Noname manuscript No.
(will be inserted by the editor)

Comments on: High-dimensional simultaneous inference with the bootstrap

Matthias Löffler · Richard Nickl

Received: date / Accepted: date

We would like to congratulate Ruben Dezeure, Peter Bühlmann and Cun-Hui Zhang for a stimulating and methodologically important contribution to the field of high-dimensional statistics. They propose a bootstrap methodology to infer the distribution of the statistic $\max_{j\in G}|\hat{\beta}_j-\beta_j|/\hat{\sigma}_j$ for any subset $G\subseteq\{1\dots,p\}$, which particularly allows the construction of simultaneous ℓ_∞-confidence 'bands' for the *whole* parameter β. Together with the suggested variance estimator this offers robustness against misspecification of the error distribution and takes possible heteroscedasticity into account. For the proofs multiplier bootstrap ideas from Chernozhukov et. al. (2013) are employed with technical virtuosity.

Here we discuss the potential shortfalls of this method for inference for relevant functionals of β. Our main caveat is that one should not be misled by the fact that one has a confidence region for the whole parameter and conclude that statistical inference for every aspect of β is feasible. Even for simple linear functionals the plug-in confidence set is possibly not useful. This is due to the fact that in the high-dimensional setting not all norms are equivalent.

Let us illustrate this in the prediction problem in random design: one observes
$(Y_1,X_1),\dots,(Y_n,X_n)$ where for simplicity X_i are *i.i.d.* Gaussian with unit variance, and is interested in predicting the linear functional $E[Y_{n+1}|X_{n+1}]=X_{n+1}^T\beta^0$ for some design vector X_{n+1}.
The plug-in confidence set for $X_{n+1}^T\beta$ is

$$C_n:=\left\{X_{n+1}^T\beta:\ \beta\in C_n^\beta\right\}, \tag{1}$$

Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, CB3 0WB Cambridge, United Kingdom

E-mail: m.loffler@statslab.cam.ac.uk; r.nickl@statslab.cam.ac.uk

where the construction of the confidence set C_n^β is given by Theorem 3 in Dezeure et. al. (2017). However, due to the use of the de-sparsified estimator in C_n^β the set C_n does not pick up the sparsity of β^0 and can easily be seen to have width of order at least $\sqrt{p\log(p)/n}$ with probability as close to one as desired. This is sub-optimal and diverges to ∞ in the high-dimensional setting $p \gg n$, rendering C_n practically useless compared to the the minimax rate of estimation of $X_{n+1}^T\beta^0$, which is $\sqrt{s_0\log(p)/n}$.

We conjecture here that it is impossible to have inference procedures that both have good ℓ_∞ and ℓ_2 behaviour simultaneously due to the need for de-sparsification in the ℓ_∞-case and the need for sparse estimators that adapt to s_0 in the ℓ_2-case.

Moreover, even when one uses a sparse estimator picking up the optimal rate $\sqrt{s_0\log(p)/n}$ is in general not achievable by confidence intervals for the prediction problem.
Using decision-theoretic principles laid out in chapter 8.3 in Giné and Nickl (2016) and carefully investigating the proofs in Nickl and van de Geer (2013) one can prove the following result: We define the space of $p^{1-\gamma}$-sparse vectors as

$$\Theta_\gamma := \left\{\beta : \beta \in \mathbb{R}^p, \sum_{i:\beta_i\neq 0} 1 \leq p^{1-\gamma}\right\}, \quad 0<\gamma<1,$$

and have:

Theorem 1 *Suppose that $p \geq n$ and that $p^{1-\gamma} = o(n/\log(p))$ for some $0<\gamma<1$. Furthermore, assume that the X_i are i.i.d. $\mathcal{N}(0,I)$ distributed, $i = 1,\cdots,n+1$. Finally, assume that for some $0<\alpha<1/3$, the confidence set C_n is honest for $X_{n+1}^T\beta$ over Θ_γ, satisfying*

$$\sup_{\beta\in\Theta_\gamma} P_\beta\left(X_{n+1}^T\beta \notin C_n\right) \leq \alpha.$$

Then, necessarily for any $1>\gamma_1>\gamma$,

$$\sup_{\beta\in\Theta_{\gamma_1}} E_\beta|C_n| \geq C\min(n^{-1/4}, \sqrt{p^{1-\gamma}\log(p)/n}). \tag{2}$$

Specifically this implies that it is impossible to achieve adaptation, i.e. obtaining the optimal rate $\sqrt{s_0\log(p)/n}$ for various values of β^0 and s_0, in the highly sparse region with $s_0 = o(\sqrt{n}/\log(p))$ assumed by Dezeure et. al. (2017). Of course, one might say that a diameter of $n^{-1/4} \vee \sqrt{s_0\log(p)/n}$ is acceptable and is much better than the one of (1) as it shrinks to 0. However, extending the proof of Theorem 1 using the results from Ingster et. al. (2010) one sees that even this is only attainable if one assumes that the variance of the errors ε_i is *known* and otherwise the rate $\sqrt{p^{1-\gamma}\log(p)/n}\wedge 1$ for an arbitrary $\gamma > 1/2$ is the best achievable rate for the prediction problem, even if the true sparsity is much smaller.

Other examples of functionals for which the above remarks apply include 'dense' functionals such as $\sum_{j=1}^{p} \beta_j^0$ (Cai and Guo, 2017) or the ℓ_2-loss $\|\hat{\beta} - \beta^0\|_2^2$ (Cai and Guo, 2016).

We want to point out here that for matrix inference problems with rank as unknown parameter the situation is more favorable. For example, in (Carpentier et. al., 2017) we consider the matrix completion problem and investigate the existence of honest and adaptive Frobenius confidence sets for the whole matrix. If 'repeated sampling' is possible we give an explicit and computable construction of such a set, even when the error variance is unknown, and thus show that in this model inference for functionals related to the Frobenius distance is possible.

To conclude, the existence of meaningful confidence statements in high-dimensional models depends highly on the statistical model, geometry of the underlying parameter space and particular aspect under consideration and requires careful case by case consideration. For practitioners the message is that extreme care has to be exercised when constructing confidence sets. Dezeure et. al. should be congratulated for singling out a set of well-posed high-dimensional inference problems.

References

Carpentier A, Klopp O, Löffler M and Nickl R (2017) Adaptive confidence sets for matrix completion. Bernoulli, to appear

Cai TT and Guo Z (2017) Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity. Ann Statist 45(2):615-646

Cai TT and Guo Z (2016) Accuracy Assessment for High-dimensional Linear Regression. Preprint at https://arxiv.org/abs/1603.03474

Chernozhukov V, Chetverikov D and Kato K (2013) Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors. Ann Statist 41(6):2786-2819

Dezeure R, Bühlmann P and Zhang CH (2017) High-dimensional simultaneous inference with the bootstrap. TEST, to appear

Giné E and Nickl R (2016) Mathematical Foundations of Infinite-Dimensional Statistical Methods. Cambridge University Press

Ingster YI, Tsybakov AB and Verzelen N (2010) Detection boundary in sparse regression. Electron J Statist 4:1476-1526

Nickl R and van de Geer S (2013) Confidence sets in sparse regression. Ann Statist 41(6):2852-2876