Journal of Statistical Planning and Inference 143 (2013) 1869–1871

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Rejoinder

Peter Bühlmann [a,*], Philipp Rütimann [a], Sara van de Geer [a], Cun-Hui Zhang [b]

[a] *Seminar for Statistics, ETH Zurich, Switzerland*
[b] *Department of Statistics, Rutgers University, United States*

We thank all discussants for their thoughtful and inspiring comments. We also thank the editors, especially Anirban DasGupta, for arranging this discussion. We structure our rejoinder according to various topics and items which have been raised.

1. Clustering based on canonical correlations

We thank Bien and Wegkamp for pointing out that "Although this method is not the focus of the paper or this discussion, we feel that this interesting method deserves more attention and may have applications in other areas". Thus, we emphasize here more clearly our main motivation for such a *novel* clustering algorithm.

First, the main reasons why one should use clustering based on canonical correlation are described in Section 1.1 of the paper with the following words. "We primarily propose to use canonical correlation for clustering the variables as this reflects the notion of linear dependence among variables: and it is exactly this notion of linear dependence which causes the identifiability problems in the linear model. Hence, this is conceptually a natural strategy for clustering variables when having the aim to address identifiability problems with variable selection in the linear model."

Second, it may be somewhat surprising at first sight that a greedy bottom-up algorithm can find an optimal solution (Theorem 2.1 in the paper). Thus, unlike as for many other clustering problems, the computation is astonishingly simple and clearly distinguishes itself from ad-hoc algorithms.

If these two points are driving factors for a certain application, one should indeed make use of our clustering method based on canonical correlations.

2. Our main target: variable screening

Our main target and interest in the paper is the property of variable screening: with high probability, the estimated set of variables $\hat{S}$ should contain the true active set S_0, i.e., $\hat{S} \supseteq S_0$. Ritov and Gershon write that for such a target "this [our proposed method] makes a lot of sense". Clearly, in a high correlation scenario for the covariables, the Lasso fails as accurate method for variable screening: our proposed methods do a much better job at the price of a larger set $\hat{S}$, and this is the reason why we are plotting the empirical results as a function of $|\hat{S}|$ (Figs. 3–6 in the paper).

Once we have a good variable screening we typically would continue with a subsequent analysis using the selected variables from $\hat{S}$ only. When Ritov and Gershon write that "when keeping the false negative rate down, may mean being very relaxed on the false positive rate" we implicitly have in mind that the issue with the false positives should be addressed in a second stage. For this task, we can use many procedures, potentially also from the low-dimensional setting if $|\hat{S}| \ll n$. Bien and Wegkamp propose the interesting CRL+Lasso method and they report good results for it: the procedure fits into the general philosophy to do re-estimation based on the selected variables only.

We note that for prediction, the plain Lasso is overall among the best methods, as reported in the paper and also found by Bien and Wegkamp. This finding is rather obvious: for highly correlated scenarios, the Lasso typically picks only one representative from a group of highly correlated variables but due to high collinearity within such a group, one

[*] Corresponding author.
E-mail address: peter.buehlmann@stat.math.ethz.ch (P. Bühlmann).

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.05.018

representative is sufficient for accurate prediction. However, prediction is not our main target (Bien and Wegkamp's empirical results are on prediction only). We included prediction performance in our paper to see whether good screening capacity could be observed from cross-validated prediction accuracy: unfortunately, this is not really the case and it seems difficult to judge for a specific data-set a method's performance for screening.

Ritov and Gershon make many general comments about high-dimensional inference, and we agree that many subtleties remain.

3. Various methods for supervised selection and estimation of clusters

A main issue, raised by both Bien and Wegkamp and Shah and Samworth, concerns the way how to select the clusters and estimate the corresponding parameters. Bien and Wegkamp present a very nice empirical study including other interesting proposals, for example the sparse group penalty which is only briefly mentioned in the paper (i.e. the "CSGL" method in Bien and Wegkamp).

3.1. The standard group Lasso penalty and the groupwise prediction penalty

Bien and Wegkamp as well as Shah and Samworth reproduce our finding that the groupwise prediction penalty (i.e., the CGL method) is often performing rather poorly. Bien and Wegkamp give a good explanation for it, namely that the within group design matrix can be rather ill-posed causing potential problems, and we do not have additional insights why CGL fails to be competitive. Our motivation to employ the groupwise prediction penalty comes from parameterization invariance. We can reparameterize (with a one-to-one mapping) within any group G_r, i.e., we can take any basis within group G_r: then, the estimated active set $\hat{S}$ and the prediction remain the same (Bühlmann and van de Geer, 2011, Section 4.5.1). This nice property seems to have its price, as clearly illustrated in the paper and the empirical results from Bien and Wegkamp and Shah and Samworth. In addition, the clustering itself is not parameterization invariant (i.e., reparameterizing groups and using the transformed variables as input for the cluster algorithm leads to a different clustering), see also the proposal by Shah and Samworth for clustering with sign-correction. In view of this, parameterization invariance with given groups of variables (as for the CGL method) might be less relevant in the overall context when taking clustering into account as well.

Bien and Wegkamp as well as Shah and Samworth report good empirical results when using the standard group Lasso penalty (denoted in Bien and Wegkamp by CUGL). Bien and Wegkamp also provide a nice interpretation in the extreme case where all variables in a group are the same: then, CRL equals CUGL while CGL is not unique. A very similar analysis is presented by Shah and Samworth for the more general case when the within-group correlation is large: again, the CUGL is then approximately equal to the CRL. Thus, in view of the empirical and methodological findings from Bien and Wegkamp and Shah and Samworth, the standard group Lasso penalty (CUGL) seems to be a very worthwhile proposal. However, it remains unclear to us how the empirical findings generalize to other scenarios (note that for the real-data design $\mathbf{X}$ in Section 5.2 of the paper, the CGL is competitive and slightly best for variable screening). The penalty of the CGL hinges on the group sizes, the within group empirical covariance and the within group structure of the regression, e.g. whether the coefficients are sparse or have the same sign within groups; the penalty of the standard group Lasso (CUGL) is independent of the design within groups. Ideally, we would have an adaptive rule which would select an appropriate penalty (and if only interested in prediction, which is not our main target of interest, cross-validation might be used towards achieving this goal).

Bien and Wegkamp mention for the CGL method that "the established oracle inequality seems wasteful as the rate is proportional to $\sum_{r\in S_0} m_r$, the total number of elements in the active groups, not the total number of active variables as is the case with the plain Lasso". We think that this is the usual drawback of a group Lasso method, whether with the groupwise prediction or the standard group Lasso penalty. Further regularization within the groups, especially if the groups are large, leads to better performance: examples include the sparsity-smoothness penalty where better oracle inequalities are established (Meier et al., 2009), or the sparse group penalty (Simon et al., 2012). In addition, another issue raised by Bien and Wegkamp, namely that "The authors assume the design matrix $\mathbf{X}$ to be fixed in the analysis of the CGL-method but assume it to be multivariate normal in the analysis of the CRL-method. These differing assumptions hamper direct comparison of the two methods", is not of fundamental nature, and we can easily state results for the CRL with fixed design as well (see also Proposition 4.2. in the paper).

3.2. The simple CRL and Shah and Samworth's sign correction

We share the sentiment expressed by Shah and Samworth that "we were initially surprised at how convincingly it [the cluster group Lasso] was trumped by the ostensibly more naive cluster representative Lasso under most of the simulation settings." Shah and Samworth come up with an interesting modification of the settings in the paper: when some of the variables are negatively correlated (and the signs of the regression coefficients are coupled with the correlation structure of the variables to obtain the same signal as in the paper), the CRL method can be very poor while other methods perform substantially better. Whether such a model is approximately reflecting some "realistic" scenarios in applications is another issue: if the variables are positively correlated, however, the cancellation phenomenon in Shah and Samworth's model does not occur.

Shah and Samworth then propose a slightly more sophisticated “sign-corrected” version of the CRL method and they illustrate impressive improvement for scenarios where CRL fails to be reasonably accurate. Thus, when sticking to a simple procedure, Shah and Samworth's sign-correction seems promising while the group Lasso with standard group Lasso penalty (CUGL) is an obvious competitor which performed as well in the considered settings.

4. The advantage of clustering

4.1. The elastic net and other methods without grouping

As mentioned by all three discussants, the elastic net is indeed a good method to cope with correlated variables. However, and in sharp contrast to the cluster Lasso procedures, the elastic net does not provide a grouping of the variables.

Ritov and Gershon point to other ideas of dimensionality reduction: “in Gershons method, a [small] subset of the variables is selected such that any other variable would be well linearly explained by the chosen ones”. This is a very interesting proposal, perhaps as or even more suitable than clustering based on canonical correlations, but it does not lead to groups or clusters of the variables.

Whether we want groups or not depends of course on the application. A good example, in favor of clusters, is the analysis of genomic data from single nucleotide polymorphisms (SNPs): single SNPs are typically not identifiable (due to the very large number of SNPs in the millions), a group of correlated SNPs corresponds well with a spatial neighborhood in the genome sequence and thus, identifying relevant groups has a direct interpretation in terms of relevant genomic regions.

4.2. Identifiability of groups and resolution of clustering

What we have primarily in mind with the cluster Lasso methods is, besides variable screening with an estimator $\hat{S}$, to infer groups of variables which are “relevant” for explaining a response variable.

Ideally, the partitioning into groups (or clustering) is as fine as possible while the groups are still sufficiently well-identifiable from data: in view of near non-identifiability of single highly correlated variables, however, the chosen clustering will largely not consist of single variables. Ideally, a method would choose the resolution of clustering, i.e., the number of clusters, in such a way that it provides the finest partition whose corresponding groups are still identifiable from data.

We have indicated in the paper that cross-validated prediction is not good enough for this task: for example, plain Lasso on single variables leads to powerful prediction yet it typically leads to non-identifiable variables in high correlation settings. An alternative route is indicated by Shah and Samworth. One could aim for some significance measure for single clusters, using for example stability selection (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) or *p*-values (Meinshausen et al., 2009; Zhang and Zhang, 2011; Bühlmann, in press; van de Geer et al., 2013). Based on such *p*-values, one can test hierarchically in a top-down manner, as described in Meinshausen (2008) and generalized by Goeman and Solari (2010), until the clusters are not significant anymore: this indeed would provide a finest resolution which provides significant groups for explaining the response (we have work in progress in this direction).

References

Bühlmann, P. Statistical significance in high-dimensional linear models. Bernoulli, in press.
Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data: Methods Theory and Applications. Springer Verlag.
Goeman, J., Solari, A., 2010. The sequential rejection principle of familywise error control. Annals of Statistics 38, 3782–3810.
Meier, L., van de Geer, S., Bühlmann, P., 2009. High-dimensional additive modeling. Annals of Statistics 37, 3779–3821.
Meinshausen, N., 2008. Hierarchical testing of variable importance. Biometrika 95, 265–278.
Meinshausen, N., Bühlmann, P., 2010. Stability selection (with discussion). Journal of the Royal Statistical Society Series B 72, 417–473.
Meinshausen, N., Meier, L., Bühlmann, P., 2009. *P*-values for high-dimensional regression. Journal of the American Statistical Association 104, 1671–1681.
Shah, R., Samworth, R., 2013. Variable selection with error control: Another look at Stability Selection. Journal of the Royal Statistical Society Series B 75, 55–80.
Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group Lasso. Journal of Computational and Graphical Statistics 22, 231–245.
van de Geer, S., Bühlmann, P., Ritov, Y., 2013. On asymptotically optimal confidence regions and tests for high-dimensional models. arxiv:1303.0518.
Zhang, C.-H., Zhang, S., 2011. Confidence intervals for low-dimensional parameters with high-dimensional data. arxiv:1110.2563v1.