Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Comment on "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression"

Po-Ling Loh

To cite this article: Po-Ling Loh (2020) Comment on "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression", Journal of the American Statistical Association, 115:532, 1715-1716, DOI: 10.1080/01621459.2020.1837141

To link to this article: https://doi.org/10.1080/01621459.2020.1837141

Published online: 18 Dec 2020.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 532, 1715–1716: Theory and Methods Discussion
https://doi.org/10.1080/01621459.2020.1837141

Comment on "A Tuning-Free Robust and Efficient Approach to High-Dimensional Regression"

Po-Ling Loh

Department of Statistics, University of Wisconsin-Madison, Madison, WI

We congratulate the authors on a very exciting and timely piece of work. The use of a penalized dispersion score function for high-dimensional linear regression is extremely innovative, as it moves beyond the usual mold of penalized M-estimation and illuminates the benefits of considering broader classes of estimators. The properties of the authors' approach that (1) the theoretically optimal tuning parameter for ℓ_2-error bounds does not depend on unknown quantities; (2) the loss function does not itself involve a tuning parameter; and (3) the objective function is convex and therefore computationally feasible to minimize make the algorithm highly attractive and potentially useful in practice.

Here is what Huber had to say about regression estimators (Huber and Ronchetti 2009, sec. 7.12):

> In the robustness literature of the 1980s, most of the action was on the regression front…. Already the discussants of Bickel (1976) had complained about the multiplicity of robust procedures and about their conceptual and computational complexity. Since then, the collection of estimates to choose from has become so extensive that it is worse than bewildering, namely counterproductive.
>
> The problem with straightforward M-estimates (this includes the L_1-estimate) is of course that they do not safeguard against possible ill effects from leverage points. The other estimates were invented to remedy this, but their mere multiplicity already shows that no really satisfactory remedy was found.

Indeed, estimators designed to achieve small error in the presence of high-leverage points generally had the drawback of having a relatively low breakdown point. Other high-breakdown estimators had the drawback of being computationally infeasible in moderate dimensions. As discussed in this article, however, the question of using non-M-estimators deserves to be revisited for high-dimensional regression problems.

The results derived in this work raise several interesting questions, which we now discuss. First, in the language of robust statistics, the estimator analyzed here corresponds to finding a regression parameter vector which minimizes an L-estimate of scale for the residuals. As the authors point out, these estimators were studied in Jaeckel (1972), but their estimator only corresponds to *one particular* choice of L-estimators as motivated by nonparametric statistics. This motivates the question of how well the theory in this article generalizes to other estimators, or—perhaps more in the spirit of classical robust statistics—which estimator(s) would be optimal (in some appropriate sense) among a suitably defined class of estimators. The objective function would be defined as an ℓ_1-penalized version of a scale estimate based on residuals. A natural class to consider would be the class of all L-estimators—this might refer to linear combinations of order statistics $\{r_{(i)}(\beta)\}_{i=1}^n$ of residuals $r_i(\beta) = y_i - x_i^T\beta$, or linear combinations of arbitrary functions of residuals. Common examples of such estimators include the least median of squares (LMS) (Rousseeuw 1984) or least α-quantile (Yohai and Zamar 1993) estimators, which minimize the $\frac{1}{2}$- and α-quantiles of the empirical cdf of $\{|r_i(\beta)|\}_{i=1}^n$, respectively; least trimmed squares (LTS), which minimizes $\sum_{i=1}^h r_{(i)}^2(\beta)$, for some $h \leq n$; or least pth power deviations (Forsythe 1972). An ℓ_1-penalized version of LTS has been shown in a recent article to perform well empirically for high-dimensional regression (Alfons, Croux, and Gelper 2013), and a more careful study of the theoretical properties of the estimator and its comparison to the estimator proposed in this article would certainly be worthwhile. In addition to obtaining bounds on the statistical error of these estimators, it would be important to understand which of these estimators are computationally feasible in high dimensions.

Second, elaborating further on the topic of optimality, it is important to carefully consider the criteria by which we should declare an estimator to be optimally robust. In Remark 5 of their article, the authors have assessed the quality of their estimators in terms of asymptotic relative efficiency when the error variables are drawn iid from various heavy-tailed distributions. As they point out, obtaining an estimator which is fully efficient with respect to the MLE would be impractical in high dimensions. On the other hand, classical robust statistics theory suggests a slightly different viewpoint—considering the (asymptotic) variance of estimators when data are drawn iid from an ϵ-ball around an uncontaminated distribution. (The asymptotic theory derived by the authors stems from the second stage of their estimation method; thus, it seems the question would be whether a slightly different objective function might be preferred over the present proposal when viewed through this lens of optimality.) Another (potentially more complex) open

CONTACT Po-Ling Loh *loh@ece.wisc.edu* Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706.

question is whether optimality results could be proved in terms of the variance of the estimators in *finite* samples, since the focus of most high-dimensional literature, including the error bounds in this article, is on nonasymptotic results. One feature the authors highlight about their proposed method is that its theoretical properties remain valid when the error distributions are asymmetric, in which case the analysis of Jaeckel (1971) may be relevant to this discussion.

It is worth mentioning that the objective studied by the authors is related to the (penalized) class of *S*-estimators: Traditionally, *S*-estimators are obtained by minimizing an *M*-estimator of scale computed from the regression residuals (Rousseeuw and Yohai 1984), whereas here, the authors consider an *L*-estimate. In the context of location estimation, a comparison of different *L*-estimators of location was provided in Jaeckel (1971), but the discussion is limited to linear combinations of order statistics and not *functions* thereof. Perhaps a theoretical comparison between the performance of (penalized) *L*-estimators and *M*-estimators, which include the commonly used Huber estimator, would also be interesting.

Another interesting question is whether the authors believe that their method should be favored over the Lasso even when the error distribution is (sub-)Gaussian. In such a scenario, the canonical Lasso would already produce regression estimators with ℓ_2-error rates of the same order. On the topic of (low-dimensional) least squares in comparison to various proposals for robust linear regression, Stigler (2010) made the following eloquent remarks:

> Ever since the statistical world fully grasped the nature of what Fisher created in the 1920s with the analysis of multiple regression models and the analysis of variance and covariance—ever since about 1950—we have seen what that analysis can do and seen the *magic* of the results it permits. The perfection of that distribution theory, the ease of assessing additional variables and partitioning sums of squares as related to potential causes—no other set of methods can touch it in these regards. Least squares is still and will remain King for these reasons—these magical properties—even if for no other reason …least squares will remain the tool of choice unless someone concocts a robust methodology that can perform the same magic.

Consequently, we wonder whether the authors' proposed method should be recommended for use in all circumstances, or if it has any drawbacks worth considering from a theoretical viewpoint (e.g., deriving results for inference, variable selection consistency, etc.), due to the fact that it is nonsmooth and not an *M*-estimator.

Finally, we end with a more speculative question: A topic of current interest, particularly in the computer science community, is to devise and analyze estimators that are robust to *adversarial* perturbations. In such a setting, an iid sample is first drawn from a distribution, and then an adversary is allowed to change an ϵ proportion of points in a way that may depend on the original sample as well as the algorithm. The goal is to estimate an underlying parameter of the uncontaminated distribution. Although the statistical analysis of estimators in such settings may be appreciably different from the analysis of iid data, an interesting observation is that estimators which perform well for heavy-tailed data are often also useful in situations where data are adversarially contaminated (Minsker 2018; Lugosi and Mendelson 2019; Prasad et al. 2020). It would be an interesting direction to explore whether the estimators studied here can be analyzed and shown to be provably effective when applied to adversarially contaminated data, as well.

The theory of robustness has continued to receive renewed attention since Stigler's historical editorial (Stigler 2010), and the topics mentioned here are only a small sample of emergent challenges which have necessitated building a broader theoretical foundation. At the same time, advances in areas such as optimization have enabled a more complete analysis of various estimators which may have seemed intractable to an earlier generation of statisticians, while opening up the possibility of analyzing still more complicated estimators. Nonetheless, the insights of classical robust statistics, such as ideas from the theory of optimal robustness, may still be helpful for informing future research progress in these newer directions. Returning to Huber's comments about the "bewildering" collection of regression estimators, we find ourselves revisiting the same question of which types of estimators are most apropos for the present problem of high-dimensional linear regression. The authors' manuscript will surely inspire further practical and theoretical innovations reaching beyond the class of traditional penalized *M*-estimators.

Funding

The author gratefully acknowledges support from NSF grant DMS-1749857.

References

Alfons, A., Croux, C., and Gelper, S. (2013), "Sparse Least Trimmed Squares Regression for Analyzing High-Dimensional Large Data Sets," *The Annals of Applied Statistics*, 7, 226–248. [1715]

Forsythe, A. B. (1972), "Robust Estimation of Straight Line Regression Coefficients by Minimizing *p*th Power Deviations," *Technometrics*, 14, 159–166. [1715]

Huber, P. J., and Ronchetti, E. M. (2009), *Robust Statistics*, Wiley Series in Probability and Statistics, New York: Wiley. [1715]

Jaeckel, L. A. (1971), "Robust Estimates of Location: Symmetry and Asymmetric Contamination," *The Annals of Mathematical Statistics*, 42, 1020–1034. [1716]

——— (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals," *The Annals of Mathematical Statistics*, 43, 1449–1458. [1715]

Lugosi, G., and Mendelson, S. (2019), "Robust Multivariate Mean Estimation: The Optimality of Trimmed Mean," arXiv no. 1907.11391. [1716]

Minsker, S. (2018), "Uniform Bounds for Robust Mean Estimators," arXiv no. 1812.03523. [1716]

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2020), "Robust Estimation via Robust Gradient Estimation," *Journal of the Royal Statistical Society*, Series B, 82, 601–627. [1716]

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880. [1715]

Rousseeuw, P. J., and Yohai, V. (1984), "Robust Regression by Means of *S*-Estimators," in *Robust and Nonlinear Time Series Analysis*, eds. J. Franke, W. Härdle, and D. Martin, New York: Springer, pp. 256–272. [1716]

Stigler, S. M. (2010), "The Changing History of Robustness," *The American Statistician*, 64, 277–281. [1716]

Yohai, V. J. and Zamar, R. H. (1993), "A Minimax-Bias Property of the Least α-Quantile Estimates," *The Annals of Statistics*, 21, 1824–1842. [1715]