TEST (2017) 26:751–758
DOI 10.1007/s11749-017-0560-4

DISCUSSION

Rejoinder on: High-dimensional simultaneous inference with the bootstrap

Ruben Dezeure[1] · **Peter Bühlmann**[1] · **Cun-Hui Zhang**[2]

Published online: 9 October 2017

Abstract We thank the discussants for their interesting, inspiring and thoughtful comments and ideas. We provide here some responses.

Keywords De-biased Lasso · De-sparsified Lasso · Double robustness · Multiple testing · Postselection inference · Stability

Mathematics Subject Classification 62J07 · 62F40

1 Non-sparse regression parameters and "double robustness" (Bradic and Zhu 2017)

Bradic and Zhu present a fascinating new way for inference in high-dimensional linear models but with *non-sparse* regression parameters, based on their own earlier work (Zhu and Bradic 2016). They provide an insightful array of simulations, covering also

Ruben Dezeure is partially supported by the Swiss National Science Foundation SNF 2-77991-14.
Cun-Hui Zhang is partially supported by NSF Grants IIS-140793, DMS-15-13378 and DMS-17-21495.

This rejoinder refers to the comments available at doi:10.1007/s11749-017-0555-1; doi: 10.1007/s11749-017-0556-0; doi:10.1007/s11749-017-0557-z; doi: 10.1007/s11749-017-0558-y; doi:10.1007/s11749-017-0559-x.

✉ Peter Bühlmann
buhlmann@stat.math.ethz.ch

[1] Seminar for Statistics, ETH Zürich, HG G17, Rämistrasse 101, 8092 Zurich, Switzerland

[2] Department of Statistics and Biostatistics, Rutgers University, 569 Hill Center, Busch Campus, Piscataway, NJ 08854-8019, USA

non-sparse cases, for comparing various bootstrap schemes: this pushes the methodology into an interesting range of settings.

They propose a so-called CorrT test-statistics which, when scaled with $\sqrt{n}$, converges to a Gaussian limit, even for settings with non-sparse regression parameters. What is required instead is a corresponding sparsity of a sub-matrix of the design: if the focus is on inference for a single regression parameter say β_1^0, the assumption is that the regression of the first covariate X_1 against all others $X_{-1} = (X_2, \ldots, X_p)$ is sparse. Their test is related to the de-biased or de-sparsified estimator $\hat{b}_1$ via

$$\left|\frac{\hat{b}_1 - \beta_1^0}{\widehat{s.e.}_1}\right| = \left|\frac{Z_1^T(Y - X_1\beta_1^0 - \mathbf{X}_{-1}\hat{\beta}_{-1})}{(Z_1^T X_1)\widehat{s.e.}_1}\right|$$
$$= \frac{\left|(X_1 - \mathbf{X}_{-1}\hat{\gamma})^T(Y - X_1\beta_1^0 - \mathbf{X}_{-1}\hat{\beta}_{-1})\right|}{\hat{\sigma}_\varepsilon \|Z_1\|_2} \Rightarrow |\mathcal{N}(0,1)|$$

with $Z_1 = X_1 - \mathbf{X}_{-1}\hat{\gamma}$. It is known that this asymptotic normality of $\hat{b}_1$ does not require the sparsity of γ for suitable Z_1 (Javanmard and Montanari 2014). A kind of double robustness property is in play here to capture the symmetry roles between the two residual vectors in the numerator of the above expression, especially for the Gaussian design: if both the regression parameter and the design are sparse, the estimator is efficient as in Geer et al. (2014); and if only one of them is sparse and the other non-sparse, the estimator is still $\sqrt{n}$ consistent. This is very interesting in theory and useful in practice in the high-dimensional setting. Loosely related ideas exist in the literature on model-misspecification (Scharfstein et al. 1999; Bang and Robins 2005, cf.).

2 Small subsets of unpenalized variables (Lockhart and Samworth 2017)

Lockhart and Samworth suggest an interesting modification of the de-biased or de-sparsified Lasso which is computationally cheaper (as we will mention below), at the expense of more stringent conditions on the design. The idea is motivated by leaving some variables (in a group G) unpenalized or using "full" adjustment (for the target of interest β_G^0), as explained below: since the estimator will be non-sparse for the covariates in G, it should be regular and with a Gaussian limit (as made rigorous by Lockhart and Samworth). They consider for $\lambda > 0$,

$$\hat{\beta} = \text{argmin}_\beta \left(\|Y - X\beta\|_2^2/n + \lambda \sum_{j \in G^c} |\beta_j| \right), \tag{1}$$

where $G \subset \{1, \ldots, p\}$ is a "small subset" with $|G| < n$ (and the cardinality of the complement $|G^c|$ is large). The estimator is not unbiased for the true parameter β_G^0, but the bias occurs only through the components of the true parameter $\beta_{G^c}^0$: it is easy to derive (and in the analysis of Lockhart and Samworth) that

$$\hat{\beta}_G = \beta_G^0 + R_G X_{G^c}(\beta_{G^c}^0 - \hat{\beta}_{G^c}) + R_G \varepsilon, \quad R_G = (X_G^T X_G)^{-1} X_G^T.$$

Thus, for every component $j \in G$ we have that

$$\hat{\beta}_j = \beta_j^0 + \sum_{k \in G^c} (R_G)_{j,\bullet} X_k (\beta_k^0 - \hat{\beta}_k) + \text{ mean zero noise term.} \tag{2}$$

The sum only involves indices in G^c: for $|G| > 1$ (or $|G^c| < p - 1$); this is in contrast to the de-biased or de-sparsified Lasso which involves the bias term

$$\sum_{k \neq j} Q_{j,k}(\beta_k^0 - \hat{\beta}_k)$$

for some $Q_{j,k}$. Therefore, we say that the estimator in (1) for β_G^0 "fully" adjusts for all the variables in G: the word "fully" means that the bias term involves only a sum (as above) with indices in G^c.

As an alternative, a form like in (2) can be easily obtained by the de-biased or de-sparsified Lasso: for $j \in G$, we consider the regularized regression of X_j versus all other covariates X_{-j}, with unpenalized variables in G:

$$\hat{\gamma} = \text{argmin}_\gamma \left(\|X_j - X_{-j}\gamma\|_2^2/n + \lambda_X \sum_{k \in G^c} |\gamma_k| \right). \tag{3}$$

The corresponding residual vector is denoted by $Z_j = X_j - X_{-j}\hat{\gamma}$, and we then proceed with the de-biased estimator $\hat{b}$ as in our paper (Dezeure et al. 2017), but now with this modified residual vector which "fully" adjusts for the variables in G. This then leads to the form: for $j \in G$,

$$\hat{b}_j = \beta_j^0 + \sum_{k \in G^c} \frac{Z_j^T X_k}{Z_j^T X_j} (\beta_k^0 - \hat{\beta}_k) + \text{ mean zero noise term.}$$

Zhang and Zhang (2014) proposed this as the restricted low-dimensional projection estimator to fully de-bias for variables in G in the case where G is the index set of X_k with high $|X_j^T X_k/n|$. This is in analogy to (2). The advantage is that due to the KKT conditions of the penalized regression (of X_j versus X_{-j}) we have that $\max_{k \in G^c} |Z_j^T X_k/n| \leq \lambda_X/2$ which leads to the same bound for the estimated bias as in the original paper (Dezeure et al. 2017). In fact, the entire theory carries over when using the de-biasing step with the residual vector Z_j in (3) which "fully" adjusts for G. This view seems very much related to what Lockhart and Samworth suggest in their Sect. 3 with their estimator $\hat{\beta}_G$ in (1). Our strategy by simply replacing the residuals Z_j with the version in (2) requires no further proofs: everything is immediate from the KKT conditions saying that

$$Z_j^T X_k = 0 \quad \text{for } k \in G, \quad |Z_j^T X_k/n| \leq \lambda_X/2 \quad \text{for } k \in G^c.$$

From a computational viewpoint, the estimator $\hat{\beta}_G$ in (1) from Lockhart and Samworth is substantially cheaper than the de-sparsified Lasso. It requires once the Lasso for a problem with dimension $|G^c|$ and twice a least squares problem in dimension $|G|$; in contrast, the ("fully" adjusted) de-sparsified Lasso requires $|G|$ times a ("fully adjusted") Lasso in dimension $|G^c|$ and once a Lasso in dimension p. In terms of computational complexity, assuming that $|G| < n \ll p$ we have: for the components in G

$$\begin{aligned} &\text{estimator in (1) for components in } G: \quad O(n^2|G^c| + n|G|^2) = O(n^2 p), \\ &\text{"fully" adjusted de-sparsified Lasso with residuals in (2)}: \quad O(|G|n^2|G^c| + n^2 p) = O(|G|n^2 p). \end{aligned}$$

Although the orders of magnitude do not reflect more refined computation times, we already see that for "somewhat larger" groups, e.g., $|G| \asymp n$, there is a substantial gain with the estimator in (1). The price to be paid for the computational speed-up is the much more stringent assumption about $\|\Theta\|_\infty$ in Corollary 2 in Lockhart and Samworth: we conjecture that their condition is "not too far" from being necessary and thus perhaps implying that the estimator is not very reliable in some scenarios.

To cope with the latter problem, Lockhart and Samworth propose to de-bias (or de-sparsify) the estimator $\hat{\beta}_{G^c}$ in (1) which is then used for constructing an estimator $\hat{b}_G =: \hat{b}_{\mathrm{LoSa};G}$ for β_G^0 (where the notation $\hat{b}_{\mathrm{LoSa};G}$ is used for the proposed estimator by Lockhart and Samworth). It would be interesting to see the (empirical) properties of $\hat{b}_{\mathrm{LoSa};G}$ in comparison to the standard de-biased (or de-sparsified) estimator $\hat{b}_G$ as considered in the original paper (Dezeure et al. 2017).

3 Higher-order asymptotic results (Chatterjee 2017)

Chatterjee points out that empirical results suggest better theoretical properties of the bootstrap than the standard (non-bootstrapped) de-biased estimator. We do not have an answer to this interesting remark. Of particular interest would be a *theoretical* result saying that bootstrapping the entire estimator including the bias correction term, as we propose in the original paper (Dezeure et al. 2017) leads to better performance than (bootstrapping) the linearized estimator studied by Zhang and Cheng (2016). Such a result would be interesting as it would not involve some studentization, BCa correction or double bootstrapping which is typically required for higher-order accuracy in the classical low-dimensional setting (Hall 1988, cf.).

4 Different norms, adaptivity and confidence sets for prediction (Löffler and Nickl 2017)

Löffler and Nickl emphasize the important point that different norms are not equivalent in high-dimensional spaces. Indeed, our work is focusing on the ℓ_∞-norm and corresponding max-type statistics. We should add here that all our results (Dezeure et al. 2017) are uniform when the $o(1)$ is uniform in conditions (A1)–(A6). Moreover, the $o(1)$ in (A1) and (A4) are uniform over ℓ_0-sparse parameters in the set

$$\Theta(s_0) = \{\beta;\ \|\beta\|_0^0 \le s_0\}$$

under conditions (B1) and (B2). However, the ℓ_∞-norm results exclude confidence sets for prediction or sum-type or "dense" functionals, as pointed out by Löffler and Nickl.

The case for inference of $\beta_G^0 = \{\beta_j^0;\ j \in G\}$ for large groups $G \subseteq \{1, \ldots, p\}$ is particularly interesting in practice. When the group size is modest with $|G| = o(n)$, ℓ_2-norm results with optimal rates have been given by using a version of the de-biased or de-sparsified Group Lasso (Mitra and Zhang 2016; van de Geer and Stucky 2016). For larger groups with $|G| \gg n$, such ℓ_2-norm results are—presumably, as also noted by Löffler and Nickl—impossible to achieve (even when dropping the request for "honesty" with uniform convergence). However, the following is worth pointing out: in practical applications (e.g., genomics, genetics) one is often interested in the easier problem of testing the statistical null hypothesis $H_{0,G} : \beta_G^0 \equiv 0$ versus the alternative $H_{A,G} : \beta_G^0 \not\equiv 0$. The max-type test statistics from the paper (Dezeure et al. 2017) should exhibit good power when the alternative is sparse with only a few nonzero coefficients. When the alternative is ("modestly") dense, one would think that a sum-type statistics should perform better in terms of power. This line of thinking might be misleading though when the covariates exhibit fairly high correlation (or small subsets of covariates are nearly linearly dependent) as is often the case in many high-dimensional real datasets. Then, even when β^0 is dense but with highly correlated variables, a max-type statistics might work quite well for detecting the alternative. Thus, the notion of a "sparse" or "dense" alternative by considering the structure of β^0 only is too short-sighted: the "correlation structure" among the covariates is relevant as well. We are not aware of a result which addresses this issue.

Löffler and Nickl point out fascinating facts for the ambitious task of constructing honest confidence sets in high-dimensional settings, namely: the difficulty for problems like prediction, the distinction whether the error variance is known or not, and their positive and encouraging results on matrix inference problems. These are important benchmarks about limitations and possibilities for high-dimensional inference.

5 Simplicity, ranking, and mean squared error (Liu and Yu 2017)

Liu and Yu raise various points, often in connection with issues about practical value and simplicity.

We agree that their proposals for bootstrapping the Lasso or LassoOLS (Liu and Yu 2013) are simpler and computationally much more efficient than the de-biased Lasso, with the disadvantage that coverage is not guaranteed for small and large nonzero coefficients, see their Figures 3 and S1 in the electronic supplementary material (Liu and Yu 2017). The empirical study in (Dezeure et al. 2015) confirms substantial disadvantage in coverage probability for directly bootstrapping the Lasso or LassoOLS; and the latter study also shows that other computationally efficient methods such as multi-sample splitting (Meinshausen et al. 2009) or bias-corrected Ridge estimation (Bühlmann 2013) perform comparatively well and robustly in terms of coverage, at the price of being less statistically efficient. We note that the LassoOLS does correct the bias within the selected model, but unlike de-biasing methods studied in

the paper (Dezeure et al. 2017), it does not remove the biased due to model selection error. We emphasize also that the inaccuracy for coverage of the bootstrapped Lasso or LassoOLS is clearly present for large non-zero coefficients, and thus, such a super-efficiency is in our view a relevant issue: if it were only for the small non-zero coefficients, we would agree with Liu and Yu that this wouldn't be a major obstacle in practice.

Liu and Yu also mention other metrics than p values. It is established statistical practice that confidence intervals are more informative than p values since they explicitly report on effect sizes (with corresponding uncertainties). In practice though one does not want to display, e.g., $O(10^4)$ confidence intervals, see also (Benjamini and Yekutieli 2005). Ranking, as proposed by Liu and Yu, is certainly a very important and informative metrics; the MSE is useful as well but one cannot easily construct the numbers from a given data set (i.e., we would need to estimate the MSE which would be cumbersome).

Regarding the MSE, the de-biased Lasso should be viewed as raw estimators with

$$\hat{b}_j \approx N(\beta_j, \sigma_j^2/n) \ (j = 1, \ldots, p),$$

that is, a correlated approximate Gaussian sequence model. This enables many possible downstream options for further use, and for loss functions like the MSE the Lasso should be compared with the *thresholded* version of the de-biased Lasso. The advantage of the thresholded de-biased Lasso is that all large coefficients, meaning $> C_0\lambda \asymp \sqrt{\log(p)/n}$, are retained while the zeroes are correctly removed, see Theorem 3 in (Zhang and Zhang 2014). On the other hand, the (thresholded) Lasso with or without OLS is not guaranteed to retain all coefficients larger than $C_0\lambda$. In fact, in the worst case scenario, the Lasso will presumably zero out a large coefficient of the order $\sqrt{s_0}\lambda$. The MSE properties of a thresholded de-biased Lasso should be much improved in comparison with the non-thresholded version shown in the numerical experiments from Liu and Yu.

6 Simultaneous versus post-selection inference (Chatterjee 2017)

Chatterjee raises an interesting issue about post-selection inference, thereby pointing to the "practitioners point of view".

In our own view, the simultaneous inference is the "cleanest" approach, although perhaps sometimes too conservative (but less conservative than guarding against all sub-models as in Berk et al. (2013), see below). When having adopted a model, such as a linear or a logistic regression model with all the covariates, simultaneous inference over all the (e.g., regression) parameters is clean and has a solid confirmatory interpretation. Building upon confidence intervals for individual parameters in the traditional sense, such simultaneous inference is also relatively easy to explain to practitioners.

Post-selection inference techniques are interesting since they address to a certain extent the issues when practitioners have chosen certain sub-models or scientific hypotheses based on data, and the same data is used again for inference. A major problem seems to us, especially in the high-dimensional context, that a chosen ran-

dom sub-model or hypothesis might look very different when the data would have been different (e.g., other realizations from the same true underlying data generating probability distribution). A prime example is the Lasso in a linear model: in the high-dimensional scenario, a data-chosen sub-model would look substantially different for different data realizations (and the entire regularization path could look very different). This is particularly true in presence of moderately large number of weak signals paired with significantly correlated covariates. When the Lasso-estimated sub-models will vary substantially, such post-selection inference would produce results which are not replicable when having other new data from the same underlying (probability) mechanism. Of course, more general post-selection inference techniques can address this instability issue of an estimated sub-model: Berk et al. (2013) protect against all possible sub-models by paying a price to be very conservative—and in fact more conservative than the simultaneous inference in the full model.

Feasibility of statistical inference for individual regression coefficients is questionable in presence of highly correlated covariates and corresponding selected models based on individual variable inclusion are instable. This selection instability in turn is unpleasant in the scientific interpretation of post-selection inference. For example, if the jth covariate is measuring the expression of a gene j, assume that the regression coefficient β_1^0 is "fairly large" and $\beta_2^0 = 0$, and that X_1 and X_2 are highly correlated. Then, a Lasso-estimated sub-model would typically either include X_1 or X_2; and if the sub-model is selected where X_1 is missing, we would miss the true relevant gene 1 in the stage of post-selection inference. This would not happen, when considering groups of variables simultaneously: there, we typically would find that the group of covariates $\{X_1, X_2\}$ is significant. For simultaneous inference, Mitra and Zhang (2016) and van de Geer and Stucky (2016) consider group inference and Mandozzi and Bühlmann (2016a, b) advocate a hierarchical group inference procedure. For post-selection inference, Lee et al. (2016) use the instable Lasso as model selector: post-selection inference after a stable model selection based on group of variables inclusion seems to be an open topic.

References

Bang H, Robins J (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61:962–972

Benjamini Y, Yekutieli D (2005) False discovery rate-adjusted multiple confidence intervals for selected parameters. J Am Stat Assoc 100:71–81

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid postselection inference. Ann Stat 41:802–837

Bradic J, Zhu Y (2017) Comments on: high-dimensional simultaneous inference with the bootstrap. TEST. doi:10.1007/s11749-017-0556-0

Bühlmann P (2013) Statistical significance in high-dimensional linear models. Bernoulli 19:1212–1242

Chatterjee A (2017) Comments on: high-dimensional simultaneous inference with the bootstrap. TEST. doi:10.1007/s11749-017-0557-z

Dezeure R, Bühhlmann P, Zhang CH (2017) High-dimensional simultaneous inference with the bootstrap (with discussion). TEST. doi:10.1007/s11749-017-0554-2

Dezeure R, Bühlmann P, Meier L, Meinshausen N (2015) High-dimensional inference: confidence intervals, p values and R-software hdi. Stat Sci 30:533–558

Hall P (1988) Theoretical comparison of bootstrap confidence intervals. Ann Stat 16:927–953

Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15:2869–2909

Lee JD, Sun DL, Sun Y, Taylor JE (2016) Exact post-selection inference, with application to the lasso. Ann Stat 44:907–927

Liu H, Yu B (2013) Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. Electron J Stat 7:3124–3169

Liu H, Yu B (2017) Comments on: high-dimensional simultaneous inference with the bootstrap. TEST. doi:10.1007/s11749-017-0559-x

Lockhart R, Samworth R (2017) Comments on: high-dimensional simultaneous inference with the bootstrap. TEST. doi:10.1007/s11749-017-0555-1

Löffler M, Nickl R (2017) Comments on: high-dimensional simultaneous inference with the bootstrap. TEST. doi:10.1007/s11749-017-0558-y

Mandozzi J, Bühlmann P (2016a) Hierarchical testing in the high-dimensional setting with correlated variables. J Am Stat Assoc 111:331–343

Mandozzi J, Bühlmann P (2016b) A sequential rejection testing method for high-dimensional regression with correlated variables. Int J Biostat 12:79–95

Meinshausen N, Meier L, Bühlmann P (2009) *P* values for high-dimensional regression. J Am Stat Assoc 104:1671–1681

Mitra R, Zhang C-H (2016) The benefit of group sparsity in group inference with de-biased scaled group Lasso. Electron J Stat 10:1829–1873

Scharfstein D, Rotnitzky A, Robins J (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). J Am Stat Assoc 94:1096–1146

van de Geer S, Bühlmann P, Ritov Y, Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Stat 42:1166–1202

van de Geer S, Stucky B (2016) χ^2-confidence sets in high-dimensional regression. In: Frigessi A, Bühlmann P, Glad IK, Langaas M, Richardson S, Vannucci M (eds) Statistical analysis for high-dimensional data, the abel symposium 2014. Springer, New York, pp 279–306

Zhang C-H, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J Roy Stat Soc B 76:217–242

Zhang X, Cheng G (2016) Simultaneous inference for high-dimensional linear models. J Am Stat Assoc. doi:10.1080/01621459.2016.1166114

Zhu Y, Bradic J (2016) Hypothesis testing in non-sparse high-dimensional linear models. arXiv:1610.02122