Supplementary Material for the paper:

Consistency factor for the MCD estimator at the Student-t distribution

Lucio Barabesi
University of Siena
Department of Economics and Statistics, Siena, Italy
and
Andrea Cerioli
University of Parma
Department of Economics and Management
and University Centre "Robust Statistics Academy" (Ro.S.A.), Parma, Italy
and
Luis Angel García-Escudero
University of Valladolid
Department of Statistics and Operations Research
and IMUVA, Valladolid, Spain
and
Agustín Mayo-Iscar
University of Valladolid
Department of Statistics and Operations Research
and IMUVA, Valladolid, Spain

1 Description

We provide details and additional empirical evidence that supplement the data analysis described in §5 of the manuscript. In particular, for each application, we give:

- the reference of the repository where the original data set was retrieved and a link through which the specific data under investigation can be accessed;
- a link to the Monte Carlo estimates of the expectation of the ordered squared robust distances $\widetilde{d}^2_{(i),\beta_n}$, computed through Algorithm 1;
- further empirical results of the analysis described in the manuscript.

All the computations described in the manuscript and in this supplementary file are performed through a Fortran code which makes use of the Fast-MCD algorithm of Rousseeuw and Van Driessen (1999), as implemented in the R package `robustbase` (Todorov and Filzmoser, 2009) and following the previous (normal-based) analysis of Cerioli et al. (2009) and Cerioli (2010). A similar implementation is also available in the FSDA toolbox for MATLAB, which is an extensive package for robust statistical analysis developed by the University of Parma and the European Commission's Joint Research Centre and freely distributed by The Matlab Central (`https://www.mathworks.com/matlabcentral/fileexchange/72999-fsda`).

Our code makes use of some numerical algorithms, which are necessary to evaluate the inverse of the regularized Beta function and to compute the integral displayed in formula (10) of the paper. We remark that the required numerical algorithms are nowadays standard and are available through virtually all programming languages. For instance, our Fortran implementation takes advantage of IMSL™ libraries for numerical computation, including function `DBETAI` for the evaluation of the incomplete Beta function ratio and function `DQDAGS` for numerical integration. Alternative non-proprietary implementations are also available: see, e.g., function `XINBTA` developed and listed by Cran et al. (1977), which is the source of many libraries. Similarly, we take advantage of the same IMSL™ libraries for random number generation and computation of the Kolmogorov-Smirnov statistic.

Portable Matlab and R implementations of the code used in this manuscript are currently under development. Calls to numerical functions similar to those listed above will still be necessary, but they will be easily available in such languages as well. See, for instance, functions `betaincinv` (Matlab) and `qbeta` (R) for inverting the cumulative distribution function of a Beta distribution, and functions `integral` (Matlab) and `integrate` (R) for performing numerical integration.

2 Image denoising

2.1 Data and simulated distances

The Image Segmentation Data Set, from which the data analyzed in §5.1 of the manuscript are derived, is available from the UCI Machine Learning Repository (Dua and Graff, 2022). At the following address

`https://univpr-my.sharepoint.com/:f:/g/personal/andrea_cerioli_unipr_it/`
`Eh6mTrZIK89KgSHexb6tWtsBSgvYBX8i9G8H1rMJJoadJQ?e=YcwrGj`

we have created a specific repository for the `GRASS` image, before and after contamination: see files

`Grass.prn`

and

`Grass_contaminated.prn`,

respectively. In these files, the first column corresponds to the variable denoted as *rawred-mean* in the original data set, the second column corresponds to *rawblue-mean*, while the third column corresponds to *rawgreen-mean*.

The Monte Carlo estimates of the expectation of the ordered squared robust distances $\widetilde{d}^2_{(i),\beta_n}$, computed through Algorithm 1, are also available at the same address, both for the original `GRASS` image ($m_n = n = 330$) and for the same image after contamination ($m_n = 300$): see files

`Robust_distances_Grass.txt`

and

`Robust_distances_Grass_contaminated.txt`,

respectively. Finally, files

`Classical_distances_Grass`

and

`Classical_distances_Grass_contaminated`

report the analogous Monte Carlo estimates computed from the full-sample estimators $\widehat{\mu}$ and $\widehat{\Sigma}$, which are obtained with $\alpha_n = 0$, $w_{i,0} = 1$ for $i = 1, \dots, n$, and $\eta_{0,p} = 1$.

2.2 Additional empirical evidence

Figure 1 shows the index plot of the ordered squared robust distances for the contaminated `GRASS` image. The choice $m_n = 300$ clearly emerges from the left-hand panel, even without the help of a formal outlier detection rule. A milder form of occasional contamination in the supposedly homogeneous part of the image may instead be suspected from the right-hand panel.

The estimates of μ and Σ computed for the original `GRASS` image under Assumption 3 are

$$\widetilde{\mu}_{\alpha_n} = (10.7, 12.1, 18.3)^\top$$

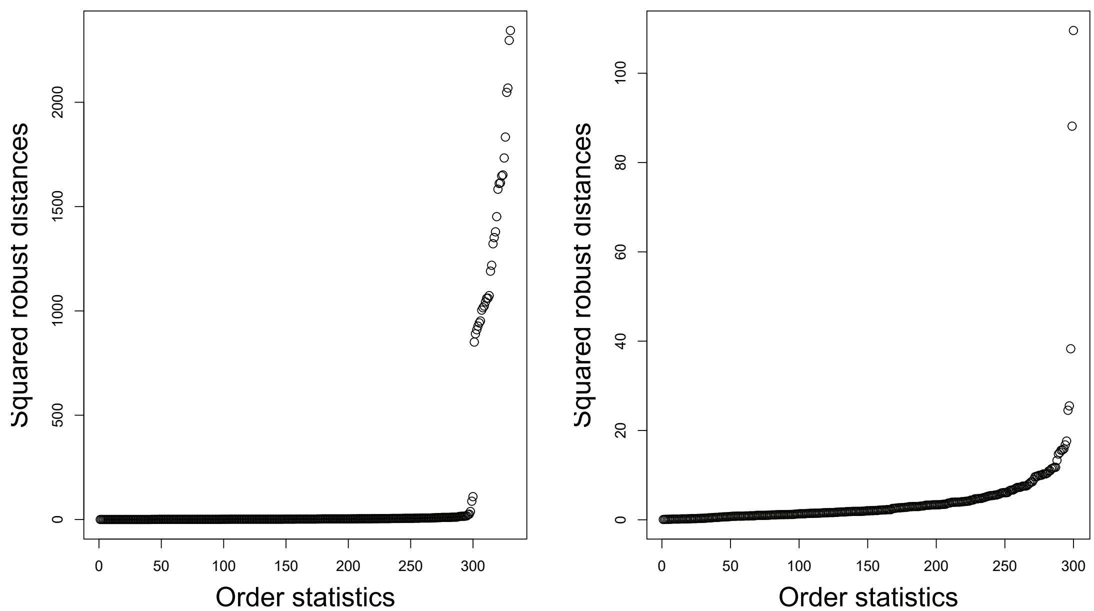

Figure 1: Index plot of the ordered squared robust distances for the contaminated `GRASS` image. Left: plot of all the distances. Right: zoom of the plot excluding the contaminated portion of the image.

and

$$\widetilde{\Sigma}_{\alpha_n}(5) = \begin{pmatrix} 40.7 & 50.9 & 56.4 \\ 50.9 & 67.9 & 71.6 \\ 56.4 & 71.6 & 80.1 \end{pmatrix},$$

where we plug in the estimate $\check{\nu} = 5$ in the consistency factor from Proposition 1. They are indeed very similar to those obtained for the contaminated `GRASS` image under the same Assumption:

$$\widetilde{\mu}_{\alpha_n} = (10.8, 12.2, 18.4)^\intercal$$

and

$$\widetilde{\Sigma}_{\alpha_n}(5) = \begin{pmatrix} 37.4 & 46.8 & 52.9 \\ 46.8 & 63.1 & 66.3 \\ 52.9 & 66.3 & 77.3 \end{pmatrix}.$$

Of course, the same conclusion does not hold for the non-robust estimates $\widehat{\mu}$ and $\widehat{\Sigma}$, which differ widely before and after contamination.

3 Stock index returns

3.1 Data and simulated distances

The data on daily stock index returns for the Canadian and UK equity markets considered in §5.2 of the manuscript are available from Li (2018):

`https://doi.org/10.1016/j.dib.2017.12.045`.

At the following address

`https://univpr-my.sharepoint.com/:f:/g/personal/andrea_cerioli_unipr_it/EsvMxJbPOCxKiwKOt3MERHUBOZDHmXqTFKoyLZxxQlZWQA?e=vOfWZW`

we have created a specific repository for the two-years data analyzed in this work, before and after contamination: see files

`Returns_1991_1992_CanUK.prn`

and

`Returns_1991_1992_CanUK_contaminated.prn`,

respectively. In these files, the first column corresponds to the stock index returns for the Canadian equity market, while the second column corresponds to the stock index returns for the UK equity market.

The Monte Carlo estimates of the expectation of the ordered squared robust distances $\widetilde{d}^2_{(i),\beta_n}$, computed through Algorithm 1, are also available at the same address, both for the original two-years data of returns ($m_n = n = 520$) and for the same data after contamination ($m_n = 475$): see files

`Robust_distances_Returns_1991_1992_CanUK.txt`

and

`Robust_distances_Returns_1991_1992_CanUK_contaminated.txt`,

respectively. Finally, files

`Classical_distances_Returns_1991_1992_CanUK.txt`

and

`Classical_distances_Returns_1991_1992_CanUK_contaminated.txt`

report the analogous Monte Carlo estimates computed from the full-sample estimators $\widehat{\mu}$ and $\widehat{\Sigma}$, which are obtained with $\alpha_n = 0$, $w_{i,0} = 1$ for $i = 1, \ldots, n$, and $\eta_{0,p} = 1$.

3.2 Additional empirical evidence

The full-sample estimates of μ and Σ for the original two-years bivariate series of uncontaminated stock index returns are

$$\widehat{\mu} = (-0.012, 0.013)^{\mathsf{T}}$$

and

$$\widehat{\Sigma} = \begin{pmatrix} 0.380 & 0.193 \\ 0.193 & 1.331 \end{pmatrix}.$$

The MCD estimates computed on the same data under Assumption 3 are

$$\widetilde{\mu}_{\alpha_n} = (-0.073, -0.073)^{\mathsf{T}}$$

and

$$\widetilde{\Sigma}_{\alpha_n}(6) = \begin{pmatrix} 0.309 & 0.213 \\ 0.213 & 1.563 \end{pmatrix},$$

where we plug in the tentative estimate $\check{\nu} = 6$ in the consistency factor from Proposition 1. The robust estimates of μ and Σ are similar to those obtained for the contaminated bivariate series of index returns:

$$\widetilde{\mu}_{\alpha_n} = (-0.035, -0.079)^{\mathsf{T}}$$

and

$$\widetilde{\Sigma}_{\alpha_n}(5) = \begin{pmatrix} 0.413 & 0.246 \\ 0.246 & 1.709 \end{pmatrix},$$

with a slight increase in variability due to the use of $\check{\nu} = 5$ as our refined estimate of ν in the consistency factor of Proposition 1. The refined estimate of ν is obtained by first inspecting the empirical distribution of the rescaled squared robust distances $\widetilde{d}^2_{i,\alpha_n,\nu}$ for values of ν close to its preliminary estimate and then selecting the contamination rate ε by comparing these distances to quantiles of F_T. The rescaled squared robust distances are given, for $\nu = 3, \ldots, 20$, in file

`Scaled_robust_distances_Returns_1991_1992_CanUK_contaminated.prn` ,

available at the data repository cited above. If we instead use $\check{\nu} = 6$ as our refined estimate of ν, we obtain

$$\widetilde{\Sigma}_{\alpha_n}(6) = \begin{pmatrix} 0.378 & 0.225 \\ 0.225 & 1.561 \end{pmatrix},$$

in close agreement with the full-sample estimate of Σ for the uncontaminated bivariate series.

Finally, the effect of the contamination induced by the simulated financial crisis is minor on the non-robust location estimate $\widehat{\mu} = (-0.104, -0.056)^{\mathsf{T}}$. It is instead huge on the non-robust dispersion estimate

$$\widehat{\Sigma} = \begin{pmatrix} 3.904 & 2.640 \\ 2.640 & 4.160 \end{pmatrix},$$

which is greatly affected by the increase of the variability of returns in both markets after Day 460. Therefore, even in this case where the shift contamination is minor, the increased volatility induced by the simulated financial shock makes $\widehat{\Sigma}$ completely unreliable for the purpose of estimating the bivariate structure of the two return series under Assumption 3.

References

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

Cerioli, A., Riani, M., and Atkinson, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*, 19:341–353.

Cran, G. W., Martin, K. J., and Thomas, G. E. (1977). Remark AS R19 and Algorithm AS 109: A remark on algorithms: AS 63: The incomplete beta integral AS 64: Inverse of the incomplete beta function ratio. *Journal of the Royal Statistical Society, Series C*, 26:111–114.

Dua, D. and Graff, C. (2022). UCI machine learning repository.

Li, L. (2018). Daily stock index return for the Canadian, UK, and US equity markets, compiled by Morgan Stanley Capital International, obtained from Datastream. *Data in Brief*, 16:947–949.

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

Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47.