DISCUSSION OF "CORRELATED VARIABLES IN REGRESSION: CLUSTERING AND SPARSE ESTIMATION"

JACOB BIEN AND MARTEN WEGKAMP

Brief Review

We congratulate Bühlmann, Rütimann, van de Geer, and Zhang (hereafter BRVZ) on their inspiring paper that addresses the important issue of correlated covariates in the high dimensional regression model

$$Y = \mathbf{X}\beta^0 + \varepsilon.$$

Here Y is the response vector in $\mathbb{R}^n$, $\mathbf{X}$ is an $n \times p$ matrix, $\beta^0 \in \mathbb{R}^p$ is the vector of coefficients, and finally $\varepsilon \in \mathbb{R}^n$ is assumed to be multivariate normal with mean zero and covariance matrix $\sigma^2 I$. While it has been shown that the lasso, and its many variants, "work" in terms of variable selection and prediction, they work best for near orthogonal cases of $\mathbf{X}$. However, if $p > n$, correlation among the covariates is obviously inevitable. It is worth pointing out that the fit of the lasso estimator $\widehat{\beta}$ always satisfies

$$2\|\mathbf{X}\widehat{\beta} - \mathbf{X}\beta^0\|_2^2 \leq 3n\lambda\|\beta^0\|_1$$

for $\lambda = 8\sigma\sqrt{\log p/n}$ with probability $1 - 2/p$, under no conditions on $\mathbf{X}$. See Bühlmann and van de Geer (2011) in the regression context or Wegkamp and Yuan (2011) in the context of sparse support vector machines. However, faster rates for the prediction error and consistent variable selection do require (compatibility and irrepresentable) conditions on $\mathbf{X}$ (Bühlmann and van de Geer, 2011; Bunea, 2008).

BRVZ propose two methods to compete with the lasso estimator. Both use as a first step a novel agglomerative hierarchical clustering algorithm based on the empirical canonical correlations of $\mathbf{X}$. It is important to note that the clustering is unsupervised—the vector Y of responses is not used. 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. It is guaranteed to find a partition with the maximal empirical canonical correlation between two clusters less than a set threshold τ. The authors advocate taking the smallest possible value for τ. If it exists, this algorithm will find the finest clustering with this property (τ-separation). When the rows of $\mathbf{X}$ are iid $N_p(0, \Sigma)$ and the maximal canonical correlation between groups is less than the minimal maximal canonical correlation within groups and $\text{rank}(\Sigma_{G_r, G_r}) = o(n)$, the algorithm

is shown to be consistent. In the same multivariate normal setting, hierarchical clustering based on the sample correlations between two covariates, consistently finds the true clusters, provided the minimal correlation within clusters exceeds the maximal correlation between clusters and $\log p = o(n)$. In the simulations it is shown to have poor performance in presence of a single large cluster.

After finding q clusters, BRVZ propose two alternative methods:

CRL: Each cluster of covariates is collapsed into one vector, the average of the covariates in that cluster, and a new $n \times q$ design matrix $\bar{\mathbf{X}}$ is formed. Fit the lasso (Tibshirani, 1996):

$$\min_{\beta \in \mathbb{R}^q} \|Y - \bar{\mathbf{X}}\beta\|_2^2 + \lambda \|\beta\|_1.$$

CGL: Taking the clusters as groups, fit the group lasso (Yuan and Lin, 2006):

$$\min_{\beta \in \mathbb{R}^p} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \sum_{r=1}^{q} \sqrt{|G_r|} \|\mathbf{X}^{(G_r)} \beta_{G_r}\|_2.$$

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. For the CRL method, the reduction in dimension from p to q is often substantial and the compatibility constant is much less. The obvious main drawback of this method is the bias incurred: the coefficient $\widehat{\beta}_i$ is the same for all covariates in the same group. In the case of Gaussian design, it is shown that the rate is good, provided the bias is small. CGL works well when there are many active variables within the few active groups. For fixed design, the compatibility constant for the group lasso is oftentimes less than that of the plain lasso. However, 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.

Related Methods and Additional Analysis

The question of how to improve upon the lasso when the features are highly correlated is a fundamental one. One of the most well-known attempts to deal with this situation is the *elastic net* (Zou and Hastie, 2005). In this method, a ridge penalty (based on the squared ℓ_2 norm) is added to the objective function:

$$\hat{\beta}^{(\mathrm{EN})} = \arg\min_{\beta} \|y - \mathbf{X}\beta\|_2^2 + \lambda\alpha\|\beta\|_1 + \lambda(1-\alpha)\|\beta\|_2^2.$$

If a group of highly correlated variables is present in the design matrix and $\alpha < 1$, the coefficients will be shrunken towards each other (whereas the

lasso may exclude all but one of the variables). More precisely, Zou and Hastie (2005) show that if ρ_{jk} is the correlation between the jth and kth feature (and the response has been centered and predictors standardized), then

$$|\hat{\beta}_j^{(\mathrm{EN})} - \hat{\beta}_k^{(\mathrm{EN})}| \leq \frac{\|y\|_1}{\lambda(1-\alpha)}\sqrt{2(1-\rho_{jk})}, \tag{1}$$

provided $\hat{\beta}_j^{(\mathrm{EN})}\hat{\beta}_k^{(\mathrm{EN})} > 0$. The CRL method can be seen as a more extreme approach, in which highly correlated variables are given exactly equal weight. Since CRL enforces the grouping more aggressively, we might expect it to do better when the clustering has succeeded in identifying the correct grouping but to do worse when working with the wrong grouping.

BRVZ perform a simulation study comparing their two methods (CRL and CGL) to the lasso. In what follows, we augment part of the simulation study to include the elastic net and several variants of CRL and CGL. This points to some future directions along this line of work and raises a few interesting points. The methods we compare are as follows (we use the R package `glmnet` for all lasso and elastic net fits):

- CRL: The cluster representative lasso as proposed by the authors.
- CRL+Lasso: We use CRL to select the nonzero groups and then perform a lasso on the remaining variables. This allows for within-group sparsity and relaxes the constraint that all coefficients be equal.
- CGL: The cluster group lasso as proposed by the authors, which makes use of a less common form of the group lasso known as the "groupwise prediction penalty" or "standardized group lasso" penalty: $\sum_r w_r \|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2$ (Bühlmann and van de Geer, 2011; Simon and Tibshirani, 2012). We use the R package `standGL`.
- CUGL: Here we perform CGL but use the "unstandardized" group lasso penalty, $\sum_r w_r \|\beta_{G_r}\|_2$. We use the R package `SGL` for this method and the next one.
- CSGL: Here we perform CGL but use the *sparse* group lasso penalty (see, e.g., Simon et al. 2012): $\alpha\|\beta\|_1 + (1-\alpha)\sum_r w_r\|\beta_{G_r}\|_2$.
- EN: The elastic net as discussed above. Unlike all other methods, EN does not require first clustering the features.

All tuning parameters were selected using a validation set of size 100 to exhibit each method's potential. In all other details we followed the paper's description as closely as possible. Table 1 shows the out-of-sample test errors for the block-diagonal and single-block models of Section 5.1.1 and 5.1.2.

Perhaps the most striking observation is the large discrepancy between CGL's performance and the rest of the methods (both in the paper's simulations and ours). Comparing CGL to CUGL, we conclude that the difference must lie in the choice of group lasso penalty. Although BRVZ claim that CGL uses "a much more appropriate penalty", we find the "unstandardized" group lasso penalty to give much better results in these simulations. We would be particularly interested in the authors' thoughts on this issue in

	A(a)	A(b)	A(c)	A(d)
Lasso	12.27 (0.17)	17.34 (0.30)	12.71 (0.14)	16.70 (0.31)
CRL	11.11 (0.16)	17.28 (0.29)	13.03 (0.13)	16.81 (0.28)
CRL+Lasso	10.88 (0.11)	16.45 (0.34)	11.66 (0.16)	15.94 (0.30)
CGL	15.96 (0.29)	38.53 (0.51)	13.76 (0.29)	27.45 (1.06)
CUGL	12.65 (0.15)	17.12 (0.28)	12.84 (0.12)	16.49 (0.27)
CSGL	12.67 (0.15)	16.75 (0.28)	12.55 (0.12)	16.07 (0.27)
EN	11.87 (0.16)	17.27 (0.30)	12.66 (0.13)	16.67 (0.31)
	B(a)	B(b)	B(c)	B(d)
Lasso	12.37 (0.16)	22.11 (0.46)	12.48 (0.16)	22.45 (0.38)
CRL	15.05 (0.19)	22.89 (0.46)	13.41 (0.18)	23.34 (0.41)
CRL+Lasso	11.45 (0.13)	21.84 (0.48)	11.91 (0.17)	22.20 (0.42)
CGL	415.92 (2.99)	65.94 (1.61)	41.91 (5.40)	38.51 (1.44)
CUGL	15.55 (0.21)	22.04 (0.42)	13.35 (0.18)	22.35 (0.40)
CSGL	15.14 (0.20)	21.41 (0.40)	12.71 (0.18)	21.45 (0.36)
EN	12.00 (0.16)	22.15 (0.44)	12.40 (0.16)	22.44 (0.38)

TABLE 1. Average test MSE for the block-diagonal and single-block models with $\sigma = 3$. Standard errors given in parentheses.

light of our empirical findings. Since the theoretical oracle bounds for CGL assume a fixed design matrix $\mathbf{X}$, we also computed the in-sample errors, shown in Table 2. In the"A" models, the in-sample errors are indeed lower and much closer to the (out-of-sample) test errors of the other methods. Still, in the "B" models, the errors remain far higher than those of other methods. One potentially important difference between the "A" and "B" scenarios is in the conditioning of $\mathbf{X}$. In the "A" scenario, the clustering method returned the correct partition, consisting of 100 groups of size 10, with the (average) condition number of each group's design matrix, $\mathbf{X}^{(G_r)}$, being 12.8. By contrast, in the "B" scenarios, the clustering returned one group of size 34 containing the 30 correlated variables and 4 noise variables. The condition number of this selected group's design matrix is quite high: 36.5. We suspect that the poor conditioning of this matrix makes the penalty $\|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2$ undesirable here. In further support that this penalty may not be performing well, we find that CGL tends to select a highly regularized model. If conditioning of $\mathbf{X}^{(G_r)}$ is indeed the problem, a better alternative to $\|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2 = \sqrt{\beta_{G_r}^T \mathbf{X}^{(G_r)T}\mathbf{X}^{(G_r)}\beta_{G_r}}$ would be to use $\sqrt{n\beta_{G_r}^T \widetilde{\Sigma}^{(G_r)}\beta_{G_r}}$, where $\widetilde{\Sigma}^{(G_r)}$ is a better-conditioned estimate of the within group covariance matrix, such as $\widetilde{\Sigma}^{(G_r)} = n^{-1}\mathbf{X}^{(G_r)T}\mathbf{X}^{(G_r)} + \rho I_{|G_r|}$ (Ledoit and Wolf, 2004). This is similar to the "ridged group lasso" suggested in Simon and Tibshirani (2012).

	A(a)	A(b)	A(c)	A(d)	B(a)	B(b)	B(c)	B(d)
CGL	13.29	17.82	12.39	16.07	339.22	65.22	37.10	36.02

TABLE 2. In-sample error (using training set $\mathbf{X}$) for the block-diagonal and single-block models with $\sigma = 3$.

It appears that our proposed method CRL+Lasso performs the best. It enjoys the group screening ability of CRL while still acting like the lasso on individual features. Comparing CSGL to CUGL, we see that adding a sparsity term to the group lasso may be (mildly) beneficial in these scenarios. Finally, we observe that the elastic net, despite its simplicity, remains quite competitive.

Analysis of the Methods in an Extreme Case

The authors imagine a situation in which the features cluster into highly correlated groups. We take this idea to the extreme here, and consider what the paper's methods do when the predictors within each group of $\mathcal{G} = \{G_1, \ldots, G_q\}$ are *identical*. That is, let

$$\mathbf{X} = [x_1 1_{|G_1|}^T : \cdots : x_q 1_{|G_q|}^T] \in \mathbb{R}^{n \times p}.$$

(Here $1_N = (1, \ldots, 1)^T \in \mathbb{R}^N$.) For simplicity, we assume that $\|x_j\|_2 = 1$ and write $\widetilde{\mathbf{X}} = [x_1 : \cdots : x_q] \in \mathbb{R}^{n \times q}$. Suppose that the columns of $\widetilde{\mathbf{X}}$ are close enough to orthogonal that canonical correlation clustering results in the partition $\mathcal{G}$. Then, for CRL, we solve

$$\hat{\gamma} = \arg\min_{\gamma} \|y - \widetilde{\mathbf{X}}\gamma\|_2^2 + \lambda\|\gamma\|_1. \tag{2}$$

In a slight notational departure from the paper, we define $\hat{\beta}_{\mathrm{CRL}}$ to be the p-vector used with the original design matrix $\mathbf{X}$. That is, the rth block of $\hat{\beta}_{\mathrm{CRL}}$ is given by

$$\hat{\beta}_{\mathrm{CRL},G_r} = |G_r|^{-1}\hat{\gamma}_r 1_{|G_r|},$$

where the factor of $|G_r|$ makes it so that $\widetilde{\mathbf{X}}\hat{\gamma} = \mathbf{X}\hat{\beta}_{\mathrm{CRL}}$.

Now, CGL does not have a unique solution in this context, but we can describe its set of solutions:

$$\begin{aligned}\mathcal{B}_{\mathrm{CGL}} &= \arg\min_{\beta} \|y - \sum_{r=1}^{q} x_r 1_{|G_r|}^T \beta_{G_r}\|_2^2 + \lambda \sum_{r=1}^{q} w_r \|x_r 1_{|G_r|}^T \beta_{G_r}\|_2 \\ &= \arg\min_{\beta} \|y - \sum_{r=1}^{q} x_r 1_{|G_r|}^T \beta_{G_r}\|_2^2 + \lambda \sum_{r=1}^{q} w_r |1_{|G_r|}^T \beta_{G_r}|,\end{aligned}$$

where we have used that $\|x_r\|_2 = 1$. It is easy to see that

$$\mathcal{B}_{\mathrm{CGL}} = \{\beta \in \mathbb{R}^p : 1_{|G_r|}^T \beta_{G_r} = \hat{\delta}_r\},$$

where

$$\hat{\delta} = \arg\min_{\delta} \|y - \widetilde{\mathbf{X}}\delta\|_2^2 + \lambda \sum_{r=1}^{q} w_r |\delta_r|.$$

Notice that if $w_1 = \cdots = w_q$ (i.e., all groups have the same size), then the above problem would be identical (up to λ) to (2). In this case, we conclude that $\hat{\beta}_{\mathrm{CRL}} \in \mathcal{B}_{\mathrm{CGL}}$. Finally, we turn to what we called earlier the unstandardized group lasso:

$$\hat{\beta}_{\mathrm{CUGL}} = \arg\min_{\beta} \|y - \sum_{r=1}^{q} x_r 1_{|G_r|}^T \beta_{G_r}\|_2^2 + \lambda \sum_{r=1}^{q} w_r \|\beta_{G_r}\|_2.$$

We rewrite the problem as

$$\min_{\beta,\delta} \left\{ \|y - \widetilde{\mathbf{X}}\delta\|_2^2 + \lambda \sum_{r=1}^{q} w_r \|\beta_{G_r}\|_2 \text{ s.t. } 1_{|G_r|}^T \beta_{G_r} = \delta_r \right\}.$$

Minimizing over β first, we get $\hat{\beta}_{G_r}(\delta) = |G_r|^{-1}\delta_r 1_{|G_r|}$. Substituting in this expression leaves a minimization over δ:

$$\min_{\delta} \|y - \widetilde{\mathbf{X}}\delta\|_2^2 + \lambda \sum_{r=1}^{q} w_r |G_r|^{-1/2} |\delta_r|.$$

Thus, taking the standard choice, $w_r = |G_r|^{1/2}$, we see that $\hat{\beta}_{\mathrm{CRL}} = \hat{\beta}_{\mathrm{CUGL}}$.

Finally, we consider the elastic net in this situation:

$$\hat{\beta}_{\mathrm{EN}} = \arg\min_{\beta} \|y - \sum_{r=1}^{q} x_r 1_{|G_r|}^T \beta_{G_r}\|_2^2 + \lambda \sum_{r=1}^{q} \left[(1-\alpha)\|\beta_{G_r}\|_2^2 + \alpha \|\beta_{G_r}\|_1 \right].$$

By (1), we have that $\hat{\beta}_{\mathrm{EN}}$ is constant within group. Building this fact in as a constraint, we may reparametrize the problem using $\beta_{G_r} = |G_r|^{-1}\delta_r 1_{|G_r|}$ to get

$$\min_{\delta} \|y - \widetilde{\mathbf{X}}\delta\|_2^2 + \lambda \sum_{r=1}^{q} \left[(1-\alpha)|G_r|^{-1}\delta_r^2 + \alpha |\delta_r| \right].$$

We see that, in contrast to the other three methods, the "grouping effect" is applied between groups as well. This of course is to be expected since the elastic net does not start with a set of known groups, but rather applies a general shrinkage that has the desired grouping effect on correlated features. This feature of the elastic net may be beneficial if $\widetilde{\mathbf{X}}$ itself has highly correlated columns; however, if this is not the case, then the elastic net would be shrinking the estimates unnecessarily.

Acknowledgment

Marten Wegkamp is supported in part by NSF Grant DMS-10-07444.

References

Bühlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Verlag.

Bunea, F. (2008), 'Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization', *Electronic Journal of Statistics* **2**, 1153–1194.

Ledoit, O. and Wolf, M. (2004), 'A well-conditioned estimator for large-dimensional covariance matrices', *Journal of Multivariate Analysis* **88**(2), 365–411.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2012), 'A sparse-group lasso', *Journal of Computational and Graphical Statistics* .

Simon, N. and Tibshirani, R. (2012), 'Standardization and the group lasso penalty', *Statistica Sinica* **22**(3), 983–1001.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Wegkamp, M. and Yuan, M. (2011), 'Support vector machines with a reject option', *Bernoulli* **17**, 1368–1385.

Yuan, M. and Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society, Series B* **67**, 301–320.

Department of Biological Statistics and Computational Biology & Department of Statistical Science, Cornell University, 1178 Comstock Hall, Ithaca, NY 14853

Department of Statistical Science & Department of Mathematics, Cornell University, 1194 Comstock Hall, Ithaca, NY 14853