Computational Statistics
https://doi.org/10.1007/s00180-019-00948-1

REJOINDER

Rejoinder on: Hierarchical inference for genome-wide association studies: a view on methodology with software

Claude Renaux[1] · Laura Buzdugan[1] · Markus Kalisch[1] · Peter Bühlmann[1]

1 Preamble

We thank the discussants, in alphabetical order, Goeman and Böhringer (GB), Frommlet and Schimek (FS), Heller (HE) and Rodrigues and Lourenco (RL) for their insightful, interesting and very valuable contributions on our paper.

2 The analysis pipeline in practice

One of our main goals of the paper is to provide an "analysis pipeline", as coined by HE. It includes software and explanation: the latter is given by some supporting mathematical theory as well as empirical evidence and experience, and they are all "incomplete". Incompleteness of empirical evidence is due to the limited amount of analyzed real and simulated data; on the other hand, the mathematical theory exhibits incompleteness due to unrealistic assumptions and a lack of understanding of the actual procedure. We will elaborate on the second issue in Sect. 3.

Despite the mentioned incompleteness above, we want to communicate and transfer the analysis pipeline to practice, with the hope that it will serve as a step towards more reliable and replicable methodology for assigning "relevance" of features as, for example, in the analysis of data from GWAS. In fact, we should re-mention here that our paper is building and reviewing on previous contributions starting with Meinshausen et al. (2009) on multiple sample splitting and inference, Meinshausen (2008), Mandozzi and Bühlmann (2016a, b) on high-dimensional hierarchical inference and Buzdugan et al. (2016), Klasen et al. (2016) on real GWAS applications: the novelty of the paper is primarily the review of earlier work, the development of modular software

This rejoinder refers to the comments available at: https://doi.org/10.1007/s00180-019-00945-4, https://doi.org/10.1007/s00180-019-00941-8, https://doi.org/10.1007/s00180-019-00942-7, https://doi.org/10.1007/s00180-019-00943-6.

✉ Peter Bühlmann
peter.buehlmann@stat.math.ethz.ch

[1] Seminar for Statistics, ETH Zürich, Zürich, Switzerland

Published online: 08 January 2020

with the R-package `hierinf` for providing the "analysis pipeline" and an extension for meta analysis.

Parts of the methodology should be extended and perhaps adapted in the future. We believe though that the principle of hierarchical inference is very natural and powerful and consider it as appropriate for a range of inference problems. This seems to be in accordance with the comments from all the discussants. They also make various suggestions how the analysis pipeline could be modified and further developed and we will discuss them below.

2.1 From marginal to multivariate regression: gaining potential stability

`FS` ask why modeling of the population structure (which is rather common in GWAS) is less an issue with a multivariate model. The simplest hypothetical case is the following: the response Y is a (linear) function of the covariates X ("Y is caused by X in a structural equation model"). Assuming that the population structure results in say shifts of the X variables or changing the proportions of discrete X variables (but not on Y; this means, the population structure is not a confounder), then the regression coefficients stay the same while the correlations can change dramatically. We quote here Tukey, explaining from a different point of view why regression in general should be much better than marginal correlation: "One of the major arguments for regression instead of correlation is potential stability. We are very sure that the correlation cannot remain the same over a wide range of situations, but it is possible that the regression coefficient might. We are seeking stability of our coefficients so that we can hope to give them theoretical significance" (Tukey 1954). Nevertheless, additional deconfounding within a multivariate model may be beneficial: using mixed models as mentioned in Sect. 5 or performing deconfounding more directly as discussed in Sect. 6.

3 The main assumptions for theory and the actual methodology

We admit that the presented theory in the paper is far from satisfactory. Yet we think that this does not rule out the usefulness of the methodology.

The theory essentially requires the screening property of the Lasso or any other high-dimensional variable selector; for example, the adaptive or thresholded Lasso have slightly better theoretical properties for variable screening than the Lasso (van de Geer et al. 2011) at the price of a two-stage and hence more complicated screening procedure.

To establish the screening property of the Lasso, the sparsity assumption (A1) in the paper seems somehow "unavoidable" for theory. As we explain in the paper in Section 2.1.1, the condition on sparsity can be relaxed to weak sparsity in terms of ℓ_q-"norm" for $0 < q < 1$ (van de Geer 2016). In addition, it could still be interesting to consider a best sparse approximation of the truth: the justification of our inferential procedure hinges then on the behavior of the Lasso in such a sparse approximation setting, see Bühlmann and van de Geer (2011).

As GB write, a main problem is that the screening property of the Lasso would typically not hold. It is known that with Lasso screening on a first split half-sample and single variable inference afterwards on the other half-sample, the resulting p values are highly variable across different sample splits, resulting in a so-called p value lottery (Meinshausen et al. 2009). It is probably the instability of the Lasso as a variable screening method which leads to such highly variable (and non-replicable) results. The issue is perhaps a bit less pronounced for p values of groups of highly correlated variables since our procedure would work fairly well if randomly any of the highly correlated variables are selected. GB propose a goodness-of-fit test regarding the screening property of the Lasso: we have not tried it out but we think that it would easily indicate violation of the screening property of the Lasso, see also the "p value lottery" mentioned above.

3.1 Two remedial views on the problematic screening property of the Lasso

Aggregation and Stability The paper takes up the old idea from Meinshausen et al. (2009) to aggregate the p values from multiple sample splits. Formula (8) in the paper, derived in Meinshausen et al. (2009), describes in fact a general result for aggregating multiple and arbitrarily dependent p values. Romano and DiCiccio (2019) have recently studied this problem more generally. With an aggregation over multiple sample splits, the results are much less depending on the realizations of sample splitting than using a single split. Unfortunately, we have no theory in absence of an approximate screening property of the Lasso, although in principle it seems that the screening property is less important when using aggregation of p values as discussed next.

The aggregation bears a relation to Stability Selection (Meinshausen and Bühlmann 2010). For example, when using median aggregation of the p values with fixed $\gamma = 0.5$ in formula (18), a significant group (at any level $\alpha < 1$) requires that at least one variable in the group G of interest must have been selected (by the Lasso screening) at least 50% times across the B sample splits. This is due to the fact that we assign a p value equal to 1 if no selected variable appears in the group G, i.e., if $\hat{S}^{[b]} \cap G = \emptyset$. Since aggregation requires some stability for significance, this view contrasts what GB wrote, namely that "We are not sure that this [aggregation] is the best way to deal with such variability, since such variability could also be a strong indicator that the screening property fails. If that would be the case, the results would have excess Type I errors for every data split, and consequently also for the aggregated result." It is perhaps instructive to indicate a bit more the role of stability. Stability Selection counts

$$f(G) = B^{-1} \sum_{b=1}^{B} I(\hat{S}^{[b]} \cap G \neq \emptyset).$$

One then establishes a threshold τ such that all G with $f(G) \geq \tau$ are inferred to be "relevant", and it controls the expected number of false positives as a measure of "type I error". The validity for such an error control with Stability Selection requires a

strong condition about exchangeability. With regard to the inference procedure in the paper, we control the more classical FWER but also requiring a strong assumption, namely the screening condition. Empirically it seems that both Stability Selection and our hierarchical inference procedure work "well" in regimes much beyond the assumptions from theory.

Post-selection inference with respect to multiple selectors Another remedy relies on an interpretation from post-selection inference as described in Sect. 7. While the inference procedure remains the same, the post-selection interpretation does not require the screening property of the Lasso.

3.2 Different methods based on old and new de-biasing of the Lasso

In the paper in Section 2.2.2, we mention a few other methods: de-biasing or de-sparsifying the Lasso do not require a screening property nor the beta-min condition (A3). The main bottleneck with the de-biased Lasso is the computational cost $\mathcal{O}\big(n \min(n, p) p^2\big)$ which is quadratic in p for $p \gg n$ whereas the hierarchical inference approach of the paper requires $\mathcal{O}\big(Bn^2 p\big)$ essential operations (Section 2.2.1 in the paper). For GWAS with very large p, de-biasing the Lasso is not really feasible.

In Guo et al. (2019), we develop a de-biasing technique for group testing which is, especially for large groups, computationally more efficient and statistically more powerful than the de-biased Lasso in a range of scenarios. This novel group testing procedure can then be plugged into the hierarchical testing framework. From a theory perspective, it does not require a screening property nor the beta-min assumption (A3). We plan to have this alternative group testing method as an option in our R-package `hierinf`.

4 Exploratory versus confirmatory inference and stability

FS raise the point that in the context of GWAS, exploratory data analysis and inference has an important role. This is certainly an important remark. As we indicated above, we think that the hierarchical inference procedure in the paper is still working "well" beyond the regime where the assumptions for theory hold and where the method would be confirmatory.

In particular, the sample splitting aspect of our procedure exploits whether some findings are sufficiently stable, see Sect. 3.1 above. Stability itself is an important requirement in inference and data analysis: this has been emphasized from different viewpoints in Meinshausen and Bühlmann (2010), Shah and Samworth (2013), Yu (2013), Kumbier and Yu (2019). We also discuss stability from the viewpoint of replicability in Sect. 6.

Significance and ranking It is perhaps a bit delicate to let the user decide whether a procedure is confirmatory or not. We all know that "$P \leq 0.05$" is a too simplified "one-dimensional" view, especially with complex and large-scale data, as FS have mentioned as well. However, ranking variable or group importance in terms of p

values might be better than ranking based on magnitude of point estimators: because the former takes some uncertainty into account. (And it might be even better to do some sort of ranking with confidence intervals than p values).

5 Adaptations of the procedure

All discussants raised various interesting points: some adaptations of our procedure would address some of them. FS ask whether mixed models could be used: this would be possible, for example using a high-dimensional ℓ_1-penalized mixed model fit for screening of the relevant fixed effects (Schelldorfer et al. 2011, 2014) followed by low-dimensional mixed model inference.

GB mention that it would be valuable to include biological insights: this is directly possible as long as it is encoded in terms of hierarchical groups. One can define a R `dendrogram`, manually or using some other function, and feed it into our software; region-wise grouping is already built into `hierinf` as discussed in the paper (end of Section 2.3 and Section 3.1). Instead of a hierarchy, other structures such as directed acyclic graphs are possible (Goeman and Mansmann 2008) but this would require a substantial extension; see also the nice overview by Goeman and Solari (2014) which covers structured inference as well.

HE points out that controlling the false discovery rate (FDR) is important. We agree but do not know how to do this in a "data adaptive" way: our hierarchical procedure proceeds in a top-down manner through the hierarchy without pre-determining the final depth of the tree. Controlling the FDR at outer nodes is certainly an important step in that direction but requires to pre-define the outer nodes (the depth of the tree).

RL and also FS bring up the issue of choosing the aggregation scheme for meta analysis. We were also a bit surprised that Tippett's method worked better than Stouffer's, even in situations where we would expect an advantage for the latter. We gave a reason for this in the paper (end of Section 4.2), namely the instability by design of the procedure (which is still present to a certain extent despite multiple sample splitting and aggregation). Other aggregation techniques are possible, of course: with Tippett and Stouffer, we tried out two of them which are at "opposite poles" of each other.

6 Replicability and transferability

HE brings up the issue of replicability, whether significant findings occur in two or more datasets. As she points out, nothing in addition is needed than the p values for the multiple datasets. We plan to implement this nice suggestion into the software `hierinf` and our analysis pipeline. We also want to point out the work by Heller and Yekutieli (2014) which uses an elegant empirical Bayes approach for replicability analysis.

One of the plaguing issues in practice is the transferability to new sub-populations or different "domains". The classical quantification of replicability is concerned for the same data generating distribution. But if the data mechanism (slightly) changes from one to another sub-population, the situation is different. Transfer learning, primarily

developed in machine learning, deals with the changing mechanism situation, mainly for prediction (Pan and Yang 2010).

Population structure, unobserved confounding and beyond `HE` mentions the issue of unobserved confounding. It is certainly one of the major problems for transferability to new sub-populations. Recent proposals to do "deconfounding" include Wang and Blei (2018), Ćevid et al. (2018): due to the difficulty of the problem, the proposals rely on some "untestable" assumptions. Spectral transformation from Ćevid et al. (2018) requires only to linearly transform the data first which can then be used as input to our procedure. We should also mention that one could easily use the classical technique (which may be somewhat inferior to spectral transformations): namely directly adjusting for the first few principal components of the design matrix by using them as control variables in our software `hierinf`.

When having observed more than one sub-population, an instrumental variables approach can be used for deconfounding (Angrist et al. 1996, cf.) where the environments encode discrete instruments. Instrumental variables regression relies on untestable and stringent assumptions. These severe conditions can be relaxed: stability in so-called anchor regression leads to replicability (in linear models with confounding) even when the data generating distributions change according to some unknown shift perturbations (Rothenhäusler et al. 2018).

7 Post-selection inference with respect to multiple selectors

Post-selection inference has been further developed in the past years by Lockhart et al. (2014), Tibshirani et al. (2016), Fithian et al. (2014). It is an appealing concept and is (in part) motivated by what "practitioners intend to do", namely select a (sub-)model and then do inference in the selected (random) model using the same data twice.

More precisely, denote by $M \subseteq \{1, \ldots, p\}$ a subset of the covariates. For the derivations in this section, we consider a random design linear model with $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^T X] = \Sigma$: it makes things easier here but implies that the t-test in low-dimensional sub-models is not exact. We denote by $\beta(M) = \Sigma_M^{-1}\mathbb{E}[X_M^T Y]$ with X_M the components of X corresponding to M and $\Sigma_M = \mathbb{E}[X_M^T X_M]$.

When having a single selector $\hat{S} \subseteq \{1, \ldots, p\}$, post-selection inference leads to p values and confidence intervals for the parameter $\beta(M)$ for $M = \hat{S}$: that is, it is conditional inference given that $\hat{S} = M$. This is the "classical" approach of selective inference, namely for one selector. When the selector is given by the Lasso or forward stepwise selection, Tibshirani et al. (2016) have elegant ways to do exact post-selection inference. Also the approach mentioned by `HE` falls into this category. On the other hand, Berk et al. (2013) do simultaneous inference for all sub-models $M \subseteq \{1, \ldots, p\}$ and this immediately implies conservative inference when any selector has been used. In terms of significance tests for groups of variables, the null-hypotheses are of the following form. For a single sub-model M,

$$H_{0,G}(M) : \ \beta_G(M) = 0.$$

It means that $\beta_j(M) = 0$ for all $j \in G \cap M$; if $G \cap M = \emptyset$, then $H_0(M)$ is considered to be always true. When considering all sub-models M simultaneously as in Berk et al. (2013), the null-hypotheses are

$$H_{0,G}(\text{all}) : \ \beta_G(M) = 0 \ \text{ for all subsets } M \subseteq \{1, \ldots, p\}.$$

Our hierarchical procedure is considering the null-hypotheses:

$$H_{0,G}(M^{[1]}, \ldots, M^{[B]}) : \ \beta_G(M^{[b]}) = 0 \ \text{ for all } b = 1, \ldots, B, \tag{1}$$

where the Lasso selector (on the first half of the split data) leads to $\hat{S}^{[b]} = M^{[b]}$, over multiple sample splitting $b = 1, \ldots, B$. It is easy to see that our hierarchical procedure leads to valid p values for the hypotheses as in (1) when modifying the estimation of the error variance; and this is true *without* requiring the screening property of the Lasso and *without* requiring the sparsity or compatibility conditions (A1) and (A2), nor do we need (A3). The price to be paid is the more difficult interpretation of the hypotheses in (1).

7.1 Instability of the selector and interpretation of the null-hypothesis in (1)

When using one selector $\hat{S}$ only, with the Lasso as a prime example, there is a substantial problem about replicability. On a new dataset, it is very likely that the selector picks a very different model: the degree of replicability is then very low because inference is done in two very different models. In our view, this is a substantial issue when doing post-selection inference with e.g. the Lasso or forward stepwise selection.

On the other end of the spectrum is the approach by Berk et al. (2013) which performs simultaneous inference over all sub-models: there is no specific selector involved and hence there is no instability of a selector. The degree of replicability should be much improved at the price of using a very conservative method.

Multi sample splitting as a way to create different selected models An important issue is how to protect against the instability of one selector. One generic way is via re- or sub-sampling: this is perhaps the most natural approach in frequentist statistics. This viewpoint implies the following: multiple sample splitting involves some efficiency loss due to lower sample size for the inference after selection but it gains in terms of replicability and instability with respect to the selector. This is a non-standard trade-off which is not really understood: earlier work from Breiman (1996a, b) convincingly indicates that "gain in stability is worthwhile, perhaps at the price of lower sample size". The "lower sample size" is not even obvious due to aggregation over many sub-sampled estimates. We thus conclude that multiple sample splitting is a "natural way" for post-selection over many selectors.

Interpretation of the inferential statements The remaining problem though is the interpretation of the null-hypothesis in (1). It seems useful to inspect, in an exploratory way, which variables and group of variables occurred in the selected

models $M^{[1]}, \ldots, M^{[B]}$: we could count the relative frequencies of single variables in all these models, of groups of two variables, groups of three variables and so on. This gives us an idea of the conditioning sets in the different selected regression models.

Beyond such exploratory reasoning, the following toy example indicates what we have in mind when relying on a so-called faithfulness assumption (see below). Consider $p = 10$ and the true active variables are $S_0 = \{1, 2, 3, 7, 8\}$. The variable $\{1\}$ has been found to be significant by our procedure. We also observe that the single variables $\{3\}$ and $\{8\}$ occurred in all the models $M^{[1]}, \ldots, M^{[B]}$. Under a faithfulness assumption described below, which is very common in directed graphical modeling, cf. Spirtes et al. (2000), we then conclude that variable $\{1\}$ is significant given the variables $\{3, 8\}$, and also that variable $\{1\}$ is significant given either $\{3\}$ or $\{8\}$ alone, and that variable $\{1\}$ is marginally significant given no other variable. Similar conclusions hold when a group, say, $\{2, 7\}$ has occurred in all the selected models.

The faithfulness assumption requires in our context that for every sub-model M,

$$\text{if } \beta_j(M) = 0 \quad \Longrightarrow \quad \beta_j(M') = 0 \text{ for all } M' \supseteq M.$$

For example, if the marginal effect $\beta_j(\emptyset) = 0$, then it must be zero as well when conditioning on additional variables.

In future research we plan to exploit how such a post selection interpretation over many selectors works and whether it leads to interesting insights.

8 Concluding words

We wholeheartedly thank all the discussants for sharing their insights and ideas. This certainly has led to clarifications and will lead to improvements of our analysis pipeline.

References

Angrist J, Imbens G, Rubin D (1996) Identification of causal effects using instrumental variables. J Am Stat Assoc 91:444–455

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

Breiman L (1996a) Bagging predictors. Mach Learn 24:123–140

Breiman L (1996b) Heuristics of instability and stabilization in model selection. Ann Stat 24:2350–2383

Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin

Buzdugan L, Kalisch M, Navarro A, Schunk D, Fehr E, Bühlmann P (2016) Assessing statistical significance in multivariable genome wide association analysis. Bioinformatics 32:1990–2000

Ćevid D, Bühlmann P, Meinshausen N (2018) Spectral deconfounding and perturbed sparse linear models. Preprint arXiv:1811.05352

Fithian W, Sun D, Taylor J (2014) Optimal inference after model selection. Preprint arXiv:1410.2597

Goeman JJ, Mansmann U (2008) Multiple testing on the directed acyclic graph of gene ontology. Bioinformatics 24:537–544

Goeman JJ, Solari A (2014) Multiple hypothesis testing in genomics. Stat Med 33:1946–1978

Guo Z, Renaux C, Bühlmann P, Cai T (2019) Group inference in high dimensions with applications to hierarchical testing. Preprint arXiv:1909.01503

Heller R, Yekutieli D (2014) Replicability analysis for genome-wide association studies. Ann Appl Stat 8:481–498

Klasen J, Barbez E, Meier L, Meinshausen N, Bühlmann P, Koornneef M, Busch W, Schneeberger K (2016) A multi-marker association method for genome-wide association studies without the need for population structure correction. Nat Commun. https://doi.org/10.1038/ncomms13299

Kumbier K, Yu B (2019) Veridical data science. Preprint arXiv:1901.08152

Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R (2014) A significance test for the lasso. Ann Stat 42:413–468

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 (2008) Hierarchical testing of variable importance. Biometrika 95:265–278

Meinshausen N, Bühlmann P (2010) Stability Selection (with discussion). J R Stat Soc Ser B 72:417–473

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

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359

Romano J, DiCiccio C (2019) Multiple data splitting for testing. Technical report no. 2019-03, Department of Statistics, Stanford University

Rothenhäusler D, Meinshausen N, Bühlmann P, Peters J (2018) Anchor regression: heterogeneous data meets causality. Preprint arXiv:1801.06229

Schelldorfer J, Bühlmann P, van de Geer S (2011) Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. Scand J Stat 38:197–214

Schelldorfer J, Meier L, Bühlmann P (2014) GLMMLasso: an algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. J Comput Graph Stat 23(2):460–477

Shah R, Samworth R (2013) Variable selection with error control: another look at Stability Selection. J R Stat Soc Ser B 75:55–80

Spirtes P, Glymour C, Scheines R (2000) Causation, prediction, and search, 2nd edn. MIT Press, Cambridge

Tibshirani R, Taylor J, Lockhart R, Tibshirani R (2016) Exact post-selection inference for sequential regression procedures. J Am Stat Assoc 514:600–620

Tukey J (1954) Causation, regression, and path analysis. In: Kempthorne O (ed) Statistics and mathematics in biology. Iowa State College Press, Ames, pp 35–66

van de Geer S (2016) Estimation and Testing Under Sparsity: École d'Été de Probabilités des Saint-Flour XLV – 2015. Lecture notes in mathematics 2159. Springer, Berlin

van de Geer S, Bühlmann P, Zhou S (2011) The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). Electron J Stat 5:688–749

Wang Y, Blei D (2018) The blessings of multiple causes. To appear in J. Amer. Statist. Assoc. Preprint arXiv:1805.06826

Yu B (2013) Stability. Bernoulli 19(4):1484–1500