Comments on: Statistical inference and large-scale multiple testing for high-dimensional regression models

Gerda Claeskens[1] and **Maarten Jansen**[2]

[1] ORStat and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[2] Departments of Mathematics and Computer Science, Université libre de Bruxelles, Boulevard du Triomphe Campus Plaine, CP213, B-1050 Brussels, Belgium
gerda.claeskens@kuleuven.be, maarten.jansen@ulb.be

The paper by Cai, Guo and Xia provides an interesting overview of how to construct confidence intervals for parameters (or certain functions thereof) in high-dimensional linear or logistic regression models using debiased or desparsified estimators and on multiple testing in such models with control of the false discovery proportion and of its expected value, the false discovery rate. We congratulate these authors for their efforts in bringing together and making connections between various recent papers on this topic.

Regularized estimation methods that make use of an ℓ_1-type regularizer have been and still are popular, mainly due to their double purpose functioning: (i) the regularization (due to the ℓ_1-norm) performs the task of variable selection among the variables in the model and (ii) the procedure outputs an estimator of the 'selected' or 'active' variables. The procedure, however, comes with its own specific theoretical problems. First, the ℓ_1-regularization, although interesting from the computational point of view, causes the estimators to be shrunk and hence biased. Second, especially in the sparse high-dimensional case where many variables are not selected and hence found to be inactive, the distribution of the estimators is complicated. The complication mainly arises due to the pointmasses at zero for the variables that are found to be inactive, and the selection uncertainty about precisely which variables should be found inactive or active for the data at hand.

The debiasing or desparsifying methods that are surveyed in this paper come to rescue. The debiased estimators are shown to have a limiting normal distribution with estimable variance and a bias that can be neglected for inference. Once the limiting normality is obtained, the construction of confidence intervals and hypothesis testing becomes feasible.

There is, however, a price to pay. Due to the very construction of debiasing, one of the effects of an ℓ_1-regularization is undone. More precisely, the selection of active components is given up by the debiasing, resulting in again a full vector of all nonzero components. Unfortunately, in applications this refilling may hinder the interpretation of the estimators. For instance, in graphical modeling the debiased estimator requires working with fully connected graphs, while the initial purpose was precisely to obtain an interpretable sparse graphical model.

In some circumstances, one would solely be interested in performing inference about the coefficients that are found to be active by the regularized estimator. In that case the approach of selective inference becomes interesting. In selective inference, the distribution of the estimators is conditioned on the event that precisely the found selection took place. In contrast to debiasing, selective inference no longer considers the non-selected variables. For the selected components, valid conditional inference is obtained. For selective inference using lasso in linear models, see Lee et al. (2016). For ℓ_1-regularized logistic regression, see Taylor and Tibshirani (2018). While these authors provided a computationally feasible approach by an additional conditioning on the signs of the estimated nonzero coefficients, more powerful methods can be constructed for some models via parametric programming (Duy and Takeuchi, 2022) or by adding additional randomization (Panigrahi et al., 2023; Huang et al., 2023).

While the debiasing approach allows to also perform inference on coefficients that were not selected by the regularized estimator, it would be interesting to compare the methods for inference on a subset of the selected coefficients using the two approaches of debiasing and selective inference.

While the ordinary ℓ_1-regularizer is the most well-studied, it would be of interest to investigate whether the inference methods via debiasing would also apply to other regularization methods such as the smoothly clipped absolute deviation (Fan and Li, 2001), the minimax concave penaly (Zhang , 2010), ℓ_q-norms, the elastic net, the group lasso, etc.

Logistic regression is not easy to study. The proposed formula uses weighted raw residuals. One might wonder whether likelihood-based derivations using score functions would come to similar bias corrections.

As the described method of Cai et al. (2023) for inference in binary outcome models using debiased estimators requires sample splitting, one might question whether the construction of debiasing still pays off as compared to a straightforward sample splitting approach where part of the sample is used for the selection of the variables and the other sample part provides the information regarding the inference.

As with all regularized estimators, there is an additional choice to be made: the value of the regularization constant. While indeed, as the authors notice, cross-validation is a typical approach, its randomness again should be taken into account in further inference. This, however, might require substantial changes to the theoretical results.

References

Cai, T.T., Guo, Z., and Ma, R. (2023). Statistical inference for high-dimensional generalized linear models with binary outcomes. *Journal of the American Statistical Association*, 118(542):1319–1332.

Duy, V. N. L. and Takeuchi, I. (2022). More powerful conditional selective inference for generalized lasso by parametric programming. *Journal of Machine Learning Research*, 23(300):1–37.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96:1348–1360.

Huang, Y., Pirenne, S., Panigrahi, S., and Claeskens, G. (2023). Selective inference using randomized group lasso estimators for general models. arXiv 2306.13829.

Lee, J., Sun, D., Sun, Y., and Taylor, J. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Panigrahi, S., MacDonald, P. W., and Kessler, D. (2023). Approximate post-selective inference for regression with the group lasso. *Journal of Machine Learning Research*, 24(79):1–49.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.