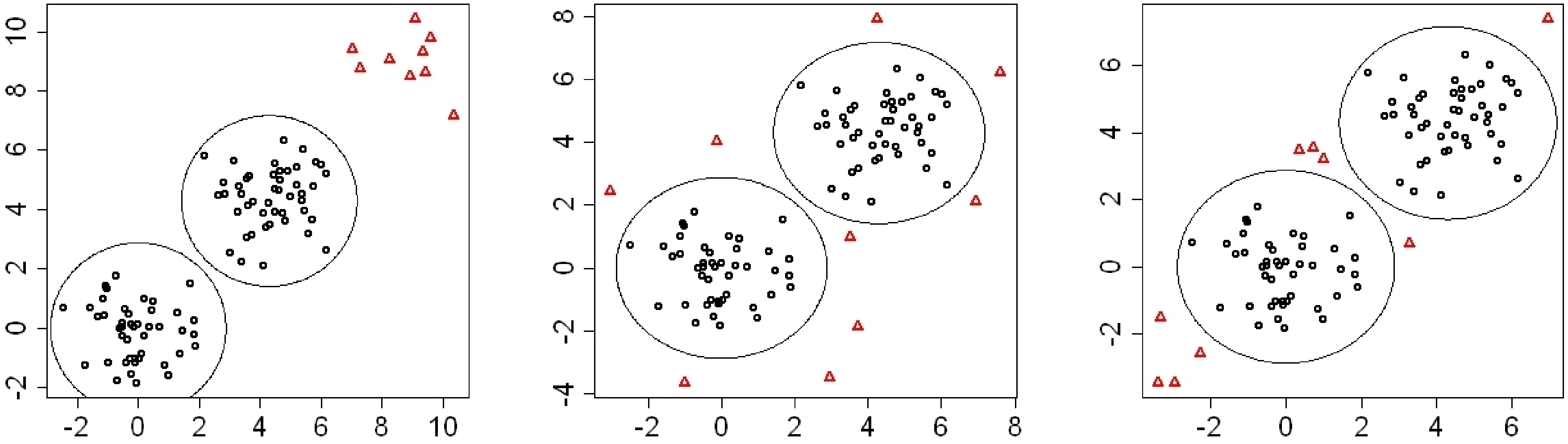
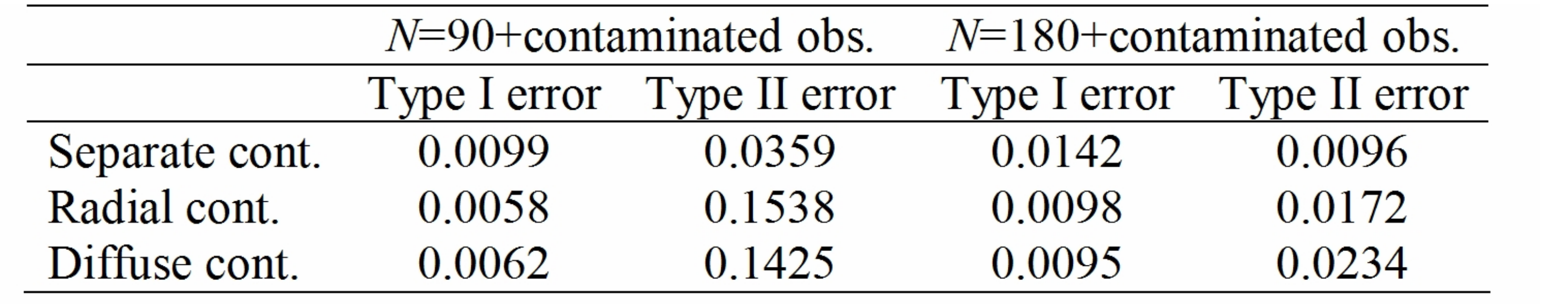


Fig 1 Two dimensional examples: separate, radial, diffuse contaminations (symbol △).

- 500 Monte Carlo replicates were performed, using ε_n=0.25 and α=0.05

A simulation experiment



	N=90+contaminated obs.		N=180+contaminated obs.	
	Type I error	Type II error	Type I error	Type II error
Separate cont.	0.0099	0.0359	0.0142	0.0096
Radial cont.	0.0058	0.1538	0.0098	0.0172
Diffuse cont.	0.0062	0.1425	0.0095	0.0234

Table 1 Expected fractions of I and II type errors (|*S*|/n and |*M*|/|*cont. data points*|).

- Swamping is negligible. Tests performed on clean data give an expected fraction of 0.0002 false rejections when n=90 and no swamping when n=180.

- Masking is evident for diffuse and radial contaminations when n=90:
  - the small sample size affects the power,
  - the small number of contaminated observations inflates the respective proportions.

- On the whole, considering the severity of settings, the performances seem acceptable.

A first application to ESA survey

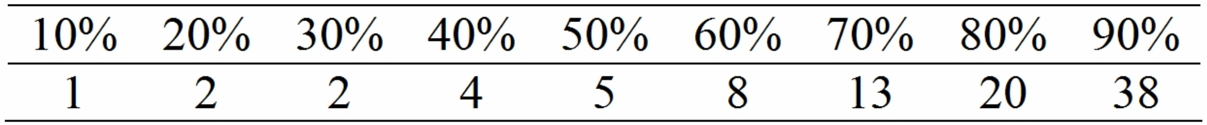
- Data of 2004 Enterprises’ System of Accounts survey are stratified by economic activity (NACE Rev. 1.1) and size class.

- A set of 10,313 records related to units selected according to the main/unique production line is analyzed.

- The assumed disclosure scenario involves six variables from the profit loss account:
  - Turnover,
  - Cost of materials, power consumptions and goods to resale,
  - Cost for services,
  - Staff costs,
  - Number of workers,
  - Earnings.

- By using NACE 4 and two size classes in term of workers, $[100, 499] \cup [500, +\infty)$, 660 strata follow**.**

A first application to ESA survey

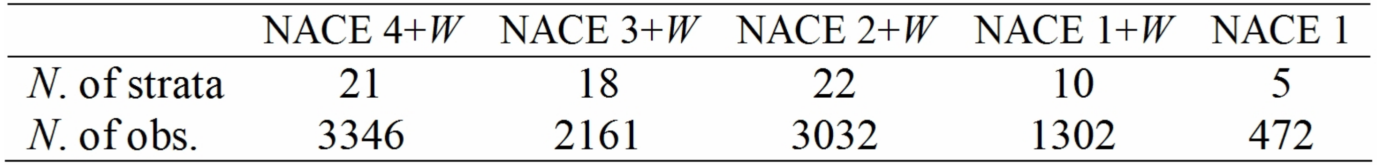


10%	20%	30%	40%	50%	60%	70%	80%	90%
1	2	2	4	5	8	13	20	38

Table 2 Quantiles of stratum sizes (four NACE digits and two size classes).

- For reducing data sparsity it is possible to collapse strata having few units into a pooled stratum.
  - When necessary, NACE digits are step by step decreased from 4 to 1; finally, if a stratum is still unsatisfactory, classes *W* are ignored; thus, the 1st digit of NACE represents the lowest admissible data resolution.
  - In a given step, a statistical unit belonging to a stratum which unfits the threshold is assigned to the next larger stratification level; the latter does not include units previously allocated into strata above the minimum number of observations.

A first application to ESA survey



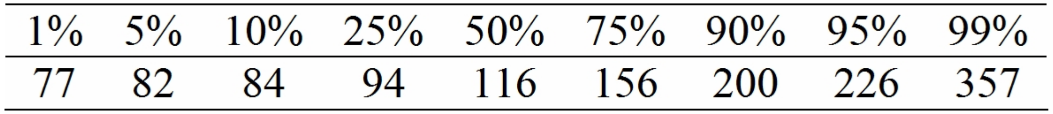
	NACE 4+W	NACE 3+W	NACE 2+W	NACE 1+W	NACE 1
N. of strata	21	18	22	10	5
N. of obs.	3346	2161	3032	1302	472

Table 3 Summary about collapsed stratification.

- A threshold of 80 observations (selected by trial and error to attenuate size differences), gives 76 strata.

- Without resorting to the collapse of lighter strata, estimations would be not feasible or misleading.

- The price to pay is the voluntary aggregation of heterogeneous observations.



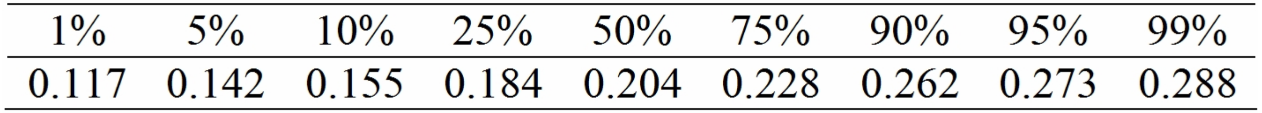
1%	5%	10%	25%	50%	75%	90%	95%	99%
77	82	84	94	116	156	200	226	357

Table 4 Quantiles of collapsed stratum sizes.

A first application to ESA survey

- The total number of enterprises declared not consistent w.r.t. the distribution of the remaining onesis 2143.

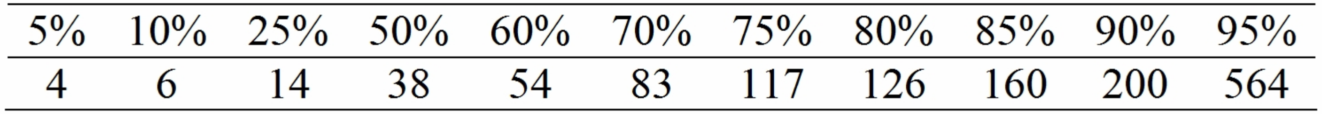
- Conditionally on respective strata, the fraction of candidate atypical units is showed in table 5. The results match the conjecture large size companies are featured by weak exchangeability.



1%	5%	10%	25%	50%	75%	90%	95%	99%
0.117	0.142	0.155	0.184	0.204	0.228	0.262	0.273	0.288

Table 5 Quantiles of the number of "atypical" units divided by the stratum size.

- By using the original strata definition, NACE 4 and two size classes in term of workers, table 6 shows some order statistics about the size of strata to whom suspect units belong. No more than 25% of those units fall in the first 70% of original strata.



5%	10%	25%	50%	60%	70%	75%	80%	85%	90%	95%
4	6	14	38	54	83	117	126	160	200	564

Table 6 Quantiles of sizes featuring original strata the "atypical" units belong to.

Conclusions

- The detection of statistical units lacking in consistency w.r.t. the process generating the majority of observations, represents a relevant first step to assess the disclosure risk of business microdata.

- In a simplified framework (ignoring survey weight and missing value issues), robust finite Gaussian mixtures and strata collapsing to perform reliable hypotheses tests are proposed.

- As tests are uninformative about the kind of data heterogeneity, any judgment requests further analyses and can imply subjective choices.

- Matters not investigated here will be dealt in future works as well further studies to achieve a global strategy on disclosure risk evaluation.

References

AA.VV. (2010). *Handbook on Statistical Disclosure Control.* ESSnet on Statistical Disclosure Control. http://neon.vb.cbs.nl/casc/handbook.htm. 18/08/2011

Becker, C., Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association.* 94. 947–955

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association.* 105. 147–156

Cerioli, A., Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Computational Statistics and Data Analysis*. 55. 544–553

Chew, V. (1966). Confidence, Prediction, and Tolerance Regions for the Multivariate Normal Distribution. *Journal of the American Statistical Association*. 61. 605–617.

Fraley, C., Raftery, A., E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association.* 97. 611–631.

Fraley, C., Raftery A., E. (2006). MCLUST version 3 for R: normal mixture modelling and model-based clustering. *Technical report no. 504, Department of Statistics, University of Washington* (revised 2009).

Hardin, J., Rocke, D., M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis.* 44. 625–638.

Hardin, J., Rocke, D., M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*. 14. 910–927.

Huber, P., J., Ronchetti, E. (2009). *Robust statistics*. John Wiley and Sons.

Lopuhaä, H., P., Rousseeuw, P., J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*. 19(1). 229–248.

Neykov, N., M., Müller, C., H. (2003). Breakdown Point and Computation of Trimmed Likelihood Estimators in Generalized Linear Models. In: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P., J. (Eds.), *Developments in robust statistics.* Physica-Verlag, Heidelberg. 277–286.

Neykov, N., M., Filzmoser, P., Dimova, R. & Neytchev, P. (2007). Robust fitting of mixtures using the Trimmed Likelihood Estimator. *Computational Statistics & Data Analysis.* 17(3). 299–308.

Pison, G., Van Aelst, S. & Willems G. (2002). Small sample corrections for LTS and MCD. *Metrika*. 55. 111–123.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3–900051–07–0. http://www.R-project.org/.

Rousseeuw, P., J., Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics.* 41, 212–223.

Tallis, G., M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics.* 34. 940–944.

Todorov, V., Templ, M. & Filzmoser, P. (2011). Detection of multivariate outliers in business survey data with incomplete information. *Advances in Data Analysis and Classification.* 5(1). 37–56.

Vandev, D., L., Neykov N., M. (1993). Robust maximum likelihood in the Gaussian case. In: Morgenthaler, S., Ronchetti, E. and Stahel, W. A (Eds.), *New Directions in Statistical Data Analysis and Robustness*. Basel. Birkhauser Verlag. 257–264.

Wilks, S., S. (1963). Multivariate Statistical Outliers. *Sankhyā, A*. 25(4). 407–426.