TEST (2015) 24:482–483
DOI 10.1007/s11749-015-0452-4

DISCUSSION

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

Roy E. Welsch[1]

Published online: 27 June 2015

The authors are to be commended for bringing the critical problem of cellwise outliers to the attention of a broader community and providing some important new estimation methods and related theory. High dimensional data analysis has become a critical area for research in statistical theory and practice and, in such situations, removing (or severely downweighting) an entire observation for a single cellwise outlier can eliminate most of the data.

In practice, it can be difficult to verify the assumptions needed for the theorems that are proved or used in this paper. For example, how can I know that the fraction of outliers tends to zero as the number of observations tends to infinity? Even after transformation, the standard normal may not be appropriate and the tails may not be heavier than the actual distribution. And we have to live with the fact that modern robust algorithms suffer from computational uncertainty (local optima) as well as the usual statistical uncertainty associated with a sample.

I tell my students to start an analysis with a robust procedure as part of data exploration to see if any downweighted observations or observations with large residuals should be examined for errors, etc. If things look reasonable, then I often say they should try a classical procedure and compare the robust and classical estimates. If these two estimates are similar, then using a classical procedure might be appropriate. Therefore, a robust procedure is being used to diagnose data problems as well as a possible final estimator.

I often first look for a diagnostic approach to data analysis (data diagnostics). With potential casewise outliers, we can compare our results with and without the case (the

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

✉ Roy E. Welsch
rwelsch@mit.edu

[1] Massachusetts Institute of Technology, Cambridge, MA 02139, USA

whole case is NA) as in leave-one-case-out covariance and regression diagnostics. This is a rather brutal approach and each element of a case could be replaced by the median of the variable associated with each cell in that case and then the analysis redone. This, of course, increases the computational cost. These methods are, naturally, affected by swamping and masking.

The approach taken in this paper is to treat a potential cellwise outlier as NA and proceed from there. We could use one of the number of missing data algorithms to fill in the NAs and, therefore, reduce the need for specialized algorithms to deal with incomplete multivariate data. However, the authors note that missing data fill-in does not address casewise outliers and is not consistent. We could address casewise diagnostics with leave-one-case-out diagnostics discussed above and I am not sure just how much I should worry about asymptotic consistency for diagnostic purposes.

An outlier, leverage point, or influential cell is not missing, but it is interesting. To avoid making it NA, I often replace it with the median of all the observations for that variable as mentioned earlier in the casewise situation. Of course, I do not know what cells are outliers, leverage points, or influential points, so I would have to do this for every cell in a variable and over all variables (nxp). With each one of the nxp replacements, I recompute the covariance matrix estimate and then compare the covariance matrix with the cell unchanged with the covariance matrix with the cell replaced by the variable median using whatever measure of distance was appropriate (e.g., Kulback-Leibler or LRT, condition number, etc.) I then look at the distribution of these differences to get a rough idea of cells that are having an unusual impact on the estimated covariance matrix and examine them for errors, etc. This method is not optimal in any particular sense and requires some computation, but it is easy to explain to students and consulting clients. Over the years, I have found communication to be as important as many of the theorems of asymptotic statistics.

There are now a number of ways to find lower rank matrix approximations for sparse data. The NA methods suggested in the paper would, in some cases, lead to at least moderately sparse matrices and bringing together sparsity ideas and robustness might provide another approach to the problems considered here. This paper is an excellent start and opens the door to some exciting new research.