TEST (2015) 24:484–488
DOI 10.1007/s11749-015-0457-z

DISCUSSION

Rejoinder on: Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination

Claudio Agostinelli[1] · Andy Leung[2] · Victor J. Yohai[3] · Ruben H. Zamar[2]

Published online: 27 June 2015

Abstract We thank the discussants, the referees, and the associate editor for their stimulating discussions and helpful remarks. We thank the editor for giving us the opportunity to discuss our paper in this journal. The rejoinder is organized in several sections, which address the main points raised by the discussants.

1 Computational issue

Christophe Croux and Viktoria Öllerer (CÖ), Ricardo A. Maronna (MA), Peter Rousseeuw and Wannes Van den Bossche (RV), Stefan Van Aelst (VA), and Roy E. Welsch (WE) comment on the high computational cost of the two-step generalized S-estimator (2SGS) and the need for faster alternatives.

MA, WE, and VA propose to separate the imputation step from the estimation to bypass the computational complexity of 2SGS. The idea is to first filter the outliers,

This rejoinder refers to the comments available at: doi:10.1007/s11749-015-0451-5; doi:10.1007/s11749-015-0452-4; doi:10.1007/s11749-015-0453-3; doi:10.1007/s11749-015-0454-2; doi:10.1007/s11749-015-0455-1; doi:10.1007/s11749-015-0456-0.

✉ Ruben H. Zamar
ruben@stat.ubc.ca

Andy Leung
andy.leung@stat.ubc.ca

[1] Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca’ Foscari di Venezia, San Giobbe, Cannaregio 873, 30121 Venezia, Italy

[2] Department of Statistics, University of British Columbia, 3182-2207 Main Mall, Vancouver, BC V6T 1Z4, Canada

[3] Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, Pabellón 1, 1426 Buenos Aires, Argentina

second impute the filtered cells, and third estimate the multivariate location and scatter matrix using some computationally efficient robust procedure, such as the fast-S or the fast-MCD. This approach is called "three-step estimator" by VA.

The first step of 2SGS (filter) is fast, but the second step is slow due to the computation of the generalized S-estimator (GSE) (Danilov et al. 2012). We notice that GSE first resamples the filtered data to compute an initial estimate and then iterates until convergence a sequence of robust imputation and estimation steps. These iterations can be computationally intensive and time consuming when a large fraction of the data has been filtered. However, the main computational burden in GSE comes from the computation of the initial robust estimator (extended minimum volume ellipsoid, EMVE) which is needed to achieve high robustness against casewise outliers.

EMVE is a generalized version of MVE, which was introduced by Rousseeuw (1985). The computation of EMVE consists of a combination of resampling and concentration steps. Once the estimators of location and scatter for a given subsample are obtained, the concentration step consists of computing the Gaussian MLE via the classical EM algorithm on the half of the observations with the smallest Mahalanobis distances. The concentration steps are time consuming, especially when there is a large number of filtered cells. This problem is aggravated by the required large number of subsamples, especially when p is large.

To save computing time, we now study the performance of a faster alternative to 2SGS, which we call Fast-2SGS. The Fast-2SGS is identical to 2SGS except that in the second step we now initalize GSE with a faster initial estimator instead of the time consuming EMVE. The faster initial estimator (inspired by the discussant's comments) is a simple three-step estimator described below.

Step I. *Filter cellwise outliers.* We flag cellwise outliers using the Gervini–Yohai filter and replace them by NA's.

Step II. *Impute filtered cells.* Following the discussants' suggestion, we impute the filtered cells using coordinate-wise medians. This simple imputation procedure seems sufficient, in our case, to obtain a fast and robust initial estimator. VA proposes a finer imputation approach: estimating the cell predictive distribution by fitting a robust simple regression of the missing variable on its most correlated counterpart. Unfortunately, in our naive implementation, this approach is very slow due to the inefficient looping in R needed to fit the different robust regressions. However, VA's suggestion seems interesting and deserves further study.

Step III. *Robustly estimate the multivariate location and scatter matrix.* We apply the MVE-S (see Maronna et al. 2006, Section 6.7.5) to the imputed data. The MVE-S is an S-estimator with bisquare ρ-function and MVE as initial estimator. We use MVE-S rather than other faster alternatives because this estimator best resembles the EMVE used in the original proposal.

This simple three-step estimator is called IS (for imputed data S-estimator) in the table and the figures below.

We conduct a simulation study to investigate the average computing time and robustness performance of Fast-2SGS and 2SGS. For comparison, we also included IS and the DetMCDScore proposed by RV. We consider several dimensions p and sample

Table 1 Average CPU time—in seconds of a 2.8 GHz Intel Xeon

Dim	ICM				THCM			
	2SGS	Fast-2SGS	DetMCDScore	IS	2SGS	Fast-2SGS	DetMCDScore	IS
10	0.70	0.10	0.01	0.08	0.36	0.08	0.01	0.06
20	6.93	0.50	0.03	0.28	1.48	0.24	0.03	0.15
30	31.51	1.72	0.07	0.68	4.26	0.59	0.06	0.22
40	91.21	4.41	0.12	1.34	8.24	1.42	0.12	0.29
50	210.05	9.37	0.20	2.30	12.83	3.36	0.20	0.38

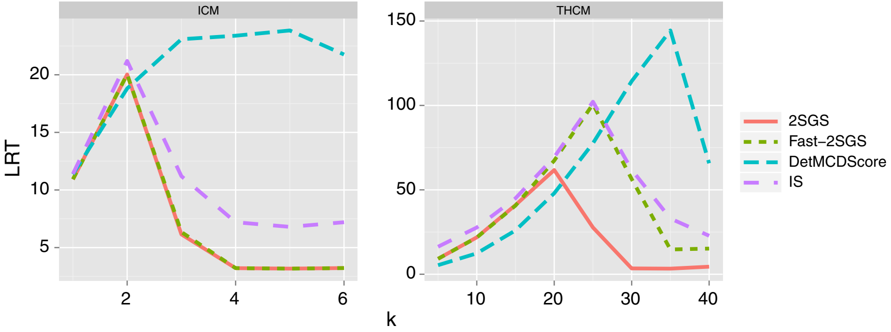

Fig. 1 Average LRT for various contamination values, k, from ICM and THCM. Results are based on $p = 40$ dimensional data sets

sizes $n = 10 \times p$. We use the random correlation structures described in Sect. 4 of the original paper. For each dimension, we average the results over 100 replications of the following setups: (a) cellwise contamination with $k = 1, 2, \dots, 6$ and (b) casewise contamination with $k = 5, 10, \dots, 100$. The range of k was chosen so that it contains the maximum average LRT. We consider the case of 10 % contamination.

Table 1 shows the mean time needed to compute the considered estimators for data with cellwise (ICM) and casewise outliers (THCM) as described in Sect. 4 of our paper. In general, DetMCDScores is the fastest and 2SGS the slowest. We also notice that Fast-2SGS is much faster than 2SGS.

Figure 1 shows the average LRT distances of the estimators for $p = 40$ dimensional data sets with different values of cellwise and casewise outliers. Similar results (not shown here) are obtained for the other considered dimensions. Under THCM, 2SGS still considerably outperforms the faster alternatives. Under ICM, Fast-2SGS performs very well (same as 2SGS). We also notice that under ICM, 2SGS, Fast-2SGS, and IS perform similarly in terms of maximum bias because the maximum occurs at a point ($k = 2$) where the filter fails to filter any cell. However, 2SGS and Fast-2SGS outperforms IS for data sets with larger cellwise outliers because the multivariate imputation of GSE gives better results in these cases.

If time considerations are important, one can use Fast-2SGS or DetMCDScore. Otherwise, we recommend to use 2SGS. Finding a fast and fully robust initial estimate for GSE would be of interest and deserves further investigation.

2 Precision matrix estimation for large and flat data sets

CÖ gives an extensive and detailed discussion of the performance of 2SGS in the case of large and flat data sets (large p and relatively small n). They focus on the context of precision matrix estimation.

Our numerical experiment confirms that, as pointed out by CÖ, 2SGS does not handle well large and flat data sets. In fact, when $n \leq 2p$, 2SGS fails to exist, likewise S-estimator and all second-generation robust estimators with breakdown point 1/2. When much data are filtered and the fraction of complete data is small, the iterations in GSE may fail to converge. In this case, GSE produces a nearly singular covariance matrix. This situation is more likely to occur for datasets with relatively small n compared to p. Danilov et al. (2012) provided a sufficient condition for the existence of GSE: the proportion of complete observations in general position to be larger than $1/2 + (p + 1)/n$. Numerical results have shown that GSE performs well for some smaller proportions of complete observations. However, no theoretical results are available for these cases.

To overcome the lack of convergence of 2SGS for large and flat data sets, we may partially impute the filtered cells to ensure a fraction $1/2 + (p + 1)/n$ of complete observations. More precisely, the procedure is to first filter outliers, then randomly select observations and impute the filtered cells using coordinate-wise medians, and finally estimate the location and scatter using GSE. Although this procedure is rather ad hoc, our initial numerical experiments suggests that it may work for $n \geq 5p$.

We now assess the performance of 2SGS with partial imputation (2SGS + PI) in the case of estimation of the precision matrix. We repeat the simulation experiment described in Sect. 1, but this time with $n = 5p$ instead of $10p$. For comparison, we also compute the GGQ precision matrix estimate proposed by CÖ. Table 2 shows the maximum average LRT distances.

It appears that partial imputation could expand the range of applicability of 2SGS to situations with $n \geq 5p$, but further research on this topic is still needed. When $n < 5p$, 2SGS + PI does not perform well and GGQ is a better alternative.

3 Controlling the robustness and efficiency of 2SGS for large p

MA makes a thoughtful remark regarding the loss of robustness of 2SGS–and in general, S-estimators with a fixed loss function ρ—when p is large. The Gaussian

Table 2 Maximum average LRT in the case of precision matrix estimation

Dim	ICM		THCM	
	2SGS + PI	GGQ	2SGS + PI	GGQ
10	5.7	6.9	7.5	7.3
20	7.6	9.4	9.1	10.1
30	9.8	11.6	11.4	12.5
40	11.7	14.1	13.3	15.0
50	13.8	17.2	15.7	17.3

efficiency of 2SGS systematically increases to one as p increases, but this gain in efficiency comes at the expense of a decrease in robustness. Hence, we agree with MA that for large p we may need to modify the GSE step to avoid the lack of robustness of S-estimators with fixed loss function. A possibility could be to use a well-calibrated MM-estimator of multivariate location and scatter (Tatsuoka and Tyler 2000) after adapting it for handling data with missing values. The resulting generalized MM-estimator would then gain robustness for p large. Another possibility could be to replace the bisquare rho function in GSE by a Rocke-type loss function, which changes with dimension in order to preserve the robustness of the estimator (Rocke 1996). A comparison of the behavior under outlier contamination of MM-estimators and Rocke type estimators with S-estimators based on a bisquare function may be found in Maronna and Yohai (2015). Again further work is needed on these topics.

4 Further development of the asymptotics

Alessio Farcomeni (FA) express interest on the asymptotics theory for 2SGS.

In the paper, we have shown that 2SGS is consistent at the central model provided the reference distribution for the filter is appropriately chosen. The development of robust inference using 2SGS, or some variation of it, would require the derivation of the asymptotic distribution of this estimator. We believe that this would be a worthwhile project for future research.

We conjecture that 2SGS inherits the desirable asymptotic properties of S-estimators established for the case of complete data (such as asymptotic normality). We, hence, believe that the proposed methodology of using a consistent filter followed by a classical robust method may open a way for robust estimation and inference under cellwise and casewise contamination.

5 Appropriate cut-off value for outlier detection

RV question our choice of cut-off values for detecting outliers in Sect. 5 of the original paper. The multiple comparison correction might be appropriate in our example because we want to determine the largest Mahalanobis distances likely to appear in the absence of cellwise, pairwise, and casewise outliers.

References

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

Maronna RA, Yohai VJ (2015) Robust and efficient estimation of high dimensional scatter and location. arXiv:1504.03389 [mathST]

Maronna RA, Martin RD, Yohai VJ (2006) Robust statistic: theory and methods. Wiley, Chichister

Rocke DM (1996) Robustness properties of S-estimators of multivariate location and shape in high dimension. Ann Stat 24:1327–1345

Rousseeuw PJ (1985) Multivariate estimation with high breakdown point. In: Grossmann W, Pflug G, Vincze I, Wertz W (eds) Mathematical statistics and applications, vol B. Reidel Publishing Company, Dordrecht, pp 256–272

Tatsuoka KS, Tyler DE (2000) On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. Ann Stat 28:1219–1243