Discussion of “the power of monitoring: how to make the most of a contaminated multivariate sample” by andrea cerioli, marco riani, anthony c. atkinson and aldo corbellini

Heritier, Stephane; Victoria-Feser, Maria-Pia

Stat Methods Appl
https://doi.org/10.1007/s10260-017-0412-0

ORIGINAL PAPER

Discussion of "The power of monitoring: how to make the most of a contaminated multivariate sample" by Andrea Cerioli, Marco Riani, Anthony C. Atkinson and Aldo Corbellini

Stephane Heritier[1] · **Maria-Pia Victoria-Feser**[2]

Accepted: 19 November 2017

Abstract This paper discusses the contribution of Cerioli et al. (Stat Methods Appl, 2018), where robust monitoring based on high breakdown point estimators is proposed for multivariate data. The results follow years of development in robust diagnostic techniques. We discuss the issues of extending data monitoring to other models with complex structure, e.g. factor analysis, mixed linear models for which S and MM-estimators exist or deviating data cells. We emphasise the importance of robust testing that is often overlooked despite robust tests being readily available once S and MM-estimators have been defined. We mention open questions like out-of-sample inference or big data issues that would benefit from monitoring.

Keywords S-estimators · Mixed models · Deviating cells · Out-of-sample inference

1 Introduction

We congratulate the authors for their comprehensive and convincing presentation of the data monitoring approach based on the forward search (FS) algorithm using high breakdown robust estimation of covariance matrices. The methods presented here can be modified to include various degree of robustness via the choice of tuning constants controlling the breakdown point and/or efficiency of the robust estimators. They are de facto used as diagnostic tools to determine important tuning parameters, such as the breakdown point, for the robust covariance estimator. In particular, Figures 2 to 8, based on the Eruptions of Old Faithful dataset, clearly illustrate the behaviour of the FS

✉ Stephane Heritier
stephane.heritier@monash.edu

1 School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia

2 Geneva School of Economics and Management, Geneva University, Geneva, Switzerland

Published online: 27 November 2017

in detecting the proportion of data that can be considered as outliers, hence providing information about the breakdown point for possible parameters tuning. Cerioli et al. (2018) also compare several robust estimators, which incidentally, do not lead to the same breakdown point estimation.

Detecting outliers in multivariate data when one assumes that the majority of the data are generated by a multivariate distribution goes back to the early stages of the development of robust theory. The "diagnostics" approach here is to first compute a robust estimator (high breakdown point) of the covariance matrix and then, given estimates, compute some "discrepancy" measures, such as the Mahalanobis distance. Contributions to high breakdown estimation of covariance matrices include the orthogonalized Gnanadesikan–Kettenring (OGK) estimator (Gnanadesikan and Kettenring 1972; Devlin et al. 1975; Maronna and Zamar 2002), the MVE and MCD estimators (Rousseeuw 1984), the Stahel–Donoho estimator (Stahel 1981; Donoho 1982), S-estimators (Rousseeuw and Yohai 1984; Davies 1987; Lopuhaä 1989), MM-estimators (Yohai 1987; Tatsuoka and Tyler 2000), Butler et al. (1993), constrained M-estimators (Kent and Tyler 1996; Rocke 1996; Liu et al. 1999; Zuo and Cui 2005; Candès et al. 2011). The reader is also referred to the books of Hampel et al. (1986); Rousseeuw and Leroy (1987); Maronna et al. (2006); Heritier et al. (2009) and Huber and Ronchetti (2009).

Algorithms to compute the estimators and, in some cases, associated diagnostic tools have been proposed by Woodruff and Rocke (1994); Rocke and Woodruff (1996); Rousseeuw and Van Driessen (1999); Olive (2004); Salibian-Barrera and Yohai (2006); Salibian-Barrera et al. (2006), Maronna et al. (2006, page 199), Critchley et al. (2010); Hubert et al. (2012) and Hubert et al. (2015).

Early attempts to estimate robustly covariance matrices when some of the data at hand are missing dates back to Little and Smith (1987) and Little and Rubin (1987). High breakdown estimation of multivariate location and scale were then proposed by Cheng and Victoria-Feser (2002); Copt and Victoria-Feser (2004); Alqallaf et al. (2009) and Danilov et al. (2012).

As Cerioli et al. (2018), we also consider estimation of a covariance matrix as a first step towards a more thorough data analysis. In what follows, we briefly mention robust methods that have been developed in multivariate settings, most of them under the multivariate normality assumption, for which the robust monitoring approach of Cerioli et al. (2018) could possibly be extended. We also discuss the important issue of robust inference (testing) and out-of-sample validity.

2 Robust estimation and outlier detection in complex structures

2.1 Deviating data cells in multivariate samples

The identification of outliers in multivariate settings has been extended to cellwise contamination. The problem is complex as this type of contamination cannot be identified easily using purely columnwise/rowwise methods such as S or MM-estimators. Substantial progress has been made in the last decade; see for instance Agostinelli et al. (2015); Öllerer (2016) and Leung et al. (2016). The snipping approach of Farcomeni

(2014a, b) developed to deal with robust clustering with cellwise contamination can also accommodate single clusters. Very recently, Rousseeuw and Van den Bossche (2017) proposed a method that can detect deviating cells while taking correlation into account. It has no restriction on the number of clean rows and can deal with high dimensions. A R package called Cellwise is now available on Cran. Monitoring cellwise contamination can certainly be of interest and could bring new insight on the data and estimators to be used.

2.2 Principal component and factor analysis

Robust estimation of the covariance matrix is an essential step for principal component analysis (PCA), or similarly, factor analysis (FA), both being dimension reduction technique. Various robust alternatives have been proposed; see for example Devlin et al. (1981); Croux and Hasebroeck (2000); Croux and Ruiz-Gazen (2005); Hubert et al. (2005); Salibian-Barrera et al. (2006); Croux et al. (2007); Hubert et al. (2009); Dupuis Lozeron and Victoria-Feser (2010); Xu et al. (2012) and Xu et al. (2013). Monitoring outliers in this setting is also important in practice, especially when the results of PCA or FA are used to compute individual scores that can be used for example to somehow classify participants according to their response pattern. For example, Mavridis and Moustaki (2008) implement the forward search algorithm in FA models and Mavridis and Moustaki (2009) extend it to FA with binary data (for the non normal cases, see also Moustaki and Victoria-Feser 2006).

2.3 Mixed linear models

Another notable multivariate setting is the framework of mixed linear models. Common examples are repeated measures taken over time on the same individual and responses from patients in group randomised trials. In this situation, the clusters are independent and a multivariate normal formulation is available at the cluster level. Copt and Victoria-Feser (2006) exploited this equivalence and proposed S-estimators in the balanced case, i.e. all clusters have the same dimension p. The difference with the multivariate model considered by Cerioli et al. (2018) is that (i) μ_i, the mean of the outcome y_i for cluster i can be written as a linear combination of covariates, i.e. $\mu_i = \mathbf{x}_i^T \alpha$; (ii) Σ, the variance of y_i also have a specific structure, i.e. $\Sigma = \sum_{j=0}^{r} \sigma_j^2 \mathbf{z}_j \mathbf{z}_j^T$ where the $\mathbf{z}_j$'s are design matrices for the r random effects. Like in the unstructured case, these S estimators have a high breakdown assuming that the $\mathbf{z}_j$'s are well controlled. In addition, once a high breakdown estimate of the covariance matrix Σ has been obtained via a S-estimator, MM-estimators follow as later suggested by Copt and Heritier (2007); see Heritier et al. (2009) and the book website and also Koller (2016) for R code. More recently, Chervoneva and Vishnyakov (2014) extended the theory developed in Copt and Victoria-Feser (2006) by relaxing the assumption of the same number of observations per cluster. Their general S-estimator shares similar properties to the ones proposed earlier but can accommodate unbalanced clustered data. A nice illustration of this approach can be found in Chervoneva and Vishnyakov (2011). The monitoring approach proposed by Cerioli et al. (2018) and particularly the plots of squared Mahalanobis distances from monitoring S or MM-estimation could valuably be used in the this setting.

It should be stressed that the above references do not only talk about high breakdown estimation in mixed linear models but also about robust testing, a point that tends to be overlooked. In the context of this discussion, we would like to emphasize that robust monitoring is nice but robust monitoring done jointly with robust inference (testing) is better, especially when the tools are available. Heritier and Ronchetti (1994) showed that robust M-estimators can be used to build Wald, score and likelihood ratio type tests that have a stable level (robustness of validity) and power (robustness of efficiency) in a neighbourhood of the model. Examples of such tests include the robust score test based on the S-estimator of Copt and Victoria-Feser (2006), the robust likelihood ratio type test derived from the MM-estimator of Copt and Heritier (2007) and their equivalent for one-way multivariate analysis of variance (Van Aelst and Willems 2012).

3 Out-of-sample inference

Cerioli et al. (2018), propose to use the information available in the sample to evaluate the proportion of outliers and to better calibrate important quantities such as the breakdown point, a tuning parameter for the robust covariance estimator. This is a less convincing approach due to the well-known related problem of overfitting. Broadly speaking, results of a statistical analysis should be generalizable to other outcomes, equivalently, to the population from which the sample is drawn. In other words, the out-of-sample validity should be somehow assessed. Without out-of-sample validation, a statistical procedure will necessarily tend to excessively focus on the sample data, creating some type of overfitting that includes the sampling error (noise) together with the true underlying quantity of interest.

Out-of-sample inference is in particular at the basis of model selection procedures where the problem of overfitting is well recognised. Model selection criteria have also been developed that contain a penalty for out-of-sample validation. Mallows (1973) C_p, Akaike (1974) Information Criterion (AIC), Schwarz (1978) Bayesian Information Criterion (BIC), and related refinements (see e.g. McQuarrie and Tsai 1998), are the most popular. Efron (2004) developed a general framework for covariance penalty criteria that allows, for a given loss function (such as the squared loss function) to derive an estimator of the penalty for out-of-sample validity. Efron (2004) in particular shows that, given a predicted value $\hat{Y}_i$, considering the squared loss function, the penalty term for the the sample prediction error is $2\mathrm{cov}\left(\hat{Y}_i; Y_i\right)$. The loss function can be chosen as a weighted prediction error as is done, for the linear regression model, in e.g. Ronchetti and Staudte (1994) who derived the penalty associated to the sample weighted prediction error, therefore proposing a robust version of the C_p.

Other robust model selection methods have been proposed in the literature which include Machado (1993) for the BIC, Ronchetti et al. (1997) for cross-validation, Khan et al. (2007) for a robust LARS algorithm, Dupuis and Victoria-Feser (2011) and Dupuis and Victoria-Feser (2013) for fast robust search and very recently Avella Medina and Ronchetti (2017) for generalized linear models.

4 Final remarks

In this discussion, we limited our extensions to situations where the multivariate normal model with, possibly, a structure on the mean and covariance matrix, is of interest. As robust monitoring is particularly effective with high breakdown estimators, it is natural to wonder what can be done beyond the multivariate normal case. References are sparse but include Bianco et al. (2005) who proposed S and MM-estimators for a class of asymmetric models, e.g. the log-gamma distribution; Salibian-Barrera and Yohai (2008) who extended S-estimators to robust regression with censored data. Once again, robust monitoring is worth considering in these settings as they would bring new insight on the data. Finally, in high dimensional data, multivariate contamination may take a different form to the one usually assumed, i.e. outlying measurements may exist in such a way that the majority of observations are contaminated in at least one of their components. Van Aelst et al. (2012) adapt the Stahel–Donoho estimator by huberizing the data before the outlyingness is computed. They show that their proposal could better withstand large numbers of outliers. It would be interesting to see whether monitoring can be adapted to this problem. Very recently, Van Aelst and Wang (2017) robustified sure independence screening, a procedure used for variable selection in ultra-high dimensional regression analysis. Their approach is a very fast screening method using least trimmed squares principal component analysis to estimate the latent factors and the factor profiled variables. Variable screening is then performed on factor profiled variables by using regression MM-estimators. This brand new procedure may be amenable to monitoring.

References

Agostinelli C, Leung A, Yohai VJ, Zamar RH (2015) Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. Test 24:441–461

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Alqallaf F, Van Aelst S, Yohai VJ, Zamar RH (2009) Propagation of outliers in multivariate data. Ann Stat 37:311–331

Avella Medina M, Ronchetti E (2017) Robust and consistent variable selection in high-dimensional generalized linear models. Biometrika **(To appear)**

Bianco AM, Garcia Ben M, Yohai VJ (2005) Robust estimation for linear regression with asymmetric errors. Can J Stat 33(4):511–528

Butler R, Davies P, Jhun M (1993) Asymptotics for the minimum covariance determinant estimator. Ann Stat 21:385–1400

Candès EJ, Li X, Ma Y, Wrigh J (2011) Robust principal component analysis? J ACM 58(3), Article number 11

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl **(in press)**

Cheng T-C, Victoria-Feser M-P (2002) High breakdown estimation of multivariate mean and covariance with missing observations. Br J Math Stat Psychol 55:317–335

Chervoneva I, Vishnyakov M (2011) Constrained S-estimators for linear mixed effects models with covariance components. Stat Med 30(14):1735–1750

Chervoneva I, Vishnyakov M (2014) Generalized S-estimators for linear mixed effect models. Stat Sin 24:1257–1276

Copt S, Heritier S (2007) A robust alternative to the F-test in mixed linear models based on MM-estimates. Biometrics 63:1045–1052

Copt S, Victoria-Feser M-P (2004) Fast algorithms for computing high breakdown covariance matrices with missing data. In: Hubert M, Pison G, Struyf A, Van Aelst S (eds) Theory and applications of recent robust methods. Statistics for industry and technology series, Birkhauser, Basel, pp 71–82

Copt S, Victoria-Feser M-P (2006) High breakdown inference for mixed linear models. J Am Stat Assoc 101:292–300

Critchley F, Schyns M, Haesbroeck G (2010) Relaxmcd: smooth optimisation for the minimum covariance determinant estimator. Comput Stat Data Anal 54:843–857

Croux C, Hasebroeck G (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika 87:603–618

Croux C, Ruiz-Gazen A (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited. J Multivar Anal 95:206–226

Croux C, Filzmoser P, Oliveira M (2007) Algorithms for projection pursuit robust principal component analysis. Chemometr Intell Lab Syst 87:218–225

Danilov M, Yohai VJ, Zamar RH (2012) Robust estimation of multivariate location and scatter in the presence of missing data. J Am Stat Assoc 107:1178–1186

Davies PL (1987) Asymptotic behaviour of S-estimators of multivariate location parametersand dispertion matrices. Ann Stat 15:1269–1292

Devlin SJ, Gnanadesikan R, Kettenring JR (1975) Robust estimation and outlier detection with correlation coefficients. Biometrika 62:531–545

Devlin SJ, Gnanadesikan R, Kettenring JR (1981) Robust estimation of dispersion matrices and principal components. J Am Stat Assoc 76:354–362

Donoho DL (1982) Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Department of Statistics, Harward University

Dupuis Lozeron E, Victoria-Feser M-P (2010) Robust estimation of constrained covariance matrices for confirmatory factor analysis. Comput Stat Data Anal 54:3020–3032

Dupuis DJ, Victoria-Feser M-P (2011) Fast robust model selection in large datasets. J Am Stat Assoc 106:203–212

Dupuis DJ, Victoria-Feser M-P (2013) Robust vif regression with application to variable selection in large datasets. Ann Appl Stat 7:319–341

Efron B (2004) The estimation of prediction error. J Am Stat Assoc 99(467):619–632

Farcomeni A (2014a) Snipping for robust k-means clustering under component-wise contamination. Stat Comput 24:909–917

Farcomeni A (2014b) Robust constrained clustering in presence of entry-wise outliers. Technometrics 56:102–111

Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 29:81–124

Hampel FR, Ronchetti E, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York

Heritier S, Ronchetti E (1994) Robust bounded-influence tests in general parametric models. J Am Stat Assoc 89(427):897–904

Heritier S, Cantoni E, Copt S, Victoria-Feser MP (2009) Robust methods in biostatistics. Wiley, New York

Huber P, Ronchetti E (2009) Robust statistics, 2nd edn. Wiley, New York

Hubert M, Rousseeuw PJ, Branden K (2005) ROBPCA: a new approach to robust principal component analysis. Technometrics 47:64–79

Hubert M, Rousseeuw PJ, Verdonck T (2009) Robust PCA for skewed data and its outlier map. Comput Stat Data Anal 53:2264–2274

Hubert M, Rousseeuw PJ, Verdonck T (2012) A deterministic algorithm for robust location and scatter. J Comput Graph Stat 21:618–637

Hubert M, Rousseeuw PJ, Segaert P (2015) Multivariate functional outlier detection. Stat Methods Appl 24:177–202

Kent JT, Tyler DE (1996) Constrained M-estimation for multivariate location and scatter. Ann Stat 24:1346–1370

Khan JA, Van Aelst S, Zamar RH (2007) Robust linear model selection based on least angle regression. J Am Stat Assoc 102:1289–1299

Koller M (2016) robustlmm: an R package for robust estimation of linear mixed-effects models. J Stat Softw 75(6):1–24

Leung A, Zhang H, Zamar R (2016) Robust regression estimation and inference in the presence of cellwise and casewise contamination. Comput Stat Data Anal 99:1–11
Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York
Little RJA, Smith PJ (1987) Editing and imputing for quantitative survey data. J Am Stat Assoc 82:58–68
Liu RY, Parelius JM, Singh K (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by liu and singh). Ann Stat 27:783–858
Lopuhaä HP (1989) On the relation between S-estimators and M-estimators of multivariatelocation and covariance. Ann Stat 17:1662–1683
Machado JAF (1993) Robust model selection and m-estimation. Econ Theory 9:478–493
Mallows CL (1973) Some comments on C_p. Technometrics 15(4):661–675
Maronna RA, Zamar RH (2002) Robust Estimates of Location and Dispersion for High-Dimensional Datasets. Technometrics 44(4):307–317
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, Chichester
Mavridis D, Moustaki I (2008) Detecting outliers in factor analysis using the forward search algorithm. Multivar Behav Res 43:453–475
Mavridis D, Moustaki I (2009) The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. J Comput Graph Stat 18:1016–1034
McQuarrie A, Tsai C (1998) Regression and time series model selection, vol 43. World Scientific, Singapore
Moustaki I, Victoria-Feser M-P (2006) Bounded-bias robust inference for generalized linear latent variable models. J Am Stat Assoc 101:644–653
Olive D (2004) A resistant estimator of multivariate location and dispersion. Comput Stat Data Anal 46:99–102
Öllerer V, Alfons A, Croux C (2016) The shooting S-estimator for robust regression. Comput Stat 31:829–844
Rocke DM (1996) Robustness properties of S-estimators of multivariate location and shape in high dimension. Ann Stat 24:1327–1345
Rocke DM, Woodruff DL (1996) Identification of outliers in multivariate data. J Am Stat Assoc 91:1047–1061
Ronchetti E, Staudte RG (1994) A robust version of Mallows's C_p. J Am Stat Assoc 89:550–559
Ronchetti E, Field C, Blanchard W (1997) Robust linear model selection by cross-validation. J Am Stat Assoc 92:1017–1023
Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880
Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York
Rousseeuw PJ, Van den Bossche W (2017) Detecting deviating data cells. Technometrics. https://doi.org/10.1080/00401706.2017.1340909. **(in press)**
Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
Rousseeuw PJ, Yohai VJ (1984) Robust regression by means of S-estimators. In: Franke JW, Hardle W, Martin RD (eds) Robust and nonlinear time series analysis. Springer, New York, pp 256–272
Salibian-Barrera M, Yohai VJ (2006) A fast algorithm for s-regression estimates. J Comput Graph Stat 15(2):414–427
Salibian-Barrera M, Yohai VJ (2008) High breakdown point robust regression with censored data. Ann Stat 36(1):118–146
Salibian-Barrera M, Van Aelst S, Willems G (2006) PCA based on multivariate MM-estimators with fast and robust bootstrap. J Am Stat Assoc 101:1198–1211
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
Stahel WA (1981) Breakdown of covariance estimators. Technical report 31, Fachgruppe für Statistik, ETH, Zurich
Tatsuoka KS, Tyler DE (2000) The uniqueness of S and M-functionals under nonel liptical distributions. Ann Stat 28:1219–1243
Van Aelst S, Wang Y (2017) Robust variable screening for regression using factor profiling, manuscript
Van Aelst S, Willems G (2012) Robust and efficient one-way MANOVA tests. J Am Stat Assoc 106(494):706–718
Van Aelst S, Vandervieren E, Willems G (2012) A Stahel Donoho estimator based on huberized outlyingness. Comput Stat Data Anal 56:531–542
Woodruff DL, Rocke DM (1994) Computable robust estimation of multivariate location and shape in highdimension using compound estimators. J Am Stat Assoc 89:888–896

Xu H, Caramanis C, Sanghavi S (2012) Robust PCA via outlier pursuit. IEEE Trans Inf Theory 58:3047–3064
Xu H, Caramanis C, Mannor S (2013) Outlier-robust PCA: the high-dimensional case. IEEE Trans Inf Theory 59:546–572
Yohai VJ (1987) High breakdown point and high efficiency robust estimates for regression. Ann Stat 15:642–656
Zuo Y, Cui H (2005) Depth weighted scatter estimators. Ann Stat 33:381–413