TEST (2015) 24:462–466
DOI 10.1007/s11749-015-0455-1

DISCUSSION

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

Christophe Croux[1] · **Viktoria Öllerer**[1]

Published online: 27 June 2015

We would like to congratulate the authors for their innovative paper, containing many stimulating ideas. The authors propose an estimator for multivariate location and scatter, robust to both cellwise and casewise contamination. The idea is simple: (i) check for large univariate outliers and replace these cellwise outliers by NA, and (ii) apply the S-estimator for missing values of Danilov et al. (2012). If there are no cellwise outliers detected in step (i), the proposed estimator equals the regular S-estimator and shares the affine equivariance property. The authors will agree that the main power of the estimator comes from the second step, where the casewise—or multivariate—outliers are detected using a robust version of the Mahalanobis distance. For every observation, this distance is computed in the dimension given by the number of non-missing components. Danilov et al. (2012) present a smart way to compute an S-estimator associated with Mahalanobis distances computed in different dimensions.

Estimation of the scatter matrix is 'a corner stone in many applications', as the authors state. However, the applications that the authors list (principal component analysis, factor analysis, and multiple linear regression) require the precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ rather than the covariance matrix $\boldsymbol{\Sigma}$. Obviously, the inverse of the proposed two-step generalized S-estimator (TSGS) yields an estimate of the precision matrix. In this discussion note we (i) investigate the performance of TSGS as precision matrix estimator by means of a modest simulation study, (ii) discuss a regularized version of

This comment refers to the invited paper available at doi:10.1007/s11749-015-0450-6.

✉ Christophe Croux
christophe.croux@econ.kuleuven.ac.be; christophe.croux@kuleuven.be

Viktoria Öllerer
viktoria.oellerer@kuleuven.be

[1] Faculty of Economics and Business, KU Leuven, Leuven, Belgium

TSGS, and (iii) make a comparison with an estimator recently proposed by Öllerer and Croux (2014), called the GGQ-estimator.

1 The GGQ-estimator

Öllerer and Croux (2014) propose a simple robust covariance matrix estimator $\mathbf{S} = (s_{jk}) \in \mathbb{R}^{p\times p}$.

(1) Compute the robust scale estimators Q_n of (Rousseeuw and Croux 1993) for each variable.
(2) Compute standard correlations from the normal scores

$$r_{\text{Gauss}}(\mathbf{X}^j, \mathbf{X}^k) = \frac{\sum_{i=1}^n \Phi^{-1}\left(\frac{R(X_{ij})}{n+1}\right)\Phi^{-1}\left(\frac{R(X_{ik})}{n+1}\right)}{\sum_{i=1}^n \left(\Phi^{-1}(\frac{i}{n+1})\right)^2},$$

where $R(X_{ij})$ denotes the rank of X_{ij} among all elements of $\mathbf{X}^j$, the jth column of the data matrix, and where Φ is the cumulative distribution function a standard normal. This correlation measure is called the Gaussian rank correlation, and its robustness properties are studied in Boudt et al. (2012).
(3) Compute the robust covariance matrix $\mathbf{S}$ as

$$s_{jk} = Q_n(\mathbf{X}^j)Q_n(\mathbf{X}^k)r_{\text{Gauss}}(\mathbf{X}^j, \mathbf{X}^k). \tag{1}$$

The estimator $\mathbf{S}$ is consistent at normal distributions and semipositive definite. In high dimensions, or when the sample size n is close to or larger than the dimension p, we perform an additional regularization step, as in Tarr et al. (2015).
(4) Use $\mathbf{S}$ as input for the graphical lasso or GLASSO of Friedman et al. (2008). In mathematical terms,

$$\hat{\boldsymbol{\Theta}}_{\mathbf{S}}(\mathbf{X}) = \underset{\substack{\boldsymbol{\Theta}=(\theta_{jk})\in\mathbb{R}^{p\times p}\\ \boldsymbol{\Theta}\succ 0}}{\arg\max} \log\det(\boldsymbol{\Theta}) - \text{tr}(\mathbf{S}\boldsymbol{\Theta}) - \rho\sum_{j,k=1}^{p}|\theta_{jk}|, \tag{2}$$

where we maximize over all positive definite symmetric matrices, so $\boldsymbol{\Theta} \succ 0$, and where ρ is a penalty parameter to be selected.

The resulting estimator of the precision matrix is called the *GlassoGaussQn (GGQ)*. It is shown in Öllerer and Croux (2014) that the proposed estimator features a high breakdown point under cellwise contamination. The estimator can be computed very fast, even in high dimensions.

2 Precision matrix estimation

To evaluate the performance of the TSGS for precision matrix estimation, we redo the Monte Carlo simulation study of Agostinelli (2015) for $p = 10$ and $n = 100$. We

Table 1 Average LRT distances of precision matrix estimators. Results are based on 500 replicates

	Estimator	ICM			THCM	
		0 %	5 %	10 %	5 %	10 %
$p = 10, n = 100$	MLE	0.67	55.15	62.24	5.82	6.62
	TSGS	0.87	0.97	1.18	0.89	1.03
	GGQ$^{\nabla}$	0.80	2.85	4.01	4.07	4.77
$p = 20, n = 100$	MLE	3.09	106.72	120.94	8.40	9.72
	TSGS	3.73	4.70	>1000	3.98	4.93
	GGQ$^{\nabla}$	3.17	5.29	6.86	7.12	8.23
	MLE + GLASSO$^{\circ}$	4.59	116.56	127.07	9.38	11.19
	TSGS + GLASSO$^{\circ}$	4.82	6.79	9.06	7.22	8.95
	GGQ$^{\circ}$	4.75	7.09	9.22	6.82	7.75
$p = 100, n = 100$	MLE+GLASSO$^{\circ}$	28.40	573.98	>1000	30.72	31.71
	TSGS	NA	NA	NA	NA	NA
	GGQ$^{\circ}$	28.98	35.81	43.21	30.75	31.81
$p = 200, n = 100$	MLE+GLASSO$^{\circ}$	58.70	>1000	>1000	60.19	61.10
	TSGS	NA	NA	NA	NA	NA
	GGQ$^{\circ}$	59.61	71.97	86.13	60.52	61.46

$^{\nabla}$ ρ fixed at zero
$^{\circ}$ ρ selected over logarithmic grid of ten values using CV

limit ourselves to a contamination size of $k = 100$. The performance of the estimator is assessed by the average of likelihood ratio test distance

$$D(\widehat{\boldsymbol{\Sigma}^{-1}}, \boldsymbol{\Sigma}_0^{-1}) = \text{trace}(\boldsymbol{\Sigma}_0 \widehat{\boldsymbol{\Sigma}^{-1}}) - \log\det(\boldsymbol{\Sigma}_0^{-1} \widehat{\boldsymbol{\Sigma}^{-1}}) - p$$

over the $N = 500$ simulation runs. We compare with the maximum likelihood estimator (MLE), i.e., the inverse of the sample covariance matrix, and with the GGQ without regularization. The TSGS is implemented in the R-package `GSE` (Leung et al. 2014).

From the first three lines of Table 1, we see that of all three estimators MLE is doing best for clean data. For any type of contamination, however, it breaks down. Under contamination, TSGS yields lowest numbers of average LRT distance, thus, giving best results. Also GGQ gives reliable results, but its average LRT distance is higher.

The second part of Table 1 shows the results for $p = 20$, $n = 100$. Again MLE is doing best for clean data, but the other two estimators are close by. For casewise contamination (THCM), TSGS is performing best. Also for 5 % of cellwise contamination (ICM) the average LRT distance of TSGS is lowest of all estimators. However, for 10 % of cellwise contamination, the TSGS precision matrix estimator runs into computational problems. The estimated covariance matrices contain sometimes eigenvalues close to zero, causing their inverses to have very large elements. Interestingly, even in this case, the estimated covariance is rather precise.

Increasing the number of variables further to $p = 30$, but keeping the number of observations fixed at $n = 100$ led to many replicates where the estimator could not be computed anymore, or where we encountered convergence problems.

3 Regularized estimation of the precision matrix

To overcome the problems caused by close to singular covariance matrices, a regularization step can be added. Similarly as for the GGQ, a regularized TSGS precision matrix estimator can be obtained replacing $\mathbf{S}$ with the TSGS estimator in (2).

We select the penalty parameter ρ through 5-fold cross-validation over a logarithmic spaced grid of ten values from $0.1\rho_{\max}$ to $\rho_{\max}$, where ρ_{max} depends on the values of the covariance $\mathbf{S}$:

$$\rho_{\max} = \max\left(\max_{(i,j)\in\{1,\ldots,p\}^2}(\mathbf{S}-\mathbf{I}_p)_{ij} - \min_{(i,j)\in\{1,\ldots,p\}^2}(\mathbf{S}-\mathbf{I}_p)_{ij}\right).$$

This grid is proposed in the R-package `huge`-package (Zhao et al. 2014) that we use for computing the GLASSO in (2). The cross-validation criterion to be minimized is then the average log likelihood

$$\frac{1}{5}\sum_{k=1}^{5}\left\{-\log\det\hat{\boldsymbol{\Theta}}_{\rho}^{(-k)} + \operatorname{tr}(\mathbf{S}^{(k)}\hat{\boldsymbol{\Theta}}_{\rho}^{(-k)})\right\}, \tag{3}$$

where $\hat{\boldsymbol{\Theta}}_{\rho}^{(-k)}$ is the precision matrix estimate on the data with block k left out using penalty ρ, and $\mathbf{S}^{(k)}$ is a covariance estimate computed from the data of block k. Both for TSGS and for GGQ we use for $\mathbf{S}^{(k)}$ the covariance matrix estimator defined in (1). The reason is that block k consists only of $n/5$ observations which will cause computational problems when trying to compute TSGS on the data of block k only. To avoid those problems, we use instead the robust covariance matrix estimator (1) that can be computed in any dimension, also for small samples. To select the penalty parameter for the regularized ML estimator, $\mathbf{S}^{(k)}$ is the sample covariance matrix computed from block k.

Table 1 gives the results of a Monte Carlo study for $p = 20$ and $n = 100$. We use the same setting as before. Now we do not invert the covariance matrix estimates obtained by MLE and TSGS to estimate the precision matrix, but instead use them as an input for GLASSO, leading to MLE+GLASSO (the original graphical lasso) and TSGS+GLASSO, respectively. We also add the GGQ, but now with a penalty parameter different from zero.

We see that the regularization performed by GLASSO solves the singularity problem of TSGS: the average LRT distance is now lowest of all three estimators for cellwise contamination, also for 10 % of contamination. Note, however, that regularization introduces a bias. In comparison to the unregularized case, the LRT distances increased in all cases where the covariance matrices estimated by TSGS were not close to singular. Therefore, regularization needs to be used with care.

Additionally, we repeated the simulation study with an even higher number of variables, $p = 100$ and $p = 200$, while keeping the number of observations fixed at

$n = 100$. We observed that TSGS cannot be computed anymore. The only precision matrix estimates that are still computable are MLE+GLASSO and GGQ. We see that the latter two estimators perform similarly for clean data, with a slight advantage for the MLE+GLASSO. Under cellwise contamination, however, the MLE+GLASSO breaks down. It is remarkable that the MLE+GLASSO still gives reasonable results under casewise contamination (the regularization imposed by GLASSO seems to result in a robustification), at least for this simulation design.

4 Conclusion

The proposed two-stage generalized S-estimator is a precise and robust estimator, both in the presence of under cellwise and casewise contamination. We have tried several types of contamination schemes in additional, unreported simulation experiments and TSGS was performing very well in all of them. However, there are some limitations: (i) If the small sample n is not so much larger than $2p$, TSGS is nearly singular. (ii) If the sample size is smaller than about $2p$, TSGS cannot be computed anymore. In case (i), regularization gives a possible solution. In case (ii), estimators as the GGQ provide an alternative.

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. doi:10.1007/s11749-015-0450-6

Boudt K, Cornelissen J, Croux C (2012) The gaussian rank correlation estimator: robustness properties. Stat Comput 22(2):471–483

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(499):1178–1186

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441

Leung A, Danilov M, Yohai V, Zamar R (2014) GSE: robust estimation of multivariate location and scatter in the presence of missing data. http://CRAN.R-project.org/package=GSE. R package version 3.1

Öllerer V, Croux C (2015) Robust high-dimensional precision matrix estimation. In: Nordhausen K, Taskinen S (eds) Modern Multivariate and Robust Methods. Springer, pp 329–354 (to appear)

Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. J Am Stat Assoc 88(424):1273–1283

Tarr G, Müller S, Weber NC (2015) Robust estimation of precision matrices under cellwise contamination. Comput Stat Data Anal. doi:10.1016/j.csda.2015.02.005

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L (2014) Huge: high-dimensional undirected graph estimation. http://CRAN.R-project.org/package=huge. R package version 1.2.6