Discussion of 'Correlated variables in regression: clustering and sparse estimation' by Peter Bühlmann, Philipp Rütimann, Sara van de Geer and Cun-Hui Zhang.

Rajen D. Shah and Richard J. Samworth

January 12, 2013

We would like to begin by congratulating the authors on their fine paper. Handling highly correlated variables is one of the most important issues facing practitioners in high-dimensional regression problems, and in some ways it is surprising that it has not received more attention up to this point. The authors have made substantial progress towards practical methodological proposals, however, and we are sure the paper will stimulate considerable future research. In this discussion, we present a possible improvement to the cluster representative Lasso, give some further insights into the cluster group Lasso and conclude with some brief remarks on one possible new direction suggested by the work.

1 Variable cancellation and the cluster representative Lasso

In terms of variable selection, the cluster representative Lasso is without doubt the clear winner from the simulation studies. It even performs rather well, relative to the competing methods, when coefficients in the same group may have opposite signs (scenarios (Ac), (Ad), (Bc) and (Bd)). One way of understanding its success is to realise that when groups consist of highly positively correlated variables, any linear combination of them can be represented reasonably well by some multiple of the average of these variables.

However, when some of the variables in a group may be negatively correlated with one another, the cluster representative Lasso can perform quite poorly. Note that the proposed bottom-up hierarchical agglomerative clustering algorithm can certainly result in groups containing negative correlations since it is invariant under multiplying any of the variables by -1. In these situations, when taking averages of variables within a group, pairs of very negatively correlated variables can almost cancel each other out. The result is that the cluster representatives may have little correlation with any of their respective group members. In Figure 1, we show an extreme case of this phenomenon. The top panel shows the results of repeating simulations (Aa)–(Ad) but with two important differences: the blocks comprising the covariance matrix of X are

$$\Gamma^{-}_{ij} = \begin{cases} 1 & \text{if } \ i = j, \\ -\rho & \text{if } \ i \equiv j + 1 \ (\text{mod } 2), \\ \rho & \text{otherwise,} \end{cases} \tag{1.1}$$

with $\rho = 0.9$; and the coefficient vector β^0 correspondingly has its components with even indices multiplied by -1. Thus the signal $X^T\beta^0$ has the same distribution as in the simulations

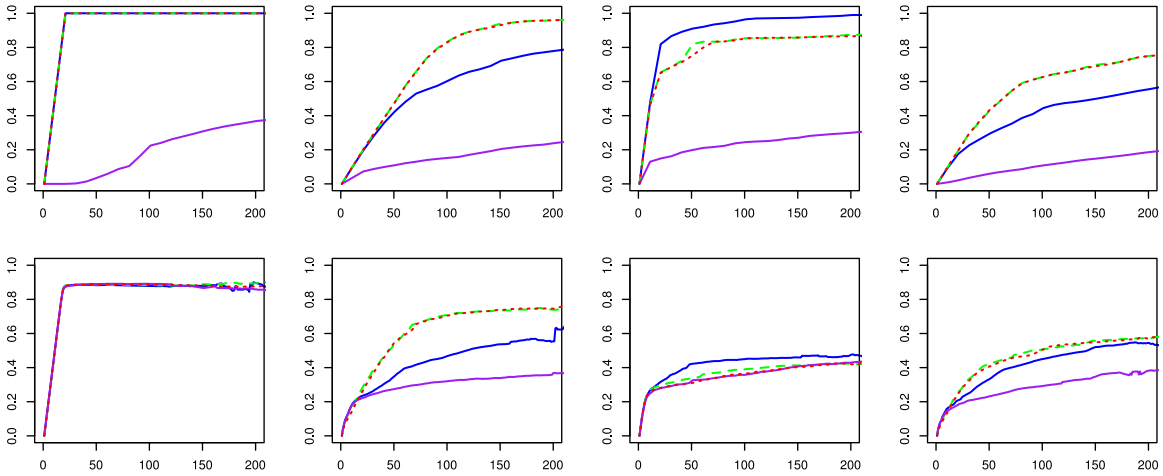

Figure 1: Plots of $|\hat{S} \cap S_0|/|S_0|$ against $|\hat{S}|$ for the cluster group Lasso with a groupwise prediction penalty (solid blue); the cluster group lasso with a standard group penalty (green dashes); the cluster representative lasso (solid purple); and the cluster representative lasso with sign corrections (red dots). Modified versions of scenarios (Aa)–(Ad) with $\sigma = 3$ are left to right, and the top and bottom panels have $\rho = 0.9$ and 0.6 respectively.

in the paper. The cluster representative Lasso is here hardly better at variable selection than random guessing. When we reduce ρ to 0.6 (bottom panel), its performance improves as the cancellation effect is not quite as dramatic, but it is still lagging behind its competitors. In some ways this problem is more serious than coefficients in the same group having opposite signs because in that case, the magnitude of the signal drops and this presents a challenging situation for almost all methods. The variable cancellation effect is, however, an issue unique to the cluster representative Lasso.

One way of solving this problem is to multiply certain variables by -1 in order maintain mostly positive correlations among variables within the same group, and then create cluster representatives from this 'sign-corrected' design matrix. This can be accomplished by adding a further step to the clustering algorithm: when two clusters to be merged have been identified, compute the scalar product between the cluster representatives of each of these clusters. If the scalar product is negative, multiply each of the variables in one of the clusters (the one containing the variable with the lowest index, for example) by -1. Thinking inductively, suppose that at the b^{th} iteration, most variables in clusters are highly positively correlated with their respective cluster representatives. Then the sign correction described above should ensure that the same holds true for the newly merged cluster at the $(b+1)^{\text{th}}$ iteration.

The greatly improved performance of the cluster representative Lasso with this 'sign correction' applied is shown by the red dots in Figure 1. One additional benefit of the sign correction is that the magnitudes of the estimated coefficients are invariant under multiplying any of the variables by -1; this property, of course, does not hold for the vanilla cluster representative Lasso.

2 The cluster group Lasso and the groupwise prediction penalty

Using a group Lasso-type penalty with groups defined by the clusters seems like a sensible idea, and we were initially surprised at how convincingly it was trumped by the ostensibly more naive cluster representative Lasso under most of the simulation settings. The cluster group Lasso did, however, improve when half of the signs of the true active coefficients were switched. One way of explaining this is to observe that the success of penalised regression methods relies on the penalty applied to the true coefficient vector (or at least some suitable surrogate), having a relatively low value. This is easiest to see when we look at the constrained form of the optimisation problems: for given values of the tuning parameters, the optimisation problem is equivalent to searching among all coefficient vectors whose penalty contribution is bounded by some value, and picking the one which minimises the empirical risk term. If the true coefficient vector (or its surrogate) is far away from this constraint set, there is no hope of recovering it or anything close to it by such a method.

Let us suppose we have q groups $G_1, \ldots, G_q$ and that, for simplicity,

$$\tfrac{1}{n}\left\{(\mathbf{X}^{(G_r)})^T\mathbf{X}^{(G_r)}\right\}_{ij} = \begin{cases} 1 & \text{if } \ i = j, \\ \rho & \text{otherwise,} \end{cases} \tag{2.1}$$

for $1 \leq r \leq q$ and $i, j \leq g_r$, where $g_r = |G_r|$ and $\rho \in (0, 1)$. Denoting the average of the components of a vector u by $\bar{u}$, we see that for a coefficient vector β, the groupwise prediction penalty is proportional to

$$\sum_{r=1}^{q} \frac{g_r^{1/2}}{n^{1/2}} \|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2 = \sum_{r=1}^{q} g_r^{1/2}\Big\{(\rho g_r + 1 - \rho)g_r\bar{\beta}_{G_r}^2 + (1-\rho)\left\|\beta_{G_r} - \bar{\beta}_{G_r}\mathbf{1}_{g_r}\right\|_2^2\Big\}^{1/2}.$$

Thus when the groupwise averages of the true coefficient vector are large and ρ is close to 1, as was the case in most of the simulation settings, the groupwise prediction penalty becomes large and so the cluster group Lasso performs poorly. Conversely, when the groupwise averages are relatively small, as occurred when half the signs of the active coefficients were switched, the performance of the cluster group Lasso improves.

In general, our experience is that the groupwise prediction penalty is not always well-suited to situations with high correlations within groups. In fact, the standard group Lasso penalty (green dashes in Figure 1) does better or at least as well in all cases except scenario (Ac) in our simulations. Interestingly, the ROC-type performance curves it traces are remarkably similar to those of the cluster representative Lasso. To see why this is the case, suppose we have a design matrix $\mathbf{X}$ with within cluster covariance structure given by (2.1). For a fixed λ, consider applying the cluster group Lasso with a standard group Lasso penalty:

$$\hat{\beta} = \hat{\beta}(\lambda) \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \left\{ \tfrac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\sum_{r=1}^{q} g_r^{1/2}\|\beta_{G_r}\|_2 \right\}.$$

Now for fixed ℓ, let $\mathbf{R} = \mathbf{Y} - \sum_{r\neq\ell}\mathbf{X}^{(G_r)}\beta_{G_r}$. Provided $\hat{\beta}$ is the unique minimiser and $\hat{\beta}_{G_\ell} \neq 0$, there exists by strong duality $\mu = \mu(\lambda)$ such that

$$\hat{\beta}_{G_\ell} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{g_\ell}} \left\{ \tfrac{1}{n}\left\|\mathbf{R} - \mathbf{X}^{(G_\ell)}\beta\right\|_2^2 + \mu\|\beta\|_2^2 \right\}.$$

Letting $\alpha = \{(\mathbf{X}^{(G_\ell)})^T\mathbf{X}^{(G_\ell)}\}^{-1}(\mathbf{X}^{(G_\ell)})^T\mathbf{R}$, we have

$$\begin{aligned}\hat{\beta}_{G_\ell} &= \{(\mathbf{X}^{(G_\ell)})^T\mathbf{X}^{(G_\ell)} + n\mu I_{g_\ell}\}^{-1}(\mathbf{X}^{(G_\ell)})^T\mathbf{X}^{(G_\ell)}\alpha \\ &= \frac{1-\rho}{1-\rho+\mu}(\alpha - \bar{\alpha}\mathbf{1}_{g_\ell}) + \frac{1+(g_\ell-1)\rho}{1+\mu+(g_\ell-1)\rho}\bar{\alpha}\mathbf{1}_{g_\ell}.\end{aligned}$$

Thus, certainly when ρ is close to 1, we see that $\hat{\beta}_{G_\ell}$ has almost constant components. This is similar to the situation with the cluster representative Lasso, where the estimated coefficient vector is constrained to be exactly constant within each group.

3 Use of clustering in high-dimensional inference

Several methods for carrying out high-dimensional inference that are currently in popular use rely on marginal measures of the significance of individual variables. An important example is the Stability Selection methodology for variable selection proposed in Meinshausen and Bühlmann (2010) and further developed in Shah and Samworth (2013). Such algorithms are vulnerable in situations with two or more highly correlated signal variables. In the case of Stability Selection, this is because these variables can 'split the vote', with the result that neither/none of the variables is chosen. Similar issues arise with other variable selection methods based on p-values for the additional effect of each individual variable. Switching the focus to the significance of clusters (determined for instance using the methodology outlined in the paper), rather than individual variables, offers the potential to alleviate these difficulties, and promises to provide a fruitful direction for future research.

References

Meinshausen, N. and Bühlmann, P. (2010) Stability selection. *J. Roy. Statist. Soc., Ser. B (with discussion)*, **72**, 417–473.

Shah, R. D. and Samworth, R. J. (2013) Variable selection with error control: Another look at Stability Selection. *J. Roy. Statist. Soc., Ser. B*, **75**, 55–80.