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

Alessio Farcomeni
Sapienza - University of Rome

1 Introduction

The authors deal with a very stimulating problem, that of robust location and scatter estimation in presence of cell-wise and case-wise outliers. As I will try to argue in the next section, this in practice corresponds to the case of robust location and scatter estimation in moderate and high dimensions. It is difficult to think about an application where data are of moderate dimensionality and contamination can be expected to arise from the THCM alone.

I am impressed by their proposed estimator, which provides an interesting solution to the problem.

The rest of this brief discussion is as follows: in the next section I give some remarks on the contamination models considered. In Section 3 I discuss `snipEM` and in Section 4 I discuss 2GSE and possible extensions.

2 Cellwise and casewise contamination

Contamination is by definition unforseen under the model, and uncontrolled at the data collection stage. Formal contamination models are useful to test and evaluate robustness, but they must be sufficiently challenging and mimick possible real situations as much as possible.

As noted also by the authors, in classical low-dimensional settings there is not much difference between a THCM and an ICM with outliers in "general position". In moderate to high dimensional data sets, departures from model assumptions can be much more complex than in small dimensional settings. Robust approaches successful under THCM may not be feasible in these cases. A reason why this happens is that many methods require that at least 50% of the observations are free from contamination. As the dimensionality is increased, under ICM all of the observations might be contaminated, albeit only partly. To better support this statement, we report in Table 1 the proportion of clean observations under ICM when the probability of contamination in a given column

is ε, for different data dimensionality p. It can be seen that the proportion of clean values is well below 50% in most cases, and could even be zero in extremal situations. An approach based on trimming like the MCD may not be used in these situations.

p	5	10	20	50	80	100
$\varepsilon = 0.005$	0.98	0.95	0.90	0.78	0.67	0.61
$\varepsilon = 0.01$	0.95	0.90	0.82	0.61	0.45	0.37
$\varepsilon = 0.05$	0.77	0.60	0.36	0.08	0.02	0.01
$\varepsilon = 0.1$	0.59	0.35	0.12	0.01	0.00	0.00

Table 1: Proportion of clean observations under ICM, for different values of component-wise contamination ε and dimensionality p

A classical motivating example is that of applications in gene discovery. Gene expression levels are repeatedly measured on independent slides. The THCM model is supported when one or more genes arise from a different pathway than the one under study. Component-wise contamination arises due to systematic errors within each slide, so that one or more measurements are outlying but the "clean" expression levels agree with the population model. In this example even *column*wise outliers are possible, where a slide might yield corrupted measurements for most of the genes.

I believe that in general moderate dimensional data will often present a combination of THCM and ICM, with dependent generating processes.

I completely agree with the authors that a modern robust estimator of location and scatter should be able to adapt to potentially any configuration of case-wise and sparse cell-wise outliers, as long as contaminated data is sufficently far from the bulk of the data.

3 Snipping

The `snipEM` estimator is a special case of a procedure, `sclust`, proposed in Farcomeni (2014a) for clustering in k groups. The original idea of snipping can be found in Farcomeni (2014b), where a robust k-means procedure is proposed. Additional details on robust clustering methods based on trimming and snipping can be found in part II of the upcoming book Farcomeni and Greco (2015).

I have few points regarding `snipEM` and its performance.

First, it seems like the authors have implemented `snipEM` constraining the condition number of the estimated covariance matrix to be at most 12, while the condition number of the true matrix is fixed to 100. The performance of `snipEM` reported might be therefore pessimistic. Note that a restriction factor on the condition number might be needed only when $k > 1$, so it can be avoided with `snipEM`. Additionally, a limitation of `snipEM` is that its performance is dependent on the initialization strategy. A better initialization strategy might be given by flagging casewise outliers through a univariate detection method (e.g., using the boxplot rule) after vectorization of the data matrix.

Secondly, the efficency of estimators based on snipping or trimming (e.g., the MCD) might always be improved by a data-driven choice of the number of discarded values. In a brief simulation study under the same setting of the authors, in which the snipping level is chosen using the G statistic of Farcomeni (2009) (and no restrictions on the estimated condition number are imposed), the efficiency is increased from about 10% to 80% when $p = 10$ and $n = 100$, and from about 30% to 95% when $p = 20$ and $n = 200$. Another useful method for choosing trimming/snipping levels is proposed in García-Escudero *et al.* (2011).

Finally, snipping is explicitely thought for cell-wise contamination. In presence of case-wise contamination the best route in my opinion is to perform trimming (case removal), and optimize over the trimming set. An estimator working well under ICM, THCM and mixtures of the two models is readily available by combining snipping and trimming: a proportion of entries ε_1 is snipped, and a proportion of rows ε_2 is trimmed. A simultaneous selection of trimmed cases and snipped cells can be performed by optimization of the appropriate objective function, obtaining a `stEM` estimator. The `stEM` estimator is seen in a brief simulation study to outperform all competitors, including 2GSE, when the trimming and snipping levels are well tuned. This happens under ICM, THCM, and even when a combination of ICM and THCM are used to generate outliers. It can be speculated that the performance of `snipEM` and `stEM` is the best given they are optimizing over the snipping/trimming set.

The `snipEM` and `stEM` functions, together with routines for robust clustering (`skmeans` and `sclust`), can be obtained from the new `snipEM` package available on `CRAN`.

4 The 2GSE

It shall be noted that the issues of tuning and initialization are crucial for `snipEM` and many similar estimators, including `stEM`. The 2GSE, in addition to being computationally very efficient, does not seem to need basically any tuning and is much less dependent on initialization than snipping methods.

A brief simulation study comparing the 2GSE for a grid of values of $\alpha \geq 0.6$ reveals that it has approximately the same performance regardless of the tuning parameter, with slightly better results for larger values (e.g., $\alpha = 0.99$).

Additionally, the authors were very convincing in showing that 2GSE seems to work equally well with both kinds of contamination models.

Given the idea is very promising, I think the authors could explore the properties of the 2GSE in other contexts. First of all, the estimated scatter matrix can be used to perform robust PCA under ICM and THCM. This would lead to a whole new class of robust PCA methods for case-wise contamination. The 2GSE estimator can be similarly embedded into discriminant analysis procedures.

I was also wondering how difficult it would be to establish the distributional properties of the 2GSE estimator. If the (asymptotic) distribution of 2GSE and related Mahalanobis distances are obtained, many other extensions would be possible including development

of robust tests, outlier detection procedures, and reweighting schemes.

References

A. Farcomeni (2009). Robust double clustering: A method based on alternating concentration steps. *Journal of Classification*, **26**, 77–101.

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

A. Farcomeni (2014b). Snipping for robust k-means clustering under component-wise contamination. *Statistics and Computing*, **24**, 909–917.

A. Farcomeni and L. Greco (2015). *Robust Methods for Data Reduction*. Chapman & Hall/CRC Press, Boca Raton, FL.

L. A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2011). Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, **21**, 585–599.