Discussion on the paper: Data Science, Big Data and Statistics, by Pedro Galeano and Daniel Peña.

Stefan Van Aelst¹, Ruben Zamar²

¹ KU Leuven, Belgium

² University of British Columbia, Canada

We praise Professors Galeano and Peña for this paper and for sharing their view on the impact of Big Data on Statistics and the emerging field of Data Science. They draw attention to seven main points which are very interesting and relevant, and present two interesting applications. We will focus our discussion on two of these topics, which are most related to our own expertise: heterogeneous data (including data quality and robustness) and automatic model selection.

Heterogeneous Data

Modern big data are often the result of administrative/operational data collection as it is the case for the two examples given by the authors. In a recent discussion paper, Professor David Hand (Hand 2018a, see also Hand 2018b) points out that this type of data also exhibits quality issues. Hence, there is a need for robust methods to automatically analyze such data. As pointed out by the authors, the standard statistical efficiency concept loses its meaning because the data form the whole population. On the other hand, data quality issues remain paramount. In fact, as mentioned in Hand (2018a) data can even be entered incorrectly for operational purposes. Hence, one should be aware that due to data quality issues different types of errors and outliers may occur in big data.

As pointed out by the authors classical robust statistics is not well suited for modern big data. In our opinion this is mainly because it has focused on the problem of casewise outliers. That is, cases that do not follow the overall pattern of the bulk of the data. First generation robust statistical procedures were built on the assumption that such anomalies affect only a minority of the cases. Classical robust methods in fact identify and downweight (or completely eliminate the effect of) outlying cases. However, downweighting complete cases becomes inefficient for high-dimensional data which contain relatively few cases in comparison to the number of variables. Often only a few cells in outlying cases are actually contaminated and the remaining cells still contain valuable information. Moreover, the increasing dimension of modern data drew attention to the phenomenon of propagation of outlying cells (Alqallaf et al. 2009). A small fraction of outlying cells can affect a very large fraction of cases in high dimensional data, easily exceeding 50% of the observations. This renders first generation robust methods incapable for reliable automatic processing of high-dimensional data. To overcome these shortcomings, new developments in robustness methodology and practice have emerged. These second generation robust procedures are more suited for big data processing such as robust estimation of location and scatter (Van Aelst et al. 2011, 2012, Agostinelli et al., 2015, Leung et al., 2017), robust regression (Öllerer et al., 2016), robust variable selection (Khan et al. 2007a, 2007b), robust screening (Wang and Van Aelst 2018), robust preprocessing of high-dimensional data (Rousseeuw and Van den Bossche, 2018).

As mentioned by the authors, heterogeneity is an important issue in big data. To address this issue, clustering methods and mixture models are commonly used. Moreover, as pointed out by the authors different subpopulations may concentrate around geometrical patterns more complex than single points, such as hyperplanes (Van Aelst et al., 2006) or manifolds. In this context outlying cases can appear as isolated observations or very small clusters. If left unnoticed, they may seriously

affect the resulting group membership and group structures obtained by clustering or mixture model techniques. Considerable research efforts has been devoted to these issues such as robust clustering with elliptical shapes (García-Escudero et al., 2008) and robust clustering around lines (García-Escudero et al., 2009). The important but difficult problem of robust methods that are resistant to both cellwise and casewise outliers in this context has not been addressed yet and requires further attention in the near future.

Automatic model selection

Big data often are also variable-rich data sets. When we have a large number of covariates available to predict a response, a common approach is to use a regularization tool such as all subsets selection, e.g. using AIC, stepwise variable selection, LARS, ridge regression, LASSO, etc. When the focus of the model is on prediction, classification or ranking, then the goal is to build a model with good out of sample generalization properties. Regularization then is a way to establish a trade off between increasing bias and a (hopefully larger) reduction in variance. Hopefully, redundant/useless variables are eliminated from the model, but some of the remaining useful variables may have their influence shrunk toward zero. However, the complexity of big data and the underlying processes imply that a single model such as a linear regression, logistic regression or CART is too simple to capture this complexity (involving interactions between the variables). Ensemble models such as random forests or boosting (see e.g. Hastie et al. 2009) have proven to be very powerful alternatives that can better capture such complexities. These ensembles are obtained by incorporating a type of randomness such as bootstrap and random variable selection. The use of random selection guarantees that individual models are sufficiently diverse. Moreover, the systematic search of optimal subset of models to combine in an ensemble may seem infeasible. However, for linear models a data-driven systematic selection of models is possible by solving a penalized optimality problem. Sparsity and diversity of the models in the ensemble are encouraged by using appropriate penalty functions (see Christidis et al. (2017)).

References

- Agostinelli, C., Leung, A., Yohai, V.J., and Zamar, R.H. (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination (with discussion). *TEST*, 24, 441–461.
- Alqallaf, F., Van Aelst, S., Yohai, V. J. and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *Annals of Statistics* 37(1), 311–331.
- Christidis, A., Lakshmanan, L.V.S., Smucler, E. and Zamar, R. (2017). Ensembles of Regularized Linear Models. [arXiv:1712.03561](https://arxiv.org/abs/1712.03561)
- García-Escudero, L.A., Gordaliza, A., San Martín, R., Van Aelst, S., and Zamar, R.H. (2009). Robust linear clustering. *Journal of the Royal Statistical Society B*, 71(1), 301–319.
- García-Escudero, L.A., Gordaliza, A., Matrán, C. and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *Annals of Statistics*, 36, 1324–1345.
- Hand, D. J. (2018a). Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society A*, 181, 555–605.
- Hand, D. J. (2018b). Hand writing: Administrative Data. *IMS Bulletin*, 47(6), 8–9.
- Hastie T., Tibshirani R. and Friedman J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer

- Khan, J.A., Van Aelst, S., and Zamar, R.H. (2007a). Building a Robust Linear Model with Forward Selection and Stepwise Procedures. *Computational Statistics and Data Analysis*, 52: 239–248.
- Khan, J.A., Van Aelst, S., and Zamar, R.H. (2007b), Robust Linear Model Selection Based on Least Angle Regression. *Journal of the American Statistical Association*, 102, 1289–1299.
- Leung, A., Yohai, V. J. and Zamar, R. (2017). Multivariate location and scatter matrix estimation under cellwise and casewise contamination, *Computational Statistics and Data Analysis*, 111, 59–76.
- Öllerer V., Alfons, A., and Croux, C. (2016). The shooting S-estimator for robust regression, *Computational Statistics*, 31(3), 829–844.
- Rousseeuw, P., and Van den Bossche, W. (2018). Detecting deviating data cells. *Technometrics*, 60(2):135–145.
- Van Aelst, S., Vandervieren, E., and Willems, G. (2011). Stahel-Donoho estimators with cellwise weights. *Journal of Statistical Computation and Simulation*, 81(1): 1–27.
- Van Aelst, S., Vandervieren, E., and Willems, G. (2012). A Stahel-Donoho estimator based on Huberized outlyingness. *Computational Statistics and Data Analysis*, 56(3): 531–542.
- Van Aelst, S., Wang, X., Zamar, R.H., Zhu, R. (2006), Linear grouping using orthogonal regression. *Computational Statistics and Data Analysis*, 50, 1287–1312.
- Wang, Y. and Van Aelst, S. (2017). Robust variable screening for regression using factor profiling. *Statistical Analysis and Data Mining*, to appear.