Journal of Statistical Planning and Inference 143 (2013) 1835–1858

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Correlated variables in regression: Clustering and sparse estimation

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

ARTICLE INFO

Available online 19 June 2013

Keywords:
Canonical correlation
Group Lasso
Hierarchical clustering
High-dimensional inference
Lasso
Oracle inequality
Variable screening
Variable selection

ABSTRACT

We consider estimation in a high-dimensional linear model with strongly correlated variables. We propose to cluster the variables first and do subsequent sparse estimation such as the Lasso for cluster-representatives or the group Lasso based on the structure from the clusters. Regarding the first step, we present a novel and bottom-up agglomerative clustering algorithm based on canonical correlations, and we show that it finds an optimal solution and is statistically consistent. We also present some theoretical arguments that canonical correlation based clustering leads to a better-posed compatibility constant for the design matrix which ensures identifiability and an oracle inequality for the group Lasso. Furthermore, we discuss circumstances where cluster-representatives and using the Lasso as subsequent estimator leads to improved results for prediction and detection of variables. We complement the theoretical analysis with various empirical results.

1. Introduction

High-dimensional regression is used in many fields of applications nowadays where the number of covariables p greatly exceeds sample size n, i.e., $p \gg n$. We focus here on the simple yet useful high-dimensional linear model

$$\mathbf{Y} = \mathbf{X}\beta^0 + \boldsymbol{\varepsilon}, \tag{1}$$

with univariate $n \times 1$ response vector $\mathbf{Y}$, $n \times p$ design matrix $\mathbf{X}$, $p \times 1$ true underlying coefficient vector β^0 and $n \times 1$ error vector $\boldsymbol{\varepsilon}$. Our primary goal is to do variable screening for the active set, i.e., the support of β^0, denoted by $S_0 = \{j;\ \beta_j^0 \neq 0,\ j = 1, \ldots p\}$: we want to have a statistical procedure $\hat{S}$ such that with high probability, $\hat{S} \supseteq S_0$ (and $|\hat{S}|$ not too large). In the case where $p \gg n$, the obvious difficulties are due to (near) nonidentifiability. While some positive results have been shown under some assumptions on the design $\mathbf{X}$, see the paragraph below, high empirical correlation between variables or near linear dependence among a few variables remain as notorious problems which are often encountered in many applications. Examples include genomics where correlation and the degree of linear dependence is high within a group of genes sharing the same biological pathway (Segal et al., 2003), or genome-wide association studies where SNPs are highly correlated or linearly dependent within segments of the DNA sequence (Balding, 2007).

An important line of research to infer the active set S_0 or for variable screening has been developed in the past using the Lasso (Tibshirani, 1996) or versions thereof (Zou, 2006; Meinshausen, 2007; Zou and Li, 2008). Lasso-type methods have

* 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.2013.05.019

proven to be successful in a range of practical problems. From a theoretical perspective, their properties for variable selection and screening have been established assuming various conditions on the design matrix $\mathbf{X}$, such as the neighborhood stability or irrepresentable condition (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006), and various forms of "restricted" eigenvalue conditions, see van de Geer (2007), Zhang and Huang (2008), Meinshausen and Yu (2009), Bickel et al. (2009), van de Geer and Bühlmann (2009), Sun and Zhang (2012). Despite these positive findings, situations where high empirical correlations between covariates or near linear dependence among a few covariables occur cannot be handled well with the Lasso: the Lasso tends to select only one variable from the group of correlated or nearly linearly dependent variables, even if many or all of these variables belong to the active set S_0. The elastic net (Zou and Hastie, 2005), OSCAR (Bondell and Reich, 2008) and "clustered Lasso" (She, 2010) have been proposed to address this problem but they do not explicitly take correlation-structure among the variables into account and still exhibit difficulties when groups of variables are nearly linearly dependent. A sparse Laplacian shrinkage estimator has been proposed (Huang et al., 2011) and proven to select a correct set of variables under certain regularity conditions. However, the sparse Laplacian shrinkage estimator is geared toward the case where highly correlated variables have similar predictive effects (which we do not require here) and its selection consistency theorem necessarily requires a uniform lower bound for the nonzero signals above an inflated noise level due to model uncertainty.

We take here the point of view that we want to avoid false negatives, i.e., to avoid not selecting an active variable from S_0: the price to pay for this is an increase in false positive selections. From a practical point of view, it can be very useful to have a selection method $\hat{S}$ which includes *all* variables from a group of nearly linearly independent variables where at least one of them is active. Such a procedure is often a good screening method, when measured by $|\hat{S} \cap S_0|/|S_0|$ as a function of $|\hat{S}|$. The desired performance can be achieved by clustering or grouping the variables first and then selecting whole clusters instead of single variables.

1.1. Relation to other work and new contribution

The idea of clustering or grouping variables and then pursuing model fitting is not new, of course. Principal component regression (cf. Kendall, 1957) is among the earliest proposals, and Hastie et al. (2000) have used principal component analysis in order to find a set of highly correlated variables where the clustering can be supervised by a response variable. Tree-harvesting (Hastie et al., 2001) is another proposal which uses supervised learning methods to select groups of predictive variables formed by hierarchical clustering. An algorithmic approach, simultaneously performing the clustering and supervised model fitting, was proposed by Dettling and Bühlmann (2004), and also the OSCAR method (Bondell and Reich, 2008) does such simultaneous grouping and supervised model fitting. A comparison of methods based on grouping variables first and pursuing classification based on clustered features is presented in Tolosi and Lengauer (2011).

Our proposal differs from previous work in various aspects. 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 in (1). 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 (1). We present in Section 2.1 an agglomerative hierarchical clustering method using canonical correlation, and we prove that it finds the finest clustering which satisfies the criterion function that between group canonical correlations are smaller than a threshold and that it is statistically consistent. Furthermore, we prove in Section 4.1 that the construction of groups based on canonical correlations leads to well-posed behavior of the group compatibility constant of the design matrix $\mathbf{X}$ which ensures identifiability and an oracle inequality for the group Lasso (Yuan and Lin, 2006). The latter is a natural choice for estimation and cluster selection; another possibility is to use the Lasso for cluster representatives. We analyze both of these methods: this represents a difference to earlier work where at the time, such high-dimensional estimation techniques have been less or not established at all.

We present some supporting theory in Section 4, describing circumstances where clustering and subsequent estimation improves over standard Lasso without clustering, and the derivations also show the limitations of such an approach. This sheds light for what kind of models and scenarios the commonly used two-stage approach in practice, consisting of clustering variables first and subsequent estimation, is beneficial. Among the favorable scenarios which we will examine for the latter approach are: (i) high within cluster correlation and weak between cluster correlation with potentially many active variables per cluster; and (ii) at most one active variable per cluster where the clusters are tight (high within correlation) but not necessarily assuming low between cluster correlation. Numerical results which complement the theoretical results are presented in Section 5.

2. Clustering covariables

Consider the index set $\{1,\ldots,p\}$ for the covariables in (1). In the sequel, we denote by $x^{(j)}$ the jth component of a vector x and by $X^{(G)} = \{X^{(j)};\ j \in G\}$ the group of variables from a cluster $G \subseteq \{1,\ldots,p\}$. The goal is to find a partition $\mathcal{G}$ into disjoint clusters $G_1,\ldots,G_q$: $\mathcal{G} = \{G_1,\ldots,G_q\}$ with $\bigcup_{r=1}^{q} G_r = \{1,\ldots,p\}$ and $G_r \cap G_\ell = \emptyset$ $(r \neq \ell)$. The partition $\mathcal{G}$ should then satisfy certain criteria.

We propose two methods for clustering the variables, i.e., finding a suitable partition: one is a novel approach based on canonical correlations while the other uses standard correlation based hierarchical clustering.

2.1. Clustering using canonical correlations

For a partition $\mathcal{G}=\{G_1,\ldots,G_q\}$ as above, we consider

$$\hat{\rho}_{\max}(\mathcal{G})=\max\{\hat{\rho}_{\mathrm{can}}(G_r,G_\ell);\ r,\ell\in\{1,\ldots,q\},\ r\neq\ell\}.$$

Here, $\hat{\rho}_{\mathrm{can}}(G_r,G_\ell)$ denotes the empirical canonical correlation (cf. Anderson, 1984) between the variables from $X^{(G_r)}$ and $X^{(G_\ell)}$. (The empirical canonical correlation is always nonnegative). A clustering with $\tau-$separation between clusters is defined as

$$\hat{\mathcal{G}}(\tau)=\text{a partition } \hat{\mathcal{G}} \text{ of } \{1,\ldots,p\} \text{ such that } \hat{\rho}_{\max}(\hat{\mathcal{G}})\le\tau\ (0<\tau<1). \tag{2}$$

Not all values of τ are feasible: if τ is too small, then there is no partition which would satisfy (2). For this reason, we define the canonical correlation of all the variables with the empty set of variables as zero: hence, the trivial partition $\mathcal{G}_{\mathrm{single}}$ consisting of the single cluster $\{1,\ldots,p\}$ has $\hat{\rho}_{\max}(\mathcal{G}_{\mathrm{single}})=0$ which satisfies (2). The fact that τ may not be feasible (except with $\mathcal{G}_{\mathrm{single}}$) can be understood from the view point that coarse partitions do not necessarily lead to smaller values of $\hat{\rho}_{\max}$: for example, when $p\gg n$ and if $\mathrm{rank}(X^{(G_r\cup G_\ell)})=n$, which would typically happen if $|G_r\cup G_\ell|>n$, then $\hat{\rho}_{\mathrm{can}}(G_r,G_\ell)=1$. In general, clustering with $\tau-$separation does not have a unique solution. For example, if $\hat{\mathcal{G}}(\tau)=\{G_1,\ldots,G_q\}$ is a clustering with $\tau-$separation and $\{G_{r;k},k=1,\ldots,q_r\}$ is a nontrivial partition of G_r with $\max_{1\le k_1<k_2\le q_r}\hat{\rho}_{\mathrm{can}}(G_{r;k_1},G_{r;k_2})\le\tau$, then $\{G_1,\ldots,G_{r-1},G_{r;k},k=1,\ldots,q_r,G_{r+1},\ldots,G_q\}$ is a strictly finer clustering with $\tau-$separation, see also Lemma A.1 below. The nonuniqueness of clustering with $\tau-$separation motivates the following definition of the finest clustering with $\tau-$separation. A clustering with $\tau-$separation between clusters, say $\hat{\mathcal{G}}(\tau)$, is finest if every other clustering with $\tau-$separation is strictly coarser than $\hat{\mathcal{G}}(\tau)$, and we denote such a finest clustering with $\tau-$separation by

$$\hat{\mathcal{G}}_{\mathrm{finest}}(\tau).$$

The existence and uniqueness of the finest clustering with $\tau-$separation are provided in Theorem 2.1 below.

A simple hierarchical bottom-up agglomerative clustering (without the need to define linkage between clusters) can be used as an estimator $\hat{\mathcal{G}}(\tau)$ which satisfies (2) and which is finest: the procedure is described in Algorithm 1.

Algorithm 1. Bottom-up, agglomerative hierarchical clustering using canonical correlations.

1: Start with the single variables as p clusters (nodes at the bottom of a tree). Set $b=0$
2: **repeat**
3: Increase b by one. Merge the two clusters having highest canonical correlation.
4: **until** Criterion (2) is satisfied.

Theorem 2.1. *The hierarchical bottom-up agglomerative clustering Algorithm 1 leads to a partition $\hat{\mathcal{G}}(\tau)$ which satisfies (2). If τ is not feasible with a nontrivial partition, the solution is the coarsest partition $\hat{\mathcal{G}}(\tau)=\mathcal{G}_{\mathrm{single}}=\{G\}$ with $G=\{1,\ldots,p\}$.*

Furthermore, if τ is feasible with a nontrivial partition, the solution $\hat{\mathcal{G}}(\tau)=\hat{\mathcal{G}}_{\mathrm{finest}}(\tau)$ is the finest clustering with $\tau-$separation.

A proof is given in Appendix A. Theorem 2.1 describes that a bottom-up greedy strategy leads to an optimal solution.

We now present consistency of the clustering Algorithm 1. Denote the population canonical correlation between $X^{(G_r)}$ and $X^{(G_\ell)}$ by $\rho_{\mathrm{can}}(G_r,G_\ell)$, and the maximum population canonical correlation by $\max_{r\neq\ell}\rho_{\mathrm{can}}(G_r,G_\ell)$, respectively, for a partition $\mathcal{G}=\{G_1,\ldots,G_q\}$ of $\{1,\ldots,p\}$. As in (2), a partition $\mathcal{G}(\tau)$ is a population clustering with $\tau-$separation if $\rho_{\max}(\mathcal{G})\le\tau$. The finest population clustering with $\tau-$separation, denoted by $\mathcal{G}_{\mathrm{finest}}(\tau)$, is the one which is finer than any other population clustering with $\tau-$separation. With the convention $\rho_{\max}(\mathcal{G}_{\mathrm{single}})=0$, the existence and uniqueness of the finest population clustering with $\tau-$separation follows from Theorem 2.1. In fact, the hierarchical bottom-up agglomerative clustering Algorithm 1 yields the finest population clustering $\mathcal{G}_{\mathrm{finest}}(\tau)$ with $\tau-$separation if the population canonical correlation is available and used in the algorithm.

Let $\mathcal{G}_{\mathrm{finest}}(\tau)$ be the finest population clustering with $\tau-$separation and $\hat{\mathcal{G}}(\tau)$ be the sample clustering with $\tau-$separation generated by the hierarchical bottom-up agglomerative clustering Algorithm 1 based on the design matrix $\mathbf{X}$. The following theorem provides a sufficient condition for the consistency of $\hat{\mathcal{G}}(\tau)$ in the Gaussian model

$$X_1,\ldots,X_n \text{ i.i.d. } \sim\mathcal{N}_p(0,\Sigma). \tag{3}$$

For any given $t>0$ and positive integers q and $d_1,\ldots,d_q$, define

$$t_r=\sqrt{d_r/n}+\sqrt{(2/n)(t+\log(q(q+1))},\quad \Delta^*_{r,\ell}=\frac{3(t_r\wedge t_\ell)+(t_r\vee t_\ell)}{(1-t_r)(1-t_\ell)}. \tag{4}$$

Theorem 2.2. *Consider* $\mathbf{X}$ *from (3) and* $\mathcal{G}^0=\{G_1,\ldots,G_q\}$ *a partition of* $\{1,\ldots,p\}$. *Let* $t>0$ *and* $d_r=\mathrm{rank}(\Sigma_{G_r,G_r})$. *Define* $\Delta^*_{r,\ell}$ *by (4). Suppose*

$$\max_{1\le r<\ell\le q}\{\rho_{\mathrm{can}}(G_r,G_\ell)+\Delta^*_{r,\ell}\}\le\tau_-\le\tau_+<\min_{1\le r\le q}\min_{\{G_{r;k}\}}\Big\{\max_{k_1<k_2}\rho_{\mathrm{can}}(G_{r;k_1},G_{r;k_2})-\Delta^*_{r,r}\Big\}, \tag{5}$$

where $\min_{\{G_{r;k}\}}$ *is taken over all nontrivial partitions* $\{G_{r;k}, k \le q_r\}$ *of* G_r. *Then,* $\mathcal{G}^0 = \mathcal{G}_{\text{finest}}(\tau)$ *is the finest population clustering with* τ*−separation for all* $\tau_- \le \tau \le \tau_+$, *and*

$$\mathbb{P}[\hat{\mathcal{G}}(\tau) = \mathcal{G}_{\text{finest}}(\tau),\ \forall \tau_- \le \tau \le \tau_+] \ge 1 - \exp(-t).$$

A proof is given in Appendix A. We note that $t = \sqrt{\log(p)}$ leads to $t_r \approx \sqrt{\text{rank}(\Sigma_{G_r,G_r})/n} + \sqrt{\log(p)/n}$: this is small if $\text{rank}(\Sigma_{G_r,G_r}) = o(n)$ and $\log(p) = o(n)$, and then, $\Delta^*_{r,\ell}$ is small as well (which means that the probability bound becomes $1 - p^{-1} \to 1$ $(p \to \infty)$ or $p \ge n \to \infty$).

The parameter τ in (2) needs to be chosen. We advocate the use of the minimal resulting τ. This can be easily implemented: we run the bottom-up agglomerative clustering Algorithm 1 and record in every iteration the maximal canonical correlation between clusters, denoted by $\hat{\rho}_{\max}(b)$ where b is the iteration number. We then use the partition corresponding to the iteration

$$\hat{b} = \operatorname*{argmin}_b \hat{\rho}_{\max}(b). \tag{6}$$

A typical path of $\hat{\rho}_{\max}(b)$ as a function of the iterations b, the choice $\hat{b}$ and the corresponding minimal $\hat{\rho}_{\max}(\hat{b})$ is shown in Fig. 1.

We conclude that the hierarchical bottom-up agglomerative clustering Algorithm 1 with the rule in (6) is fully data-driven, and there is no need to define linkage between clusters.

2.2. Ordinary hierarchical clustering

As an alternative to the clustering method in Section 2.1, we consider in Section 5.2 ordinary hierarchical agglomerative clustering based on the dissimilarity matrix D with entries $D_{r,\ell} = 1 - |\hat{\rho}(X^{(r)}, X^{(\ell)})|$, where $\hat{\rho}(X^{(r)}, X^{(\ell)})$ denotes the sample correlation between $X^{(r)}$ and $X^{(\ell)}$. We choose average-linkage as dissimilarity between clusters.

As a cutoff for determining the number of clusters, we proceed according to an established principle. In every clustering iteration b, proceeding in an agglomerative way, we record the new value h_b of the corresponding linkage function from the merged clusters (in iteration b): we then use the partition corresponding to the iteration

$$\hat{b} = \operatorname*{argmax}_b (h_{b+1} - h_b).$$

3. Supervised selection of clusters

From the design matrix $\mathbf{X}$, we infer the clusters $G_1, \ldots, G_q$ as described in Section 2. We select the variables in the linear model (1) in a groupwise fashion where all variables from a cluster G_r $(r = 1, \ldots, q)$ are selected or not: this is denoted by

$$\hat{S}_{\text{cluster}} = \{r;\ \text{cluster } G_r \text{ is selected},\ r = 1, \ldots, q\}.$$

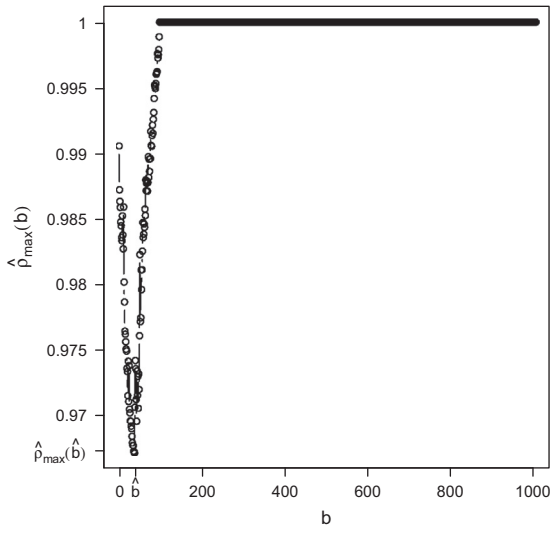

Fig. 1. Path of $\hat{\rho}_{\max}(b)$ as a function of the iterations b: from real data described in Section 5.2 with $p = 1000$ and $n = 71$.

The selected set of variables is then the union of the selected clusters:

$$\hat{S} = \bigcup_{r \in \hat{S}_{\text{cluster}}} G_r. \tag{7}$$

We propose two methods for selecting the clusters, i.e., two different estimators $\hat{S}_{\text{cluster}}$.

3.1. Cluster representative Lasso (CRL)

For each cluster we consider the representative

$$\overline{\mathbf{X}}^{(r)} = \frac{1}{|G_r|} \sum_{j \in G_r} \mathbf{X}^{(j)}, \quad r = 1, \ldots, q,$$

where $\mathbf{X}^{(j)}$ denotes the jth $n \times 1$ column-vector of $\mathbf{X}$. Denote by $\overline{\mathbf{X}}$ the $n \times q$ design matrix whose rth column is given by $\overline{\mathbf{X}}^{(r)}$. We use the Lasso based on the response $\mathbf{Y}$ and the design matrix $\overline{\mathbf{X}}$:

$$\hat{\beta}_{\text{CRL}} = \arg\min_{\beta} (\|\mathbf{Y} - \overline{\mathbf{X}}\beta\|_2^2/n + \lambda_{\text{CRL}}\|\beta\|_1).$$

The selected clusters are then given by

$$\hat{S}_{\text{cluster,CRL}} = \hat{S}_{\text{cluster,CRL}}(\lambda_{\text{CRL}}) = \{r; \hat{\beta}_{\text{CRL},r}(\lambda_{\text{CRL}}) \neq 0,\ r = 1, \ldots, q\},$$

and the selected variables are obtained as in (7), denoted as

$$\hat{S}_{\text{CRL}} = \hat{S}_{\text{CRL}}(\lambda_{\text{CRL}}) = \bigcup_{r \in \hat{S}_{\text{cluster,CRL}}} G_r.$$

3.2. Cluster group Lasso (CGL)

Another obvious way to select clusters is given by the group Lasso. We partition the vector of regression coefficients according to the clusters: $\beta = (\beta_{G_1}, \ldots, \beta_{G_q})^T$, where $\beta_{G_r} = (\{\beta_j;\ j \in G_r\})^T$. The cluster group Lasso is defined as

$$\hat{\beta}_{\text{CGL}} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda_{\text{CGL}} \sum_{r=1}^{q} w_r \|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2 n^{-1/2}, \tag{8}$$

where w_r is a multiplier, typical pre-specified as $w_r = \sqrt{|G_r|}$. It is well known that the group Lasso enjoys a group selection property where either $\hat{\beta}_{\text{CGL},G_r} \neq 0$ (all components are different from zero) or $\hat{\beta}_{\text{CGL},G_r} \equiv 0$ (the zero-vector). We note that the estimator in (8) is different from using the usual penalty $\lambda \sum_{r=1}^{q} w_r \|\beta_{G_r}\|_2$: the penalty in (8) is termed as "groupwise prediction penalty" in Bühlmann and van de Geer (2011, Section 4.5.1): it has nice parameterization invariance properties, and it is a much more appropriate penalty when $\mathbf{X}^{(G_r)}$ exhibits strongly correlated columns.

The selected clusters are then given by

$$\hat{S}_{\text{cluster,CGL}} = \hat{S}_{\text{cluster,CGL}}(\lambda_{\text{CGL}}) = \{r; \hat{\beta}_{\text{CGL},G_r}(\lambda_{\text{CGL}}) \neq 0, r = 1, \ldots, q\},$$

and the selected variables are as in (7):

$$\hat{S}_{\text{CGL}} = \hat{S}_{\text{CGL}}(\lambda_{\text{CGL}}) = \bigcup_{r \in \hat{S}_{\text{cluster,CGL}}} G_r = \{j; \hat{\beta}_{\text{CGL},j} \neq 0, j = 1, \ldots, p\},$$

where the latter equality follows from the group selection property of the group Lasso.

4. Theoretical results for cluster Lasso methods

We provide here some supporting theory, first for the cluster group Lasso (CGL) and then for the cluster representative Lasso (CRL).

4.1. Cluster group Lasso (CGL)

We will show first that the compatibility constant of the design matrix $\mathbf{X}$ is well-behaved if the canonical correlation between groups is small, i.e., a situation which the clustering Algorithm 1 is exactly aiming for.

The CGL method is based on the whole design matrix $\mathbf{X}$, and we can write the model in group structure form where we denote by $\mathbf{X}^{(G_r)}$ the $n \times |G_r|$ design matrix with variables $\{\mathbf{X}^{(j)};\ j \in G_r\}$:

$$\mathbf{Y} = \mathbf{X}\beta^0 + \boldsymbol{\varepsilon} = \sum_{r=1}^{q} \mathbf{X}^{(G_r)}\beta^0_{G_r} + \boldsymbol{\varepsilon}, \tag{9}$$

where $\beta^0_{G_r} = (\{\beta^0_j; j \in G_r\})^T$. We denote by $S_{0,\text{Group}} = \{r; \beta^0_{G_r} \neq 0, r = 1, \ldots, q\}$.

We introduce now a few other notations and definitions. We let $m_r := |G_r|$, $r = 1, \ldots, q$, denote the group sizes, and define the average group size

$$\overline{m} := \frac{1}{q} \sum_{r=1}^{q} m_r,$$

and the average size of active groups

$$\overline{m}_{S_{0,\mathrm{Group}}} := \frac{1}{|S_{0,\mathrm{Group}}|} \sum_{r \in S_{0,\mathrm{Group}}} m_r.$$

Furthermore, for any $S \subseteq \{1, \ldots, q\}$, we let $\mathbf{X}^{(S)}$ be the design matrix containing the variables in $\bigcup_{r \in S} G_r$. Moreover, define

$$\|\beta_S\|_{2,1} := \sum_{r \in S} \|\mathbf{X}^{(G_r)} \beta_{G_r}\|_2 \sqrt{m_r/\overline{m}}.$$

We denote in this section by

$$\hat{\Sigma}_{r,\ell} := (\mathbf{X}^{(G_r)})^T \mathbf{X}^{(G_\ell)}/n, \quad r, \ell \in \{1, \ldots, q\}.$$

We assume that each $\hat{\Sigma}_{r,r}$ is nonsingular (otherwise one may use generalized inverses) and we write, assuming here for notational clarity that $S_{0,\mathrm{Group}} = \{1, \ldots, s_0\}$ ($s_0 = |S_{0,\mathrm{Group}}|$) is the set of indices of the first s_0 groups:

$$\hat{R}_{S_{0,\mathrm{Group}}} := \begin{pmatrix} I & \hat{\Sigma}_{1,1}^{-1/2} \hat{\Sigma}_{1,2} \hat{\Sigma}_{2,2}^{-1/2} & \cdots & \hat{\Sigma}_{1,1}^{-1/2} \hat{\Sigma}_{1,s_0} \hat{\Sigma}_{s_0,s_0}^{-1/2} \\ \hat{\Sigma}_{2,2}^{-1/2} \hat{\Sigma}_{2,1} \hat{\Sigma}_{1,1}^{-1/2} & I & \cdots & \hat{\Sigma}_{2,2}^{-1/2} \hat{\Sigma}_{2,s_0} \hat{\Sigma}_{s_0,s_0}^{-1/2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}_{s_0,s_0}^{-1/2} \hat{\Sigma}_{s_0,1} \hat{\Sigma}_{1,1}^{-1/2} & \hat{\Sigma}_{s_0,s_0}^{-1/2} \hat{\Sigma}_{s_0,2} \hat{\Sigma}_{2,2}^{-1/2} & \cdots & I \end{pmatrix}.$$

The group compatibility constant given in Bühlmann and van de Geer (2011) is a value $\phi^2_{0,\mathrm{Group}}(\mathbf{X})$ that satisfies

$$\phi^2_{0,\mathrm{Group}}(\mathbf{X}) \le \min\left\{ \frac{\overline{m}_{S_0} |S_{0,\mathrm{Group}}|}{\overline{m}} \frac{\|\mathbf{X}\beta\|_2^2}{\|\beta_{S_{0,\mathrm{Group}}}\|_{2,1}^2};\ \|\beta_{S^c_{0,\mathrm{Group}}}\|_{2,1} \le 3 \|\beta_{S_{0,\mathrm{Group}}}\|_{2,1} \right\}.$$

The constant 3 here is related to the condition $\lambda \ge 2\lambda_0$ in Proposition 4.1 below: in general it can be taken as $(\lambda + \lambda_0)/(\lambda - \lambda_0)$.

Theorem 4.1. *Suppose that $\hat{R}_{S_{0,\mathrm{Group}}}$ has smallest eigenvalue $\hat{\Lambda}^2_{\min} > 0$. Assume moreover the incoherence conditions*:

$$\rho := \max_{r \in S_0, \ell \in S_0^c} \frac{\overline{m}}{\sqrt{m_r m_\ell}} \hat{\rho}_{\mathrm{can}}(G_r, G_\ell) \le C \frac{\hat{\Lambda}^2_{\min} \overline{m}/\overline{m}_{S_{0,\mathrm{Group}}}}{3|S_{0,\mathrm{Group}}|} \quad \text{for some } 0 < C < 1,$$

$$\rho_{S_{0,\mathrm{Group}}} := \max_{r,\ell \in S_{0,\mathrm{Group}},\ r \ne \ell} \frac{\overline{m}}{\sqrt{m_r m_\ell}} \hat{\rho}_{\mathrm{can}}(G_r, G_\ell) < \frac{1}{|S_{0,\mathrm{Group}}|}.$$

Then, the group Lasso compatibility holds with compatibility constant

$$\phi^2_{0,\mathrm{Group}}(\mathbf{X}) \ge (\hat{\Lambda}^2_{\min} \overline{m}/\overline{m}_{S_{0,\mathrm{Group}}} - 3|S_{0,\mathrm{Group}}|\rho)^2 / (\hat{\Lambda}^2_{\min} \overline{m}^2/\overline{m}^2_{S_{0,\mathrm{Group}}}) \ge (1-C)^2 \hat{\Lambda}^2_{\min} \ge (1-C)^2 (1 - |S_{0,\mathrm{Group}}| \rho_{S_{0,\mathrm{Group}}}) > 0.$$

A proof is given in Appendix A. We note that small enough canonical correlations between groups ensure the incoherence assumptions for ρ and $\rho_{S_{0,\mathrm{Group}}}$, and in turn that the group Lasso compatibility condition holds. The canonical correlation based clustering Algorithm 1 is tailored for this situation.

Remark 1. One may in fact prove a group version of Corollary 7.2 in Bühlmann and van de Geer (2011) which says that under the eigenvalue and incoherence condition of Theorem 4.1, a group irrepresentable condition holds. This in turn implies that, with large probability, the group Lasso will have no false positive selection of groups, i.e., one has a group version of the result of Problem 7.5 in Bühlmann and van de Geer (2011).

Known theoretical results can then be applied for proving an oracle inequality of the group Lasso.

Proposition 4.1. *Consider model (1) with fixed design* $\mathbf{X}$ *and Gaussian error* $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 I)$. *Let*

$$\lambda_0 = \sigma \frac{2}{\sqrt{n}} \sqrt{1 + \sqrt{\frac{4t + 4\log(p)}{m_{\min}}} + \frac{4t + 4\log(p)}{m_{\min}}},$$

where $m_{\min} = \min_{r=1,\ldots,q}|m_r|$. *Assume that* $\hat{\Sigma}_{r,r}$ *is nonsingular for all* $r = 1,\ldots,q$. *Then, for the group Lasso estimator* $\hat{\beta}(\lambda)$ *in* (8) *with* $\lambda \geq 2\lambda_0$, *and with probability at least* $1-\exp(-t)$:

$$\|\mathbf{X}(\hat{\beta}(\lambda)-\beta^0)\|_2^2/n + \lambda \sum_{r=1}^{q} \sqrt{m_r}\|\hat{\beta}_{G_r}(\lambda)-\beta^0_{G_r}\|_2 \leq 24\lambda^2 \sum_{r\in S_{0,\text{Group}}} m_r/\phi^2_{0,\text{Group}}(\mathbf{X}), \tag{10}$$

where $\phi^2_{0,\text{Group}}(\mathbf{X})$ *denotes the group compatibility constant.*

Proof. We can invoke the result in Bühlmann and van de Geer (2011, Theorem 8.1): using the groupwise prediction penalty in (8) leads to an equivalent formalization where we can normalize to $\hat{\Sigma}_{r,r} = I_{m_r\times m_r}$. The requirement $\lambda \geq 4\lambda_0$ in Bühlmann and van de Geer (2011, Therorem 8.1) can be relaxed to $\lambda \geq 2\lambda_0$ since the model (1) is assumed to be true, see also Bühlmann and van de Geer (2011, p. 108, Section 6.2.3). □

4.2. Linear dimension reduction and subsequent Lasso estimation

For a mean zero Gaussian random variable $Y\in\mathbb{R}$ and a mean zero Gaussian random vector $X\in\mathbb{R}^p$, we can always use a random design Gaussian linear model representation:

$$\begin{aligned} &Y = \sum_{j=1}^{p} \beta_j^0 X^{(j)} + \varepsilon, \\ &X\sim\mathcal{N}_p(0,\Sigma),\ \varepsilon\sim\mathcal{N}(0,\sigma^2), \end{aligned} \tag{11}$$

where ε is independent of X.

Consider a linear dimension reduction

$$Z = A_{q\times p}X$$

using a matrix $A_{q\times p}$ with $q < p$. We denote in the sequel by

$$\mu_X = \mathbb{E}[Y|X] = \sum_{j=1}^{p} \beta_j^0 X^{(j)}, \quad \mu_Z = \mathbb{E}[Y|Z] = \sum_{r=1}^{q} \gamma_r^0 Z^{(r)}.$$

Of particular interest is $Z = (\overline{X}^{(1)},\ldots,\overline{X}^{(q)})^T$, corresponding to the cluster representatives $\overline{X}^{(r)} = |G_r|^{-1}\sum_{j\in G_r} X^{(j)}$. Due to the Gaussian assumption in (11), we can always represent

$$Y = \sum_{r=1}^{q} \gamma_r^0 Z^{(r)} + \eta = \mu_Z + \eta, \tag{12}$$

where η is independent of Z. Furthermore, since μ_Z is the linear projection of Y on the linear span of Z and also the linear projection of μ_X on the linear span of Z,

$$\xi^2 = \text{Var}(\eta) = \text{Var}(\varepsilon + \mu_X - \mu_Z) = \sigma^2 + \mathbb{E}[(\mu_X-\mu_Z)^2].$$

For the prediction error, when using the dimension-reduced Z and for any estimator $\hat{\gamma}$, we have, using that $\mu_X-\mu_Z$ is orthogonal to the linear span of Z:

$$\mathbb{E}[(X^T\beta^0 - Z^T\hat{\gamma})^2] = \mathbb{E}[(Z^T\gamma^0 - Z^T\hat{\gamma})^2] + \mathbb{E}[(\mu_X-\mu_Z)^2]. \tag{13}$$

Thus, the total prediction error consists of an error due to estimation of γ^0 and a squared bias term $B^2 = \mathbb{E}[(\mu_X-\mu_Z)^2]$. The latter has already appeared in the variance $\xi^2 = \text{Var}(\eta)$ above.

Let $\mathbf{Y}, \mathbf{X}$ and $\mathbf{Z}$ be n i.i.d. realizations of the variables Y, X and Z, respectively. Then,

$$\mathbf{Y} = \mathbf{X}\beta^0 + \boldsymbol{\varepsilon} = \mathbf{Z}\gamma^0 + \boldsymbol{\eta}.$$

Consider the Lasso, applied to $\mathbf{Y}$ and $\mathbf{Z}$, for estimating γ^0:

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma}(\|\mathbf{Y}-\mathbf{Z}\gamma\|_2^2/n + \lambda\|\gamma\|_1).$$

The cluster representative Lasso (CRL) is a special case with $n\times q$ design matrix $\mathbf{Z} = \overline{\mathbf{X}}$.

Proposition 4.2. *Consider* n *i.i.d. realizations from the model* (11). *Let*

$$\lambda_0 = 2\|\hat{\sigma}_{\mathbf{Z}}\|_\infty \xi \sqrt{\frac{t^2 + 2\log(q)}{n}},$$

where $\|\hat{\sigma}_{\mathbf{Z}}\|_\infty = \max_{r=1,\ldots,q}\sqrt{(n^{-1}\mathbf{Z}^T\mathbf{Z})_{rr}}$, *and* $\xi^2 = \sigma^2 + \mathbb{E}[(\mu_X-\mu_Z)^2]$. *Then, for* $\lambda \geq 2\lambda_0$ *and with probability at least* $1-2\exp(-t^2/2)$, *conditional on* $\mathbf{Z}$:

$$\|\mathbf{Z}(\hat{\gamma}-\gamma^0)\|_2^2/n + \lambda\|\hat{\gamma}-\gamma^0\|_1 \leq 4\lambda^2 s(\gamma^0)/\phi_0^2(\mathbf{Z}),$$

where $s(\gamma^0)$ *equals the number of nonzero coefficients in* γ^0 *and* $\phi_0^2(\mathbf{Z})$ *denotes the compatibility constant of the design matrix* $\mathbf{Z}$ (Bühlmann and van de Geer, 2011, (6.4)).

A proof is given in Appendix A. The choice $\lambda = 2\lambda_0$ leads to the convergence rates:

$$\|\mathbf{Z}(\hat{\gamma}(\lambda)-\gamma^0)\|_2^2/n = O_P\left(\frac{s(\gamma^0)\log(q)}{n\phi_0^2(\mathbf{Z})}\right),$$
$$\|\hat{\gamma}(\lambda)-\gamma^0\|_1 = O_P\left(\frac{s(\gamma^0)}{\phi_0^2(\mathbf{Z})}\sqrt{\frac{\log(q)}{n}}\right). \quad (14)$$

The second result about the ℓ_1−norm convergence rate implies a variable screening property as follows: assume a so-called beta-min condition requiring that

$$\min_{r\in S(\gamma^0)}|\gamma_r^0| \geq Cs(\gamma^0)\sqrt{\frac{\log(q)}{n}}/\phi_0^2(\mathbf{Z}) \quad (15)$$

for a sufficiently large $C>0$, then, with high probability, $\hat{S}_{\mathbf{Y},\mathbf{Z}}(\lambda)\supseteq S(\gamma^0)$, where $\hat{S}_{\mathbf{Y},\mathbf{Z}}(\lambda)=\{r;\hat{\gamma}_r(\lambda)\neq 0\}$ and $S(\gamma^0)=\{r;\ \gamma_r^0\neq 0\}$.

The results in Proposition 4.2, (14) and (15) describe inference for $\mathbf{Z}\gamma^0$ and for γ^0. Their meaning for inferring $\mathbf{X}\beta^0$ and for (groups) of β^0 are further discussed in Section 4.3 for specific examples, representing γ^0 in terms of β^0 and the correlation structure of X, and analyzing the squared bias $B^2=\mathbb{E}[(\mu_X-\mu_Z)^2]$. The compatibility constant $\phi_0^2(\mathbf{Z})$ in Proposition 4.2 is typically (much) better behaved, if $q\ll p$, than the corresponding constant $\phi_0^2(\mathbf{X})$ of the original design $\mathbf{X}$. Bounds of $\phi_0^2(\mathbf{X})$ and $\phi_0^2(\mathbf{Z})$ in terms of their population covariance Σ and $A\Sigma A^T$, respectively, can be derived from Bühlmann and van de Geer (2011, Corollary 6.8). Thus, loosely speaking, we have to deal with a trade-off: the term $\phi_0^2(\mathbf{Z})$, coupled with a $\log(q)$−instead of a $\log(p)$−factor (and also to a certain extent the sparsity factor $s(\gamma^0)$) are favorable for the dimensionality reduced $\mathbf{Z}$. The price to pay for this is the bias term $B^2=\mathbb{E}[(\mu_X-\mu_Z)^2]$, discussed further in Section 4.3, which appears in the variance ξ^2 entering the definition of λ_0 in Proposition 4.2 as well as in the prediction error (13); furthermore, the detection of γ^0 instead of β^0 can be favorable for some cases and not favorable for others, as discussed in Section 4.3.

Finally, note that Proposition 4.2 makes a statement conditional on $\mathbf{Z}$: with high probability $1-\alpha$,

$$\mathbb{P}[\|\mathbf{Z}\hat{\gamma}-\mathbf{Z}\gamma^0\|_2^2/n \leq 4\lambda^2 s(\gamma^0)/\phi_0^2(\mathbf{Z})|\mathbf{Z}]\geq 1-\alpha.$$

Assuming that $\phi_0^2(\mathbf{Z})\geq\phi_0^2(A,\Sigma)$ is bounded with high probability (Bühlmann and van de Geer, 2011, Lemma 6.17), we obtain (for a small but different α than above):

$$\mathbb{P}[\|\mathbf{Z}\hat{\gamma}-\mathbf{Z}\gamma^0\|_2^2/n \leq 4\lambda^2 s(\gamma^0)/\phi_0^2(A,\Sigma)]\geq 1-\alpha.$$

In view of (13), we then have for the prediction error:

$$\mathbb{E}[\|\mathbf{X}\beta^0-\mathbf{Z}\hat{\gamma}\|_2^2/n] = \mathbb{E}[\|\mathbf{Z}\hat{\gamma}-\mathbf{Z}\gamma^0\|_2^2/n + \mathbb{E}[(\mu_X-\mu_Z)^2], \quad (16)$$

where $\mathbb{E}[\|\mathbf{Z}\hat{\gamma}-\mathbf{Z}\gamma^0\|_2^2/n]\approx O(\xi s(\gamma^0)\sqrt{\log(p)/n})$ when choosing $\lambda=2\lambda_0$.

4.3. The parameter γ^0 for cluster representatives

In the sequel, we consider the case where $Z=\overline{X}=(\overline{X}^{(1)},\ldots,\overline{X}^{(q)})^T$ encodes the cluster representatives $\overline{X}^{(r)}=|G_r|^{-1}\sum_{j\in G_r}X^{(j)}$. We analyze the coefficient vector γ^0 and discuss it from the view-point of detection. For specific examples, we also quantify the squared bias term $B^2=\mathbb{E}[(\mu_X-\mu_{\overline{X}})^2]$ which plays mainly a role for prediction.

The coefficient γ^0 can be exactly described if the cluster representatives are independent.

Proposition 4.3. *Consider a random design Gaussian linear model as in (11) where* $\mathrm{Cov}(\overline{X}^{(r)},\overline{X}^{(\ell)})=0$ *for all* $r\neq\ell$. *Then,*

$$\gamma_r^0 = |G_r|\sum_{j\in G_r} w_j\beta_j^0, \quad r=1,\ldots,q,$$
$$w_j = \frac{\sum_{k=1}^p \Sigma_{j,k}}{\sum_{\ell\in G_r}\sum_{k=1}^p \Sigma_{\ell,k}},\ \sum_{j\in G_r} w_j = 1.$$

Moreover:

1. *If, for* $r\in\{1,\ldots,q\}$, $\Sigma_{j,k}\geq 0$ *for all* $j,k\in G_r$, *then* $w_j\geq 0$ $(j\in G_r)$, *and* $\gamma_r^0/|G_r|$ *is a convex combination of* β_j^0 $(j\in G_r)$. *In particular, if* $\beta_j^0\geq 0$ *for all* $j\in G_r$, *or* $\beta_j^0\leq 0$ *for all* $j\in G_r$, *then*

$$|\gamma_r^0|\geq |G_r|\min_{j\in G_r}|\beta_j^0|$$

2. *If, for* $r\in\{1,\ldots,q\}$, $\Sigma_{j,j}\equiv 1$ *for all* $j\in G_r$ *and* $\sum_{k\neq j}\Sigma_{j,k}\equiv\zeta$ *for all* $j\in G_r$, *then* $w_j\equiv|G_r|^{-1}$ *and*

$$\gamma_r^0 = \sum_{j\in G_r}\beta_j^0.$$

A concrete example is where Σ has a block-diagonal structure with equi-correlation within blocks: $\Sigma_{j,k}\equiv\rho_r$ $(j,k\in G_r,\ j\neq k)$ *with* $-1/(|G_r|-1)<\rho_r<1$ (*where the lower bound for* ρ_r *ensures positive definiteness of the block-matrix* Σ_{G_r,G_r}).

A proof is given in Appendix A. The assumption of uncorrelatedness across $\{\overline{X}^{(r)};\ r=1,\ldots,q\}$ is reasonable if we have tight clusters corresponding to blocks of a block-diagonal structure of Σ.

We can immediately see that there are benefits or disadvantages of using the group representatives in terms of the size of the absolute value $|\gamma_r^0|$: obviously a large value would make it easier to detect the group G_r. Taking a group representative is advantageous if all the coefficients within a group have the same sign. However, we should be a bit careful since the size of a regression coefficient should be placed in context to the standard deviation of the regressor: here, the standardized coefficients are

$$\gamma_r^0\sqrt{\mathrm{Var}(\overline{X}^{(r)})}. \tag{17}$$

For e.g. high positive correlation among variables within a group, $\mathrm{Var}(\overline{X}^{(r)})$ is much larger than for independent variables: for the equi-correlation scenario in statement 2 of Proposition 4.3 we obtain for the standardized coefficient

$$\gamma_r^0\sqrt{\mathrm{Var}(\overline{X}^{(r)})}=\sum_{j\in G_r}\beta_j^0\sqrt{\rho+|G_r|^{-1}(1-\rho)}\approx\sum_{j\in G_r}\beta_j^0, \tag{18}$$

where the latter approximation holds if $\rho\approx 1$.

The disadvantages occur if rough or near cancellation among β_j^0 $(j\in G_r)$ takes place. This can cause a reduction of the absolute value of $|\gamma_r^0|$ in comparison to $\max_{j\in G_r}|\beta_j^0|$: again, the scenario in statement 2 of Proposition 4.3 is most clear in the sense that the sum of β_j^0 $(j\in G_r)$ is equal to γ_r^0, and near cancellation would mean that $\sum_{j\in G_r}\beta_j^0\approx 0$.

An extension of Proposition 4.3 can be derived for covering the case where the regressors $\{\overline{X}^{(r)};\ r=1,\ldots,q\}$ are only approximately uncorrelated.

Proposition 4.4. *Assume the conditions of Proposition 4.3 but instead of uncorrelatedness of* $\{\overline{X}^{(r)};\ r=1,\ldots,q\}$ *across r, we require*: *for* $r\in\{1,\ldots,q\}$,

$$|\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(\ell)};\ \ell\neq r\})|\le\nu\quad \text{for all } j\in G_r,\ i\notin G_r.$$

Moreover, assume that $\mathrm{Var}(\overline{X}^{(r)}|\{\overline{X}^{(\ell)};\ \ell\neq r\})\ge C>0$. *Then,*

$$\gamma_r^0=|G_r|\sum_{j\in G_r}w_j\beta_j^0+\Delta_r,\ |\Delta_r|\le\nu\|\beta^0\|_1/C.$$

Furthermore, if $\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(\ell)};\ell\neq r\})\ge 0$ *for all* $j\in G_r,\ i\notin G_r$, *and* $\mathrm{Var}(\overline{X}^{(r)}|\{\overline{X}^{(k)};\ k\neq r\})\ge C>0$, *then*

$$|\gamma_r^0|\ge|G_r|\min_{j\in G_r}|\beta_j^0|+\Delta_r,\ |\Delta_r|\le\nu\|\beta^0\|_1/C.$$

A proof is given in Appendix A. The assumption that $|\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(k)};\ k\neq r\})|\le\nu$ for all $j\in G_r$, $i\notin G_r$ is implied if the variables in G_r and G_ℓ $(r\neq\ell)$ are rather uncorrelated. Furthermore, if we require that $\nu\|\beta^0\|_1\ll|G_r|\min_{j\in G_r}|\beta_j^0|$ and $C\approx 1$ (which holds if $\Sigma_{j,j}\equiv 1$ for all $j\in G_r$ and if the variables within G_r have high conditional correlations given $\{\overline{X}^{(\ell)};\ell\neq r\}$), then

$$|\gamma_r^0|\ge|G_r|\min_{j\in G_r}|\beta_j^0|(1+o(1)).$$

Thus, also under clustering with only moderate independence between the clusters, we can have beneficial behavior for the representative cluster method. This also implies that the representative cluster method works if the clustering is only approximately correct, as shown in Section 4.6.

We discuss in the next subsections two examples in more detail.

4.3.1. Block structure with equi-correlation

Consider a partition with groups G_r $(r=1,\ldots,q)$. The population covariance matrix Σ is block-diagonal having $|G_r|\times|G_r|$ block-matrices Σ_{G_r,G_r} with equi-correlation: $\Sigma_{j,j}\equiv 1$ for all $j\in G_r$, and $\Sigma_{j,k}\equiv\rho_r$ for all $j,k\in G_r$ $(j\neq k)$, where $-1/(|G_r|-1)<\rho_r<1$ (the lower bound for ρ_r ensures positive definiteness of Σ_{G_r,G_r}). This is exactly the setting as in statement 2 of Proposition 4.3.

The parameter γ^0 equals, see Proposition 4.3:

$$\gamma_r^0=\sum_{j\in G_r}\beta_j^0.$$

Regarding the bias, observe that

$$\mu_X-\mu_{\overline{X}}=\sum_{r=1}^{q}(\mu_{X;r}-\mu_{\overline{X};r}),$$

where $\mu_{X;r}=\sum_{j\in G_r}\beta_j^0 X^{(j)}$ and $\mu_{\overline{X};r}=\overline{X}^{(r)}\gamma_r^0$, and due to block-independence of X, we have

$$\mathbb{E}[(\mu_X-\mu_{\overline{X}})^2]=\sum_{r=1}^{q}\mathbb{E}[(\mu_{X;r}-\mu_{\overline{X};r})^2].$$

For each summand, we have $\mu_{X;r}-\mu_{\overline{X};r}=\sum_{j\in G_r}X^{(j)}(\beta_j^0-\overline{\beta}_r^0)$, where $\overline{\beta}_r^0=|G_r|^{-1}\sum_{j\in G_r}\beta_j^0$. Thus,

$$\mathbb{E}[(\mu_{X;r}-\mu_{\overline{X};r})^2]=\sum_{j\in G_r}(\beta_j^0-\overline{\beta}_r^0)^2+2\rho_r\sum_{j,k\in G_r;j\neq k}(\beta_j^0-\overline{\beta}_r^0)(\beta_k^0-\overline{\beta}_r^0)=(1-\rho_r)\sum_{j\in G_r}(\beta_j^0-\overline{\beta}_r^0)^2.$$

Therefore, the squared bias equals

$$B^2=\mathbb{E}[(\mu_{X;r}-\mu_{\overline{X};r})^2]=\sum_{r=1}^{q}(1-\rho_r)\sum_{j\in G_r}(\beta_j^0-\overline{\beta}_r^0)^2. \tag{19}$$

We see from the formula that the bias is small if there is little variation of β^0 within the groups G_r or/and the ρ_r's are close to one. The latter is what we obtain with tight clusters: there is a large within and small between groups correlation. Somewhat surprising is the fact that the bias is becoming large if ρ_r tends to negative values (which is related to the fact that detection becomes bad for negative values of ρ_r, see (18)).

Thus, in summary, in comparison to an estimator based on $\mathbf{X}$, using the cluster representatives $\overline{\mathbf{X}}$ and subsequent estimation leads to equal (or better, due to smaller dimension) prediction error if all ρ_r's are close to 1, regardless of β^0. When using the cluster representative Lasso, if $B^2=\sum_{r=1}^{q}(1-\rho_r)\sum_{j\in G_r}(\beta_j^0-\overline{\beta}_r^0)^2=O(s(\gamma^0)\log(q)/n)$, then the squared bias has no disturbing effect on the prediction error as can be seen from (14).

With respect to detection, there can be a substantial gain for inferring the cluster G_r if β_j^0 ($j\in G_r$) have all the same sign, and if ρ_r is close to 1. Consider the active groups

$$S_{0,\mathrm{Group}}=\{r;\ \beta_{G_r}^0\neq 0,\ r=1,\ldots,q\}.$$

For the current model and assuming that $\sum_{j\in G_r}\beta_j^0\neq 0$ ($r=1,\ldots,q$) (i.e., no exact cancellation of coefficients within groups), $S_{0,\mathrm{Group}}=S(\gamma^0)=\{r;\ \gamma_r^0\neq 0\}$. In view of (18) and (15), the screening property for groups holds if

$$\min_{r\in S_{0,\mathrm{Group}}}\left|\sum_{j\in G_r}\beta_j^0\sqrt{\rho+|G_r|^{-1}(1-\rho)}\right|\geq C\frac{s(\gamma^0)}{\phi_0^2(\overline{\mathbf{X}})}\sqrt{\frac{\log(q)}{n}}.$$

This condition holds even if the nonzero β_j^0's are very small but their sum within a group is sufficiently strong.

4.3.2. One active variable per cluster

We design here a model with at most one active variable in each group. Consider a low-dimensional $q\times 1$ variable U and perturbed versions of $U^{(r)}$ ($r=1,\ldots,q$) which constitute the $p\times 1$ variable X:

$$\begin{aligned}&X^{(r,1)}=U^{(r)},\quad r=1,\ldots,q,\\&X^{(r,j)}=U^{(r)}+\delta^{(r,j)},\quad j=2,\ldots,m_r,\ \ r=1,\ldots,q,\\&\delta^{(r,2)},\ldots,\delta^{(r,m_r)}\ \text{i.i.d. }\mathcal{N}(0,\tau_r^2)\ \text{and independent among } r=1,\ldots,q.\end{aligned} \tag{20}$$

The index $j=1$ has no specific meaning, and the fact that one covariate is not perturbed is discussed in Remark 2 below. The purpose is to have at most one active variable in every cluster by assuming a low-dimensional underlying linear model

$$Y=\sum_{r=1}^{q}\tilde{\beta}_r^0U^{(r)}+\varepsilon, \tag{21}$$

where $\varepsilon\sim\mathcal{N}(0,\sigma^2)$ is independent of U. Some of the coefficients in $\tilde{\beta}^0$ might be zero, and hence, some of the $U^{(r)}$'s might be noise covariates. We construct a $p\times 1$ variable X by stacking the variables $X^{(r,j)}$ as follows: $X^{(\sum_{\ell=1}^{r-1}m_\ell+j)}=X^{(r,j)}$ ($j=1,\ldots,m_r$) for $r=1,\ldots,q$. Furthermore, we use an augmented vector of the true regression coefficients

$$\beta_j^0=\begin{cases}\tilde{\beta}_r^0 & \text{if } j=\sum_{\ell=1}^{r-1}m_\ell+1,\ r=1,\ldots,q,\\ 0 & \text{otherwise.}\end{cases}$$

Thus, the model in (21) can be represented as

$$Y=\sum_{j=1}^{p}\beta_j^0X^{(j)}+\varepsilon,$$

where ε is independent of X.

Remark 2. Instead of the model in (20)–(21), we could consider

$$\begin{aligned}&X^{(r,j)}=U^{(r)}+\delta^{(r,j)},\quad j=1,\ldots,m_r,\ \ r=1,\ldots,q,\\&\delta^{(r,1)},\ldots,\delta^{(r,m_r)}\ \text{i.i.d. }\mathcal{N}(0,\tau_r^2)\ \text{and independent among } r=1,\ldots,q,\end{aligned}$$

and

$$Y = \sum_{j=1}^{p} \beta_j^0 X^{(j)} + \varepsilon,$$

where ε is independent of X. Stacking the variables $X^{(r,j)}$ as before, the covariance matrix $\mathrm{Cov}(X)$ has q equi-correlation blocks but the blocks are generally dependent if $U^{(1)}, \ldots, U^{(q)}$ are correlated, i.e., $\mathrm{Cov}(X)$ is not of block-diagonal form. If $\mathrm{Cov}(U^{(r)}, U^{(\ell)}) = 0$ $(r \neq \ell)$, we are back to the model in Section 4.3.1. For more general covariance structures of U, an analysis of the model seems rather cumbersome (but see Proposition 4.4) while the analysis of the "asymmetric" model (20)–(21) remains rather simple as discussed below.

We assume that the clusters G_r are corresponding to the variables $\{X^{(r,j)};\ j = 1, \ldots, m_r\}$ and thus, $m_r = |G_r|$. In contrast to the model in Section 4.3.1, we do not assume uncorrelatedness or quantify correlation between clusters. The cluster representatives are

$$\overline{X}^{(r)} = m_r^{-1} \sum_{j \in G_r} X^{(r,j)} = U^{(r)} + W^{(r)},$$
$$W^{(r)} \sim \mathcal{N}\left(0, \tau_r^2 \frac{m_r - 1}{m_r^2}\right) \text{ and independent among } r = 1, \ldots, q.$$

As in (12), the dimension-reduced model is written as $Y = \sum_{r=1}^{q} \gamma_r^0 \overline{X}^{(r)} + \eta$.

For the bias, we immediately find

$$B^2 = \mathbb{E}[(\mu_X - \mu_{\overline{X}})^2] \leq \mathbb{E}[(U^T \tilde{\beta}^0 - \overline{X} \tilde{\beta}^0)^2] = \mathbb{E}|W^T \tilde{\beta}^0|^2 \leq s_0 \max_j |\beta_j^0|^2 \max_r \frac{m_r - 1}{m_r^2} \tau_r^2. \tag{22}$$

Thus, if the cluster sizes m_r are large and/or the perturbation noise τ_r^2 is small, the squared bias B^2 is small.

Regarding detection, we make use of the following result.

Proposition 4.5. *For the model in (21) we have*

$$\|\tilde{\beta}^0 - \gamma^0\|_2^2 \leq 2B^2 / \lambda_{\min}^2(\mathrm{Cov}(U)) = 2\mathbb{E}|W^T \tilde{\beta}^0|^2 / \lambda_{\min}^2(\mathrm{Cov}(U))$$
$$\leq 2 s_0 \max_j |\beta_j^0|^2 \max_r \frac{m_r - 1}{m_r^2} \tau_r^2 / \lambda_{\min}^2(\mathrm{Cov}(U)),$$

where $\lambda_{\min}^2(\mathrm{Cov}(U))$ *denotes the minimal eigenvalue of* $\mathrm{Cov}(U)$.

A proof is given in Appendix A. Denote by $\tilde{S}_0 = \{r; \tilde{\beta}_r^0 \neq 0\}$. We then have

$$\text{if } \min_{r \in \tilde{S}_0} |\tilde{\beta}_r^0| > 2\sqrt{2 s_0 \max_j |\beta_j^0| \max_r \frac{m_r - 1}{m_r^2} \tau_r^2 / \lambda_{\min}^2(\mathrm{Cov}(U))},$$
$$\text{then : } \min_{r \in \tilde{S}_0} |\gamma_r^0| > \sqrt{2 s_0 \max_j |\beta_j^0| \max_r \frac{m_r - 1}{m_r^2} \tau_r^2 / \lambda_{\min}^2(\mathrm{Cov}(U))}. \tag{23}$$

This follows immediately: if the implication would not hold, it would create a contradiction to Proposition 4.5.

We argue now that

$$\max_r \frac{\tau_r^2}{m_r} = O(\log(q)/n) \tag{24}$$

is a sufficient condition to achieve prediction error and detection as in the q-dimensional model (21). Thereby, we implicitly assume that $\max_j |\beta_j^0| \leq C < \infty$ and $\lambda_{\min}^2(\mathrm{Cov}(U)) \geq L > 0$. Since we have that $s(\gamma^0) \geq s_0$ (excluding the pathological case for a particular combination of β^0 and $\Sigma = \mathrm{Cov}(X)$), using (22) the condition (24) implies

$$B^2 = O(s_0 \log(q)/n) \leq O(s(\gamma^0) \log(q)/n)$$

which is at most of the order of the prediction error in (14), where $\phi_0^2(\mathbf{Z}) = \phi_0^2(\overline{\mathbf{X}})$ can be lower-bounded by the population version $\phi_0^2(\mathrm{Cov}(\overline{X})) \geq \phi_0^2(\mathrm{Cov}(U))$ (Bühlmann and van de Geer, 2011, Cor.6.8). For detection, we note that (24) implies that the bound in (23) is at most

$$\min_{r \in \tilde{S}_0} |\gamma_r^0| > \sqrt{2 s_0 \max_j |\beta_j^0| \max_r \frac{m_r - 1}{m_r^2} \tau_r^2 / \lambda_{\min}^2(\mathrm{Cov}(U))} \leq O(s(\gamma^0) \sqrt{\log(q)/n}).$$

The right-hand side is what we require as beta-min condition in (15) for group screening such that with high probability $\hat{S} \supseteq S(\gamma^0) \supseteq S_0$ (again excluding a particular constellation of β^0 and Σ).

The condition (24) itself is fulfilled if $m_r \approx n/\log(q)$, i.e., when the cluster sizes are large, or if $\tau_r^2 = O(\log(q)/n)$, i.e., the clusters are tight. An example where the model (20)–(21) with (24) seems reasonable is for genome-wide association studies with SNPs where $p \approx 10^6$, $n \approx 1000$ and m_r can be in the order of 10^3 and hence $q \approx 1000$ when e.g. using the order of

magnitude of number of target SNPs (Carlson et al., 2004). Note that this is a scenario where the group sizes $m_r \gg n$ where the cluster group Lasso seems inappropriate.

The analysis in Sections 4.2 and 4.3 about the bias B^2 and the parameter γ^0 has immediate implications for the cluster representative Lasso (CRL), as discussed in the next section.

4.4. A comparison

We compare now the results of the cluster representative Lasso (CRL), the cluster group Lasso (CGL) and the plain Lasso, at least on a "rough scale".

For the plain Lasso we have with high probability,

$$\|\mathbf{X}(\hat{\beta}_{\text{Lasso}}-\beta^0)\|_2^2/n = O\left(\frac{\log(p)s_0}{n\phi_0^2(\mathbf{X})}\right),$$
$$\|\hat{\beta}_{\text{Lasso}}-\beta^0\|_1 = O\left(\frac{s_0}{\phi_0^2(\mathbf{X})}\sqrt{\frac{\log(p)}{n}}\right), \tag{25}$$

which both involve $\log(p)$ instead of $\log(q)$ and more importantly, the compatibility constant $\phi_0^2(\mathbf{X})$ of the $\mathbf{X}$ design matrix instead of $\phi_0^2(\overline{\mathbf{X}})$ of the matrix $\overline{\mathbf{X}}$. If p is large, then $\phi_0^2(\mathbf{X})$ might be close to zero; furthermore, it is exactly in situations like model (20) and (21), having a few ($s_0 \le q$) active variables and noise covariates being highly correlated with the active variables, which leads to very small values of $\phi_0^2(\mathbf{X})$, see van de Geer and Lederer (2012). For variable screening $\hat{S}_{\text{Lasso}}(\lambda) \supseteq S_0$ with high probability, the corresponding (sufficient) beta-min condition is

$$\min_{j\in S_0}\left|\beta_j^0\right| \ge Cs_0\sqrt{\frac{\log(p)}{n}}/\phi_0^2(\mathbf{X}) \tag{26}$$

for a sufficiently large $C > 0$.

For comparison with the cluster group Lasso (CGL) method, we assume for simplicity equal group sizes $|G_r| = m_r \equiv m$ for all r and $\log(q)/m \le 1$, i.e., the group size is sufficiently large. We then obtain with high probability,

$$\|\mathbf{X}(\hat{\beta}_{\text{CGL}}-\beta^0)\|_2^2/n = O\left(\frac{|S_{0,\text{Group}}|m}{n\phi_{0,\text{Group}}^2(\mathbf{X})}\right),$$
$$\sum_{r=1}^{q}\|\hat{\beta}_{G_r}-\beta_{G_r}^0\|_2 = O\left(\frac{|S_{0,\text{Group}}|\sqrt{m}}{\sqrt{n}\phi_{0,\text{Group}}^2(\mathbf{X})}\right). \tag{27}$$

For variable screening $\hat{S}_{\text{CGL}}(\lambda) \supseteq S_0$ with high probability, the corresponding (sufficient) beta-min condition is

$$\min_{r\in S_{0,\text{Group}}}\|\beta_{G_r}^0\|_2 \ge C\frac{|S_{0,\text{Group}}|\sqrt{m}}{\sqrt{n}\phi_{0,\text{Group}}^2(\mathbf{X})}$$

for a sufficiently large $C > 0$. The compatibility constants $\phi_0^2(\mathbf{X})$ in (25) and $\phi_{0,\text{Group}}^2(\mathbf{X})$ in (27) are not directly comparable, but see Theorem 4.1 which is in favor of the CGL method. "On the rough scale", we can distinguish two cases: if the group-sizes are large with only none or a few active variables per group, implying $s_0 = |S_0| \approx |S_{0,\text{Group}}|$, the Lasso is better than the CGL method because the CGL rate involves $|S_{0,\text{Group}}|m$ or $|S_{0,\text{Group}}|\sqrt{m}$, respectively, instead of the sparsity s_0 appearing in the rate for the standard Lasso; for the case where we have either none or many active variables within groups, the CGL method is beneficial, mainly for detection, since $|S_{0,\text{Group}}|m \approx |S_0| = s_0$ but $|S_{0,\text{Group}}|\sqrt{m} < |S_0| = s_0$. The behavior in the first case is to be expected since in the group Lasso representation (9), the parameter vectors $\beta_{G_r}^0$ are very sparse within groups, and this sparsity is not exploited by the group Lasso. A sparse group Lasso method (Simon et al., 2013) would address this issue. On the other hand, the CGL method has the advantage that it works without bias, in contrast to the CRL procedure. Furthermore, the CGL can also lead to good detection if many β_j^0's in a group are small in absolute value. For detection of the group G_r, we only need that $\|\beta_{G_r}^0\|_2$ is sufficiently large: the signs of the coefficients of β_{G_r} can be different and (near) cancellation does not happen.

For the cluster representative Lasso (CRL) method, the range of scenarios, with good performance of CRL, is more restricted. The method works well and is superior over the plain Lasso (and group Lasso) if the bias B^2 is small and the detection is well-posed in terms of the dimension-reduced parameter γ^0. More precisely, if

$$B^2 = O\left(\frac{s(\gamma^0)\log(q)}{n\phi_0^2(\overline{\mathbf{X}})}\right),$$

Table 1
Comparison of cluster representative Lasso (CRL) with plain Lasso in terms of prediction and variable screening. The symbol "+" encodes better theoretical results for the CRL in comparison to Lasso; an "NA" means that no comparative statement can be made.

Model	Assumption	Prediction	Screening
Equi-corr. blocks (Section 4.3.1)	Small value in (19)	+	NA
Equi-corr. blocks (Section 4.3.1)	E.g. same sign for β_j^0 $(j \in G_r)$	NA	+
≤ 1 Var. per group (Section 4.3.2)	(24)	+	+

the CRL is better than plain Lasso for prediction since the corresponding oracle inequality for the CRL becomes, see (16): with high probability,

$$\|\overline{\mathbf{X}}\hat{\gamma} - \mathbf{X}\beta^0\|_2^2/n + \lambda\|\hat{\gamma} - \gamma^0\|_1 \leq O\left(\frac{s(\gamma^0)\log(q)}{n\phi_0(\overline{\mathbf{X}})}\right).$$

We have given two examples and conditions ensuring that the bias is small, namely (19) and (24). The latter condition (24) is also sufficient for better screening property in the model from Section 4.3.2. For the equi-correlation model in Section 4.3.1, the success of screening crucially depends on whether the coefficients from active variables in a group nearly cancel or add-up (e.g. when having the same sign), see Propositions 4.3 and 4.4. The following Table 1 recapitulates the findings.

Summarizing, both the CGL and CRL are useful and can be substantially better than plain Lasso in terms of prediction and detection in the presence of highly correlated variables. If the cluster sizes are smaller than sample size, the CGL method is more broadly applicable, in the sense of consistency but not necessarily efficiency, as it does not involve the bias term B^2 and constellation of signs or of near cancellation of coefficients in $\beta_{G_r}^0$ is not an issue. For group sizes which are larger than sample size, the CGL is not appropriate: one would need to take a version of the group Lasso with regularization within groups (Meier et al., 2009; Simon et al., 2013). The CGL method benefits when using canonical correlation based clustering as this improves the compatibility constant, see Theorem 4.1. The CRL method is particularly suited for problems where the variables can be grouped into tight clusters and/or the cluster sizes are large. There is gain if there is at most one active variable per cluster and the clusters are tight, otherwise the prediction performance is influenced by the bias B^2 and detection is depending on whether the coefficients within a group add-up or exhibit near cancellation. If the variables are not very highly correlated within large groups, the difficulty is to estimate these groups, and in case of correct grouping, as assumed in the theoretical results above, the CRL method may still perform (much) better than plain Lasso.

4.5. Estimation of the clusters

The theoretical derivations above assume that the groups G_r $(r = 1, \ldots, q)$ correspond to the correct clusters. For the canonical correlation based clustering as in Algorithm 1, Theorem 2.2 discusses consistency in finding the true underlying population clustering. For hierarchical clustering, the issue is much simpler.

Consider the $n \times p$ design matrix $\mathbf{X}$ as in (3) and assume for simplicity that $\Sigma_{j,j} = 1$ for all j. It is well-known that

$$\max_{j,k}|\hat{\Sigma}_{j,k} - \Sigma_{j,k}| = O_P(\sqrt{\log(p)/n}), \tag{28}$$

where $\hat{\Sigma}$ is the empirical covariance matrix (cf. Bühlmann and van de Geer, 2011, p. 152). Tightness and separation of the true clusters is ensured by

$$\min\{|\Sigma_{j,k}|; j, k \in G_r (j \neq k),\ r = 1, \ldots q\} > \max\{|\Sigma_{j,k}| : j \in G_r, k \in G_\ell, r, \ell = 1, \ldots, q\ (r \neq \ell)\}. \tag{29}$$

Assuming (29) and using (28), a standard clustering algorithm, using e.g. single-linkage and dissimilarity $1 - |\hat{\Sigma}_{j,k}|$ between variables $X^{(j)}$ and $X^{(k)}$, will consistently find the true clusters if $\log(p)/n \to 0$.

In summary, and rather obvious, the higher the correlation within and uncorrelatedness between clusters, the better we can estimate the true underlying grouping. In this sense, and following the arguments in Sections 4.1–4.4, strong correlation within clusters "is a friend" when using cluster Lasso methods, while it is "an enemy" (at least for variable screening and selection) for plain Lasso.

4.6. Some first illustrations

We briefly illustrate some of the points and findings mentioned above for the CRL and the plain Lasso. Throughout this subsection, we show the results from a single realization of each of different models. More systematic simulations are shown in Section 5. We analyze scenarios with $p = 1000$ and $n = 100$. Thereby, the covariates are generated as in model (20) where $U \sim \mathcal{N}_q(0, I)$ with $q = 5$ and $\tau = 0.5$, and thus, $\mathrm{Cov}(X) = \Sigma$ is of block-diagonal structure. The response is as in the linear model (1) with $\varepsilon \sim \mathcal{N}_n(0, I)$. We consider the following.

Correct and incorrect clustering: The correct clustering consists of $q=5$ clusters each having $m_r=|G_r|\equiv 200$ variables, corresponding to the structure in model (20). An incorrect clustering was constructed as 5 clusters where the first half (100) of the variables in each constructed cluster correspond to the first half (100) of the variables in each of the true 5 clusters, and the remaining second half (100) of the variables in the constructed clusters are chosen randomly from the total of 500 remaining variables. We note that $m_r\equiv 200 > n=100$ and thus, the CGL method is inappropriate (see e.g. Proposition 4.1).

Active variables and regression coefficients: We always consider 3 active groups (a group is called active if there is at least one active variables in the group). The scenarios are as follows:

(a) one active variable within each of 3 active groups, namely $S_0=\{1,201,401\}$. The regression coefficients are $\beta_1^0=-1$, $\beta_{201}^0=-1$, $\beta_{401}^0=1$;
(b) 4 active variables within each of 3 active groups, namely $S_0=\{1,2,3,4,201,202,203,204,401,402,403,404\}$. The regression coefficients are $\beta_j^0\equiv 0.25$ for $j\in S_0$;
(c) as in (b) but with regression coefficients $\{\beta_j^0;\ j\in S_0\}$ i.i.d.$\sim$Unif$([-0.5,0.5]$;
(d) as in (b) but with exact cancellation of coefficients: $\beta_1^0=\beta_3^0=2$, $\beta_2^0=\beta_4^0=-2$, $\beta_{201}^0=\beta_{203}^0=2$, $\beta_{202}^0=\beta_{204}^0=-2$, $\beta_{401}^0=\beta_{403}^0=2$, $\beta_{402}^0=\beta_{404}^0=-2$.

For the scenario in (d), we had to choose large coefficients, in absolute value equal to 2, in order to see clear differences, in favor of the plain Lasso. Fig. 2 using the R-package `glmnet` (Friedman et al., 2010) shows the results.

The results seem rather robust against approximate cancellation of coefficients (Fig. 2(c)) and incorrect clustering (right panels in Fig. 2). Regarding the latter, the number of chosen clusters is worse than for correct clustering, though. A main message of the results in Fig. 2 is that the predictive performance (using cross-validation) is a good indicator whether the group representative Lasso (with correct or incorrect clustering) works.

We can complement the rules from Section 2 for determining the number of clusters as follows: take the representative cluster Lasso method with the largest clusters (the least refined partition of $\{1,\ldots,p\}$) such that predictive performance is still reasonable (in comparison to the best achieved performance where we would always consider plain Lasso among the competitors as well). In the extreme case of Fig. 2(d), this rule would choose the plain Lasso (among the alternatives of correct clustering and incorrect clustering) which is indeed the least refined partition such that predictive performance is still reasonable.

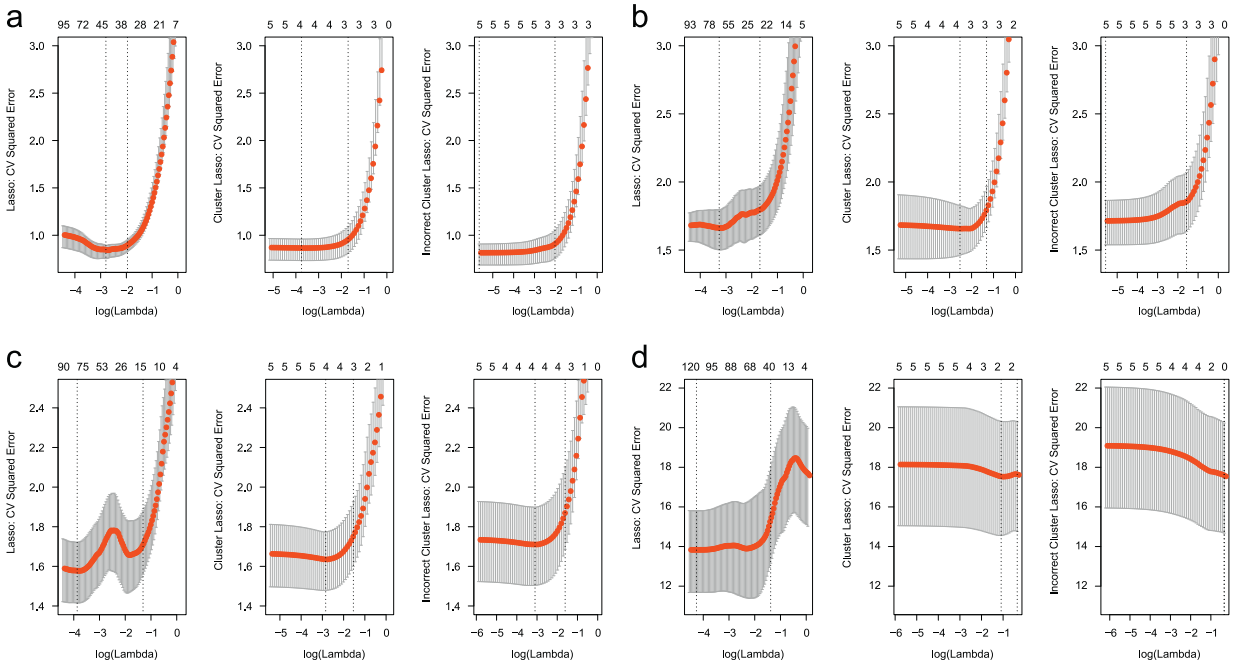

Fig. 2. Left, middle and right panel of each subfigure: Lasso, cluster representative Lasso with correct clustering, and cluster representative Lasso with incorrect clustering as described in Section 4.6: 10-fold CV squared error (y-axis) versus $\log(\lambda)$ (x-axis). Grey bars indicate the region 10-fold CV squared error $\pm$ estimated standard error (s.e.) of 10-fold CV squared error. The left vertical bar indicates the minimizer of the CV error and the right vertical bar corresponds to the largest value such that the CV error is within one standard deviation (s.d.) of the minimum. The numbers on top of each plot report the number of selected variables (in case of the cluster representative Lasso, the number of selected representatives): the number of active groups is always equal to 3, and the number of active variables is 3 for (a) and 12 for (b)–(d). Subfigures (a)–(d) correspond to the scenarios (a)–(d) in Section 4.6.

5. Numerical results

In this section we look at three different simulation settings and a pseudo real data example in order to empirically compare the proposed cluster Lasso methods with plain Lasso.

5.1. Simulated data

Here, we only report results for the CRL and CGL methods where the clustering of the variables is based on canonical correlations using Algorithm 1 (see Section 2.1). The corresponding results using ordinary hierarchical clustering, based on correlations and with average-linkage (see Section 2.2), are almost exactly the same because for the considered simulation settings, both clustering methods produce essentially the same partition.

We simulate data from the linear model in (1) with fixed design $\mathbf{X}$, $\boldsymbol{\varepsilon}\sim\mathcal{N}_n(0,\sigma^2 I)$ with $n=100$ and $p=1000$. We generate the fixed design matrix $\mathbf{X}$ once, from a multivariate normal distribution $\mathcal{N}_p(0,\Sigma)$ with different structures for Σ, and we then keep it fixed. We consider various scenarios, but the sparsity or size of the active set is always equal to $s_0=20$.

In order to compare the various methods, we look at two performance measures for prediction and variable screening. For each model, our simulated data consists of a training and an independent test set. The models were fitted on the training data and we computed the test set mean squared error $n^{-1}\sum_{i=1}^{n}\mathbb{E}[(Y_{\text{test}}(X_i)-\hat{Y}(X_i))^2]$, denoted by MSE. For variable screening we consider the true positive rate as a measure of performance, i.e., $|\hat{S}\cap S_0|/|S_0|$ as a function of $|\hat{S}|$.

For each of the methods we choose a suitable grid of values for the tuning parameter. All reported results are based on 50 simulation runs.

5.1.1. Block diagonal model

Consider a block diagonal model where we simulate the covariables $X\sim\mathcal{N}_p(0,\Sigma_A)$ where Σ_A is a block diagonal matrix. We use a 10×10 matrix Γ, where

$$\Gamma_{i,j}=\begin{cases}1, & i=j,\\ 0.9 & \text{else.}\end{cases}$$

The block-diagonal of Σ_A consists of 100 such block matrices Γ. Regarding the regression parameter β^0, we consider the following configurations:

(Aa) $S_0=\{1,2,\ldots,20\}$ and for any $j\in S_0$ we sample β_j^0 from the set $\{2/s_0,4/s_0,\ldots,2\}$ without replacement (anew in each simulation run).
(Ab) $S_0=\{1,2,11,12,21,22,\ldots,91,92\}$ and for any $j\in S_0$ we sample β_j^0 from the set $\{2/s_0,4/s_0,\ldots,2\}$ without replacement (anew in each simulation run).
(Ac) β^0 as in (Aa) but we switch the sign of half and randomly chosen active parameters (anew in each simulation run).
(Ad) β^0 as in (Ab) but we switch the sign of half and randomly chosen active parameters (anew in each simulation run).

The set-up (Aa) has all the active variables in the first two blocks of highly correlated variables. In the second configuration (Ab), the first two variables of each of the first ten blocks are active. Thus, in (Aa), half of the active variables appear in the same block while in the other case (Ab), the active variables are distributed among 10 blocks. The remaining two configurations (Ac) and (Ad) are modifications in terms of random sign changes. The models (Ab) and (Ad) come closest to the model (20)–(21) considered for theoretical purposes: the difference is that the former models have two active variables per active block (or group) while the latter model has only one active variable per active group.

Table 2 and Fig. 3 describe the simulation results. From Table 2 we see that over all the configurations, the CGL method has lower predictive performance than the other two methods. Comparing the two methods Lasso and CRL, we cannot distinguish a clear difference with respect to prediction. We also find that sign switches of half of the active variables (Ac,Ad) do not have a negative effect on the predictive performance of the CRL method (which in principle could suffer severely from sign switches). The CGL method even gains in predictive performance in (Ac) and (Ad) compared to the no sign-switch configurations (Aa) and (Ab).

Table 2
MSE for the block diagonal model with standard deviations in brackets.

σ	Method	(Aa)	(Ab)	(Ac)	(Ad)
3	CRL	10.78 (1.61)	15.57 (2.43)	13.08 (1.65)	15.39 (2.35)
	CGL	14.97 (2.40)	37.05 (5.21)	13.34 (2.06)	24.31 (6.50)
	Lasso	11.94 (1.97)	16.23 (2.47)	12.72 (1.67)	15.34 (2.53)
12	CRL	161.73 (25.74)	177.90 (25.87)	157.86 (20.63)	165.30 (23.56)
	CGL	206.19 (29.97)	186.61 (25.69)	160.31 (23.04)	168.26 (24.70)
	Lasso	168.53 (25.88)	179.47 (25.77)	158.02 (20.31)	166.50 (23.74)

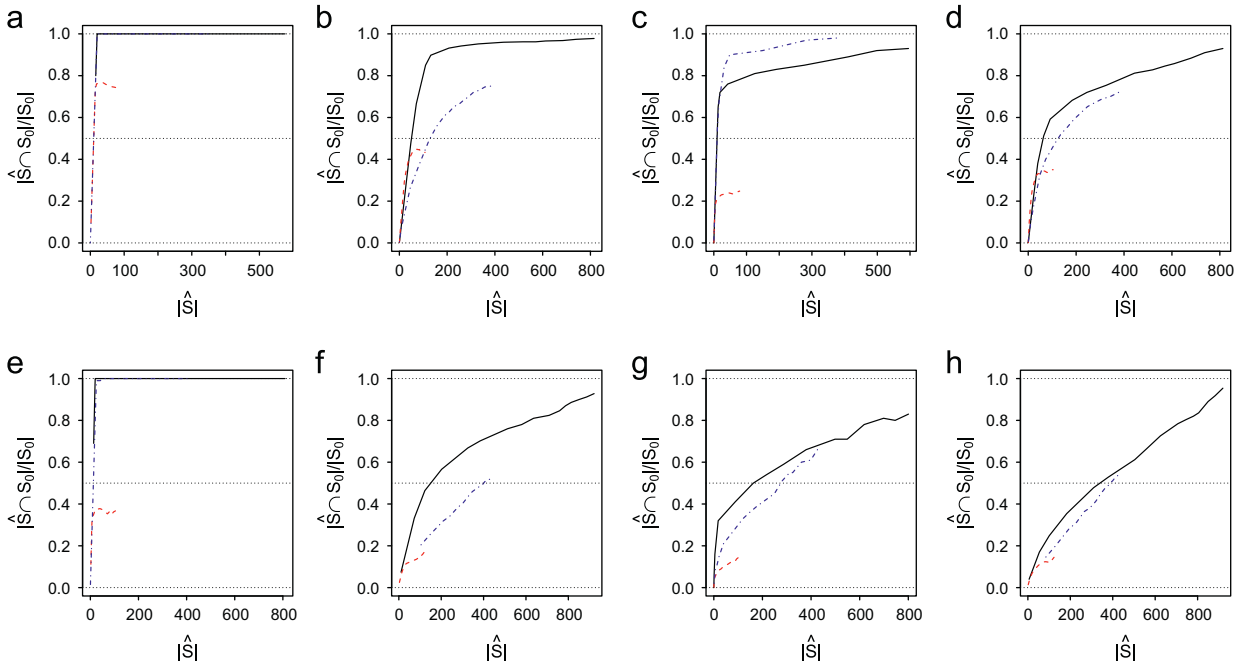

Fig. 3. Plot of $|\hat{S}\cap S_0|/|\hat{S}|$ versus $|\hat{S}|$ for block diagonal model. Cluster representative Lasso (CRL, black solid line), cluster group Lasso (CGL, blue dashed-dotted line), and Lasso (red dashed line). (a) (Aa), $\sigma=3$. (b) (Ab), $\sigma=3$. (c) (Ac), $\sigma=3$. (d) (Ad), $\sigma=3$. (e) (Aa), $\sigma=12$. (f) (Ab), $\sigma=12$. (g) (Ac), $\sigma=12$. (h) (Ad), $\sigma=12$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

From Fig. 3 we infer that for the block diagonal simulation model the two methods CRL and CGL outperform the Lasso concerning variable screening. Taking a closer look, the Cluster Lasso methods CRL and CGL benefit more when having a lot of active variables in a cluster as in settings (Aa) and (Ac).

5.1.2. *Single block model*

We simulate the covariables $X\sim\mathcal{N}_p(0,\Sigma_B)$ where

$$\Sigma_{B;i,j}=\begin{cases}1, & i=j,\\ 0.9, & i,j\in\{1,\ldots,30\}\text{ and }i\neq j,\\ 0, & \text{else.}\end{cases}$$

Such a Σ_B corresponds to a single group of strongly correlated variables of size 30. The rest of the 970 variables are uncorrelated. For the regression parameter β^0 we consider the following configurations:

(Ba) $S_0=\{1,2,\ldots,15\}\cup\{31,32,\ldots,35\}$ and for any $j\in S_0$ we sample β_j^0 from the set $\{2/s_0,4/s_0,\ldots,2\}$ without replacement (anew in each simulation run).
(Bb) $S_0=\{1,2,\ldots,5\}\cup\{31,32,\ldots,45\}$ and for any $j\in S_0$ we sample β_j^0 from the set $\{2/s_0,4/s_0,\ldots,2\}$ without replacement (anew in each simulation run).
(Bc) β^0 as in (Ba) but we switch the sign of half and randomly chosen active parameters (anew in each simulation run).
(Bd) β^0 as in (Bb) but we switch the sign of half and randomly chosen active parameters (anew in each simulation run).

In the fist set-up (Ba), a major fraction of the active variables are in the same block of highly correlated variables. In the second scenario (Bb), most of the active variables are distributed among the independent variables. The remaining two configurations (Bc) and (Bd) are modifications in terms of random sign changes. The results are described in Table 3 and Fig. 4.

In Table 3 we see that over all the configurations the CRL method performs as well as the Lasso, and both of them outperform the CGL. We again find that the CGL method gains in predictive performance when the signs of the coefficient vector are not the same everywhere, and this benefit is more pronounced when compared to the block diagonal model.

The plots in Fig. 4 for variable screening show that the CRL method performs better than the Lasso for all of the configurations. The CGL method is clearly inferior to the Lasso especially in the setting (Ba) where the CGL seems to have severe problems in finding the true active variables.

Table 3
MSE for the single block model with standard deviations in brackets.

σ	Method	(Ba)	(Bb)	(Bc)	(Bd)
3	CRL	16.73 (2.55)	27.91 (4.80)	15.49 (2.93)	22.17 (4.47)
	CGL	247.52 (28.74)	54.73 (10.59)	21.37 (9.51)	31.58 (14.17)
	Lasso	17.13 (3.01)	27.18 (4.51)	15.02 (2.74)	21.91 (4.48)
12	CRL	173.89 (23.69)	181.62 (24.24)	161.01 (23.19)	175.49 (23.61)
	CGL	384.78 (48.26)	191.26 (25.55)	159.40 (23.88)	174.49 (25.40)
	Lasso	173.37 (23.23)	178.86 (23.80)	160.55 (22.80)	174.14 (23.14)

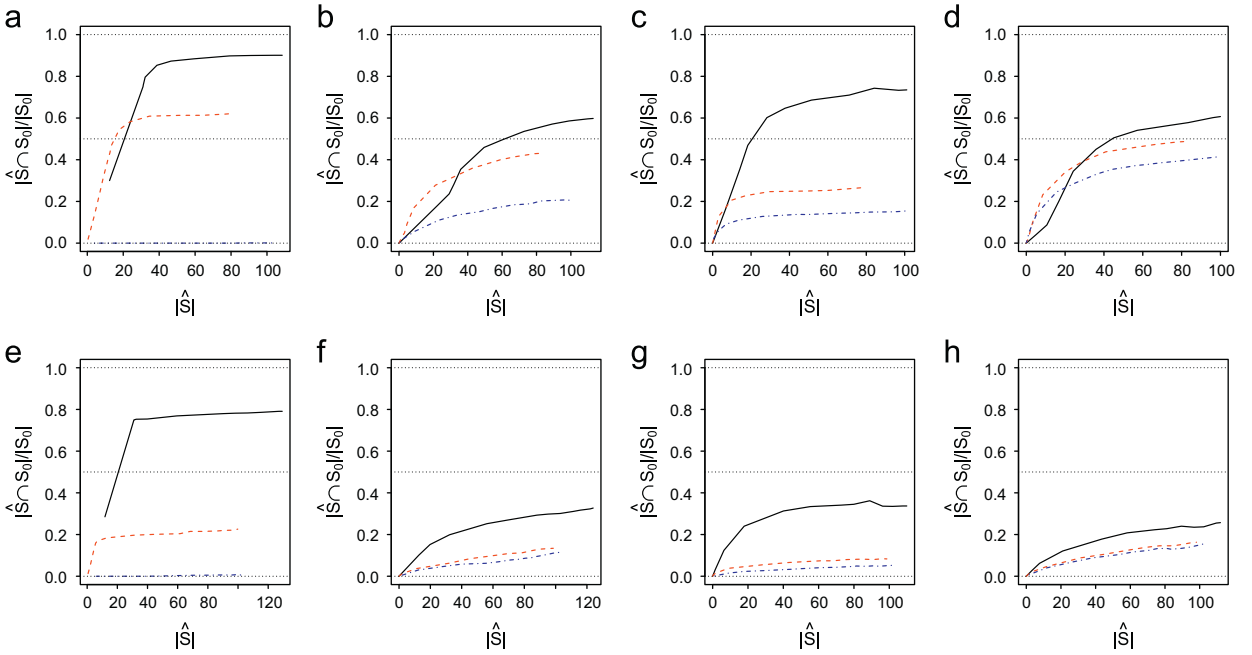

Fig. 4. Plot of $|\hat{S}\cap S_0|/|\hat{S}|$ versus $|\hat{S}|$ for single block model. Cluster representative Lasso (CRL, black solid line), cluster group Lasso (CGL, blue dashed-dotted line), and Lasso (red dashed line). (a) (Ba), $\sigma=3$. (b) (Bb), $\sigma=3$. (c) (Bc), $\sigma=3$. (d) (Bd), $\sigma=3$. (e) (Ba), $\sigma=12$. (f) (Bb), $\sigma=12$. (g) (Bc), $\sigma=12$. (h) (Bd), $\sigma=12$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

5.1.3. *Duo block model*

We simulate the covariables X according to $\mathcal{N}_p(0,\Sigma_C)$ where Σ_C is a block diagonal matrix. We use the 2×2 block matrices

$$\Gamma=\begin{bmatrix}1 & 0.9\\ 0.9 & 1\end{bmatrix},$$

and the block diagonal of Σ_C consists now of 500 such block matrices Γ. In this setting we only look at one set-up for the parameter β^0:

(C) $S_0=\{1,\ldots,20\}$ with

$$\beta_j^0=\begin{cases}2, & j\in\{1,3,5,7,9,11,13,15,17,19\},\\ \frac{1}{3}\sqrt{\frac{\log p}{n}}\sigma, & j\in\{2,4,6,8,10,12,14,16,18,20\}.\end{cases}$$

The idea of choosing the parameters in this way is given by the fact that the Lasso would typically not select the variables from $\{2,4,6,\ldots,20\}$ but selecting the other from $\{1,3,5,\ldots,19\}$. Table 4 and Fig. 5 show the simulation results for the duo block model.

Table 4
MSE for duo block model with standard deviations in brackets.

σ	Method	(C)
3	CRL	22.45 (4.26)
	CGL	32.00 (6.50)
	Lasso	22.45 (4.64)
12	CRL	190.93 (25.45)
	CGL	193.97 (27.05)
	Lasso	190.91 (25.64)

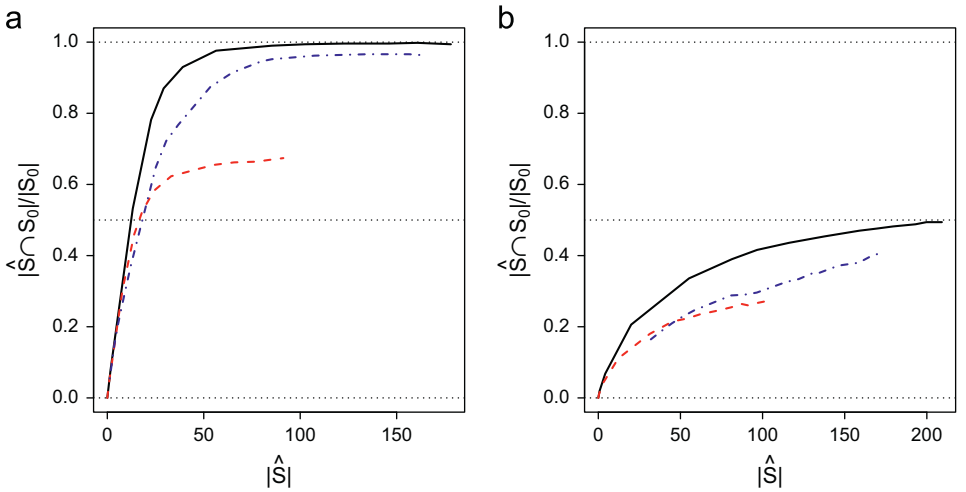

Fig. 5. Plot of $|\hat{S} \cap S_0|/|\hat{S}|$ versus $|\hat{S}|$ for duo block model. Cluster representative Lasso (CRL, black solid line), cluster group Lasso (CGL, blue dashed-dotted line), and Lasso (red dashed line). (a) (C), $\sigma = 3$. (b) (C), $\sigma = 12$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

From Table 4 we infer that for the duo block model, all three estimation methods have a similar prediction performance. Especially for $\sigma = 12$ we see no difference between the methods. But in terms of variable screening, we see in Fig. 5 that the two techniques CRL and CGL are clearly better than the Lasso.

5.2. Pseudo-real data

For the pseudo real data example described below, we also consider the CRL method with ordinary hierarchical clustering as detailed in Section 2.2. We denote the method by CRLcor.

We consider here an example with real data design matrix $\mathbf{X}$ but synthetic regression coefficients β^0 and simulated Gaussian errors $\mathcal{N}_n(0, \sigma^2 I)$ in a linear model as in (1). For the real data design matrix $\mathbf{X}$ we consider a data set about riboflavin (vitamin B2) production by bacillus subtilis. That data has been provided by DSM (Switzerland) and can be obtained from Bühlmann et al. (in press). The covariates are measurements of the logarithmic expression level of 4088 genes (and the response variable $\mathbf{Y}$ is the logarithm of the riboflavin production rate, but we do not use it here). The data consists of $n = 71$ samples of genetically engineered mutants of bacillus subtilis. There are different strains of bacillus subtilis which are cultured under different fermentation conditions, which make the population rather heterogeneous.

We reduce the dimension to $p = 1000$ covariates which have largest empirical variances and choose the size of the active set as $s_0 = 10$.

(D1) S_0 is chosen as a randomly selected variable k and the nine covariates which have highest absolute correlation to variable k (anew in each simulation run). For each $j \in S_0$ we use $\beta_j^0 = 1$.

(D2) S_0 is chosen as one randomly selected entry in each of the five biggest clusters of both clustering methods (using either Algorithm 1 or hierarchical clustering as described in Section 2.2) resulting in $s_0 = 10$ (anew in each simulation run). For each $j \in S_0$ we use $\beta_j^0 = 1$.

The results are given in Table 5 and Fig. 6, based on 50 independent simulation runs.

Table 5 shows that we do not really gain any predictive power when using the proposed cluster Lasso methods CRL or CGL: this finding is consistent with the reported results for simulated data in Section 5.1. The method CRLcor, using standard

Table 5
Prediction error for the pseudo real riboflavin data with standard deviations in brackets.

σ	Method	(D1)	(D2)
3	CRL	2.47 (0.94)	2.99 (0.72)
	CGL	2.36 (0.93)	3.13 (0.74)
	Lasso	2.47 (0.94)	2.96 (0.60)
	CRLcor	39.02 (25.15)	7.08 (2.76)
12	CRL	19.62 (10.11)	14.80 (4.91)
	CGL	17.49 (9.28)	14.90 (5.44)
	Lasso	19.63 (10.00)	15.66 (4.84)
	CRLcor	50.40 (27.68)	15.46 (5.74)

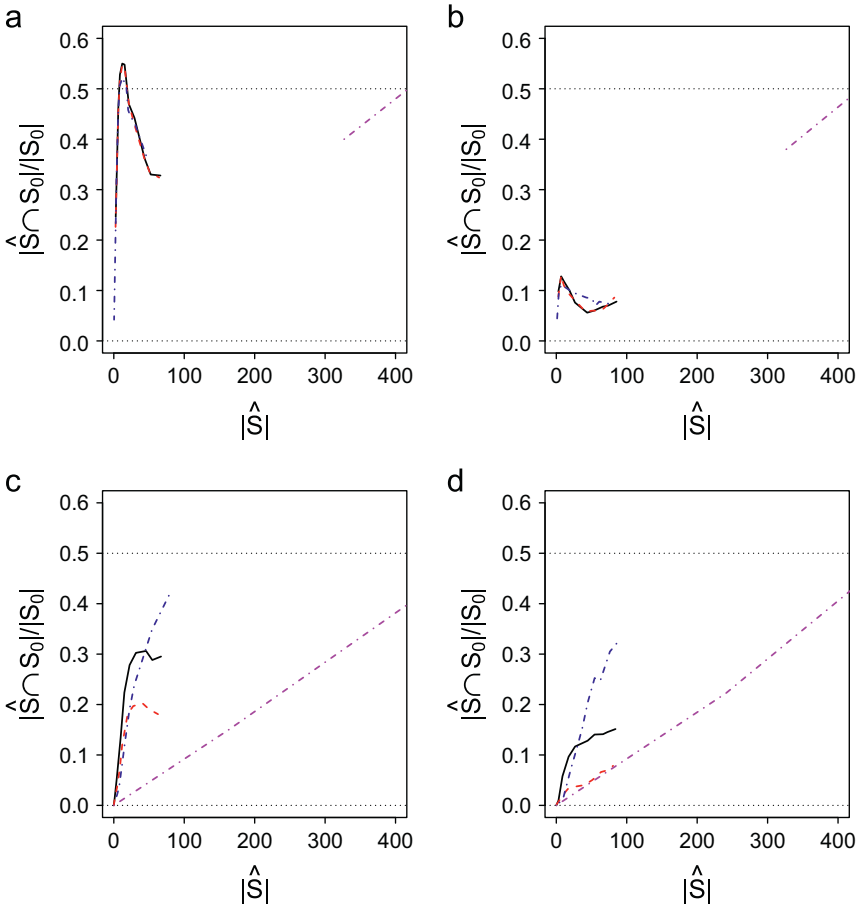

Fig. 6. Plots of $|\hat{S}\cap S_0|/|S_0|$ versus $|\hat{S}|$ for the pseudo real riboflavin data. Cluster representative Lasso (CRL, black solid line), cluster group Lasso (CGL, blue dashed-dotted line), Lasso (red dashed line) and CRLcor (magenta dashed-dotted line). (a) (D1), $\sigma=3$. (b) (D1), $\sigma=12$. (c) (D2), $\sigma=3$. (d) (D2), $\sigma=12$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

hierarchical clustering based on correlations (see Section 2.2) performs very poorly: the reason is that the automatically chosen number of clusters results in a partition with one very large cluster, and the representative mean value of such a very large cluster seems to be inappropriate. Using the group Lasso for such a partition (i.e., clustering) is ill-posed as well since the group size of such a large cluster is larger than sample size n.

Fig. 6 shows a somewhat different picture for variable screening. For the setting (D1), all methods except CRLcor perform similarly, but for (D2), the two cluster Lasso methods CRL and CGL perform better than plain Lasso. Especially for the low noise level $\sigma=3$ case, we see a substantial performance gain of the CRL and CGL compared to the Lasso. Nevertheless, the improvement over plain Lasso is less pronounced than for quite a few of the simulated models in Section 5.1. The CRLcor method is again performing very poorly: the reason is the same as mentioned above for prediction while in addition, if the large cluster is selected, it results in a large contribution of the cardinality $|\hat{S}|$.

5.3. Summarizing the empirical results

We clearly see that in the pseudo real data example and most of the simulation settings, the cluster Lasso techniques (CRL and CGL) outperform the Lasso in terms of variable screening; the gain is less pronounced for the pseudo real data example. Considering prediction, the CRL and the Lasso display similar performance while the CGL is not keeping up with them. Such a deficit of the CGL method seems to be caused for cases where we have many nonactive variables in an active group, leading to an efficiency loss: it might be repaired by using a sparse group Lasso (Simon et al., 2013). The difference between the clustering methods, Algorithm 1 and standard hierarchical clustering based on correlations (see Section 2.2), is essentially nonexistent for the simulation models in Section 5.1 while for the pseudo real data example in Section 5.2, the disagreement is huge and our novel Algorithm 1 leads to much better results.

6. Conclusions

We consider estimation in a high-dimensional linear model with strongly correlated variables. In such a setting, single variables cannot (or are at least very difficult to) be identified. We propose to group or cluster the variables first and do subsequent estimation with the Lasso for cluster-representatives (CRL: cluster representative Lasso) or with the group Lasso using the structure of the inferred clusters (CGL: Cluster group Lasso). Regarding the first step, we present a new bottom-up agglomerative clustering algorithm which aims for small canonical correlations between groups: we prove that it finds an optimal solution, that it is statistically consistent, and we give a simple rule for selecting the number of clusters. This new algorithm is motivated by the natural idea to address the problem of almost linear dependence between variables, but if preferred, it can be replaced by another suitable clustering procedure.

We present some theory which: (i) shows that canonical correlation based clustering leads to a (much) improved compatibility constant for the cluster group Lasso; and (ii) addresses bias and detection issues when doing subsequent estimation on cluster representatives, e.g. as with the CRL method. Regarding the second issue (ii), one favorable scenario is for (nearly) uncorrelated clusters with potentially many active variables in a cluster: the bias due to working with cluster representatives is small if the within group correlation is high, and detection is good if the regression coefficients within a group do not cancel. The other beneficial setting is for clusters with at most one active variable per cluster but the between cluster correlation does not need to be very small: if the cluster size is large or the correlation within the clusters is large, the bias due to cluster representatives is small and detection works well. We note that large cluster sizes cannot be properly handled by the cluster group Lasso while they can be advantageous for the cluster representative Lasso; instead of the group Lasso, one should take for such cases a sparse group Lasso (Simon et al., 2013) or a smoothed group Lasso (Meier et al., 2009). Our theoretical analysis sheds light when and why estimation with cluster representatives works well and leads to improvements, in comparison to the plain Lasso.

We complement the theoretical analysis with various empirical results which confirm that the cluster Lasso methods (CRL and CGL) are particularly attractive for improved variable screening in comparison to the plain Lasso. In view of the fact that variable screening and dimension reduction (in terms of the original variables) is one of the main applications of Lasso in high-dimensional data analysis, the cluster Lasso methods are an attractive and often better alternative for this task.

Appendix A. Proofs

A.1. Proof of Theorem 2.1

We first show an obvious result.

Lemma A.1. *Consider a partition* $\mathcal{G}=\{G_1,\ldots,G_q\}$ *which satisfies* (2). *Then, for every* $r,\ell\in\{1,\ldots,q\}$ *with* $r\neq\ell$:

$$\hat{\rho}_{\mathrm{can}}(J_1,J_2)\leq\tau \quad \text{for all subsets } J_1\subseteq G_r,\ J_2\subseteq G_\ell.$$

Proof. The proof follows immediately from the inequality

$$\hat{\rho}_{\mathrm{can}}(J_1,J_2)\leq\hat{\rho}_{\mathrm{can}}(G_r,G_\ell). \qquad \square$$

For proving Theorem 2.1, the fact that we obtain a solution satisfying (2) is a straightforward consequence of the definition of the algorithm which continues to merge groups until all canonical correlations between groups are less or equal to τ.

We now prove that the obtained partition is the finest clustering with $\tau-$separation. Let $\hat{\mathcal{G}}(\tau)=\{G_1,\ldots,G_q\}$ be an arbitrary clustering with $\tau-$separation and $\hat{\mathcal{G}}_b, b=1,\ldots,b^*$, be the sequence of partitions generated by the algorithm (where b^* is the stopping (first) iteration where $\tau-$separation is reached). We need to show that $\hat{\mathcal{G}}_{b^*}$ is a finer partition of $\{1,\ldots,p\}$ than $\hat{\mathcal{G}}(\tau)$. Here, the meaning of "finer than" is not strict, i.e., including "equal to". To this end, it suffices to prove by induction that $\hat{\mathcal{G}}_b$ is finer than $\hat{\mathcal{G}}(\tau)$ for $b=1,\ldots,b^*$. This is certainly true for $b=1$ since the algorithm begins with the finest partition of $\{1,\ldots,p\}$. Now assume the induction condition that $\hat{\mathcal{G}}_b$ is finer than $\hat{\mathcal{G}}(\tau)$ for $b<b^*$. The algorithm computes $\hat{\mathcal{G}}_{b+1}$ by merging two

members, say G' and G'', of $\mathcal{G}_b$ such that $\hat{\rho}_{\mathrm{can}}(G', G'') > \tau$. Since $\hat{\mathcal{G}}_b$ is finer than $\hat{\mathcal{G}}(\tau)$, there must exist members G_{j_1} and G_{j_2} of $\hat{\mathcal{G}}(\tau)$ such that $G' \subseteq G_{j_1}$ and $G'' \subseteq G_{j_2}$. This implies $\hat{\rho}_{\mathrm{can}}(G_{j_1}, G_{j_2}) \geq \hat{\rho}_{\mathrm{can}}(G', G'') > \tau$, see Lemma A.1. Since $\hat{\rho}_{\mathrm{can}}(G_j, G_k) \leq \tau$ for all $j \neq k$, we must have $j_1 = j_2$. Thus, the algorithm merges two subsets of a common member (namely $G_{j_1} = G_{j_2}$) of $\hat{\mathcal{G}}(\tau)$. It follows that $\hat{\mathcal{G}}_{b+1}$ is still finer than $\hat{\mathcal{G}}(\tau)$. □

A.2. Proof of Theorem 2.2

For ease of notation, we abbreviate a group index G_r by r. The proof of Theorem 2.2 is based on the following bound for the maximum difference between the sample and population correlations of linear combination of variables. Define

$$\Delta_{r,\ell} = \max_{u \neq 0, v \neq 0} |\hat{\rho}(\mathbf{X}^{(r)}u, \mathbf{X}^{(\ell)}v) - \rho(u^T X^{(r)}, v^T X^{(\ell)})|.$$

Lemma A.2. *Consider* $\mathbf{X}$ *as in* (3) *and* $\mathcal{G}^0 = \{G_1, \ldots, G_q\}$ *a partition of* $\{1, \ldots, p\}$. *Let* $\Sigma_{r,\ell} = \mathrm{Cov}(X^{(r)}, X^{(\ell)})$, $t > 0$ *and* $d_r = \mathrm{rank}(\Sigma_{r,r})$. *Define* $\Delta^*_{r,\ell}$ *by* (4). *Then,*

$$\mathbb{P}\left[\max_{1 \leq r < \ell \leq q} (\Delta_{r,\ell} - \Delta^*_{r,\ell}) \geq 0\right] \leq \exp(-t).$$

Proof. Taking a new coordinate system if necessary, we may assume without loss of generality that $\Sigma_{r,r} = I_{d_r}$ and $\Sigma_{\ell,\ell} = I_{d_\ell}$. Let $\hat{\Sigma}_{r,\ell}$ be the sample versions of $\Sigma_{r,\ell}$. For $r \neq \ell$, we write a linear regression model $\mathbf{X}^{(\ell)} = \mathbf{X}^{(r)}\Sigma_{r,\ell} + \boldsymbol{\varepsilon}^{(r,\ell)}V_{r,\ell}^{1/2}$ such that $\boldsymbol{\varepsilon}^{(r,\ell)}$ is an $n \times d_\ell$ matrix of i.i.d.N(0, 1) entries independent of $\mathbf{X}^{(r)}$ and $\|V_{r,\ell}^{1/2}\|_{(S)} \leq 1$, where $\|\cdot\|_{(S)}$ is the spectrum norm. This gives $\hat{\Sigma}_{r,\ell} = \hat{\Sigma}_{r,r}\Sigma_{r,\ell} + (\mathbf{X}^{(r)})^T\boldsymbol{\varepsilon}^{(r,\ell)}V_{r,\ell}^{1/2}/n$. Let $U_r U_r^T$ be the SVD of the projection $\mathbf{X}^{(r)}\hat{\Sigma}_{r,r}^{-1}(\mathbf{X}^{(r)})^T/n$, with $U_r \in \mathbb{R}^{n \times d_r}$. Since $\mathbf{X}^{(r)}$ is independent of $\boldsymbol{\varepsilon}^{(r,\ell)}$, $U_r^T\boldsymbol{\varepsilon}^{(r,\ell)}$ is a $d_r \times d_\ell$ matrix of i.i.d. N(0, 1) entries. Let

$$\Omega_r = \{\|\hat{\Sigma}_{r,r}^{1/2} - I_{d_r}\|_{(S)} \leq t_r\}, \quad \Omega_{r,\ell} = \{\|U_r^T\boldsymbol{\varepsilon}^{(r,\ell)}/\sqrt{n}\|_{(S)} \leq t_r \wedge t_\ell\}.$$

It follows from (4) and Theorem II.13 of Davidson and Szarek (2001) that

$$\max\{\mathbb{P}[\Omega_r^c], \mathbb{P}[\Omega_{r,\ell}^c]\} \leq 2\mathbb{P}[N(0,1) > \sqrt{2(t + \log(q(q+1)))}] \leq 2e^{-t}/\{q(q+1)\}.$$

In the event Ω_r, we have

$$\begin{aligned}
\Delta_{r,r} &= \max_{\|u\|_2 = \|v\|_2 = 1} \left| \frac{u\hat{\Sigma}_{r,r}v}{\|\hat{\Sigma}_{r,r}^{1/2}u\|_2 \|\hat{\Sigma}_{r,r}^{1/2}v\|_2} - u^T v \right| \\
&\leq \|\hat{\Sigma}_{r,r} - I_{d_r}\|_{(S)}/(1-t_r)^2 + \max_{\|u\|_2 = \|v\|_2 = 1} \left| \frac{u^T v}{\|\hat{\Sigma}_{r,r}^{1/2}u\|_2 \|\hat{\Sigma}_{r,r}^{1/2}v\|_2} - u^T v \right| \\
&\leq \{(1+t_r)^2 - 1\}/(1-t_r)^2 + 1/(1-t_r)^2 - 1 = \Delta^*_{r,r}.
\end{aligned}$$

For $r \neq \ell$, the variable change $u \to \hat{\Sigma}_{r,r}^{-1/2}u$ gives

$$\Delta_{r,\ell} = \max_{\|u\|_2 = \|v\|_2 = 1} \left| u^T \hat{\Sigma}_{r,r}^{-1/2} \left(\frac{\hat{\Sigma}_{r,\ell}}{\|\hat{\Sigma}_{\ell,\ell}^{1/2}v\|_2} - \frac{\Sigma_{r,\ell}}{\|\hat{\Sigma}_{r,r}^{-1/2}u\|_2} \right) v \right|.$$

In the event $\Omega_r \cap \Omega_\ell \cap \Omega_{r,\ell}$, we have

$$\begin{aligned}
\Delta_{r,\ell} &\leq \max_{\|u\|_2 = \|v\|_2 = 1} \{|u\hat{\Sigma}_{r,r}^{-1/2}(\hat{\Sigma}_{r,\ell} - \Sigma_{r,\ell})v| / \|\hat{\Sigma}_{\ell,\ell}^{1/2}v\|_2 \\
&\quad + |u^T\hat{\Sigma}_{r,r}^{-1/2}\Sigma_{r,\ell}v(1/\|\hat{\Sigma}_{\ell,\ell}^{1/2}v\|_2 - 1/\|\hat{\Sigma}_{r,r}^{-1/2}u\|_2)|\} \\
&\leq \max_{\|u\|_2 = \|v\|_2 = 1} \{\|\hat{\Sigma}_{r,r}^{-1/2}(\hat{\Sigma}_{r,\ell} - \Sigma_{r,\ell})\|_{(S)} / \|\hat{\Sigma}_{\ell,\ell}^{1/2}v\|_2 + |\|\hat{\Sigma}_{r,r}^{-1/2}u\|_2 / \|\hat{\Sigma}_{\ell,\ell}^{1/2}v\|_2 - 1|\} \\
&\leq \|(\hat{\Sigma}_{r,r}^{1/2} - \hat{\Sigma}_{r,r}^{-1/2})\Sigma_{r,\ell} + \hat{\Sigma}_{r,r}^{-1/2}(\mathbf{X}^{(r)})^T\boldsymbol{\varepsilon}^{(r,\ell)}V_{r,\ell}^{1/2}/n\|_{(S)}/(1-t_\ell) \\
&\quad + 1/\{(1-t_r)(1-t_\ell)\} - 1 \\
&\leq (\|\hat{\Sigma}_{r,r}^{1/2} - \hat{\Sigma}_{r,r}^{-1/2}\|_{(S)} + \|U_r^T\boldsymbol{\varepsilon}^{(r,\ell)}/\sqrt{n}\|_{(S)})/(1-t_\ell) + (t_r + t_\ell)/\{(1-t_r)(1-t_\ell)\} \\
&\leq \{1/(1-t_r) - (1-t_r) + t_r \wedge t_\ell\}/(1-t_\ell) + (t_r + t_\ell)/\{(1-t_r)(1-t_\ell)\} \\
&= (1 - (1-t_r)^2 + (1-t_r)(t_r \wedge t_\ell) + t_r + t_\ell)/\{(1-t_r)(1-t_\ell)\}.
\end{aligned}$$

Thus, for $t_r \leq t_\ell$, we find $\Delta_{r,\ell} \leq \Delta^*_{r,\ell}$. Since $\Delta_{r,\ell} = \Delta_{\ell,r}$, the above bounds hold simultaneously in the intersection of all Ω_r and those $\Omega_{r,\ell}$ with either $t_r < t_\ell$ or $r < \ell$ for $t_r = t_\ell$. Since there are totally $q + q(q-1)/2 = q(q+1)/2$ such events, $\mathbb{P}[\Delta_{r,\ell} \leq \Delta^*_{r,\ell}\ \forall r, \ell] \geq 1 - \exp(-t)$. □

Proof of Theorem 2.2. It follows from (5) that $\mathcal{G}^0$ is the finest population clustering with $\tau-$separation for all $\tau_- \le \tau \le \tau_+$. Since $\rho_{\mathrm{can}}(G_r, G_\ell) = \max_{u,v} \rho(u^T X^{(r)}, v^T X^{(\ell)})$ and $\hat{\rho}_{\mathrm{can}}(G_r, G_\ell) = \max_{u,v} \hat{\rho}(\mathbf{X}^{(r)}u, \mathbf{X}^{(\ell)}v)$, (5) and Lemma A.2 implies that with at least probability $1-e^{-t}$, the inequalities

$$\hat{\rho}_{\mathrm{can}}(G_r, G_\ell) \le \rho_{\mathrm{can}}(G_r, G_\ell) + \Delta^*_{r,\ell} \le \tau_-,$$
$$\max_{k_1 < k_2} \hat{\rho}_{\mathrm{can}}(G_{r;k_1}, G_{r;k_2}) \ge \max_{k_1 < k_2} \rho_{\mathrm{can}}(G_{r;k_1}, G_{r;k_2}) - \Delta^*_{r,r} \ge \tau_+,$$

hold simultaneously for all $r < \ell$ and nontrivial partitions $\{G_{r;k}, k \le q_r\}$ of G_r, $r = 1, \ldots, q$. In this case, $\mathcal{G}^0$ is also the finest sample clustering with $\tau-$separation for all $\tau_- \le \tau \le \tau_+$. The conclusion follows from Theorem 2.1. □

A.3. Proof of Theorem 4.1

We write for notational simplicity $S_0 := S_{0,\mathrm{Group}}$ and $s_0 := |S_{0,\mathrm{Group}}|$. Note first that for all r,

$$\|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2^2 = n\beta_{G_r}^T \hat{\Sigma}_{r,r}\beta_{G_r} = n\|\gamma_{G_r}\|_2^2,$$

where $\gamma_{G_r} := \Sigma_{r,r}^{1/2}\beta_{G_r}$. Moreover,

$$\|\mathbf{X}^{(S_0)}\beta_{S_0}\|_2^2 = n\gamma_{S_0}^T \hat{R}_{S_0}\gamma_{S_0}.$$

It follows that

$$\sum_{r \in S_0} \|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2^2 = n\|\gamma_{S_0}\|_2^2 \le n\gamma_{S_0}^T \hat{R}_{S_0}\gamma_{S_0}/\hat{\Lambda}^2_{\min} = \|X^{(S_0)}\beta_{S_0}\|_2^2/\hat{\Lambda}^2_{\min}. \tag{30}$$

Furthermore,

$$\begin{aligned}
\left|\beta_{S_0}^T (\mathbf{X}^{(S_0)})^T \mathbf{X}^{(S_0^c)}\beta_{S_0^c}\right| &= n\left|\sum_{r \in S_0}\sum_{\ell \in S_0^c} \beta_{G_r}^T \hat{\Sigma}_{r,\ell}\beta_{G_\ell}^T\right| \\
&= n\left|\sum_{r \in S_0}\sum_{\ell \in S_0^c} \gamma_{G_r}^T \hat{\Sigma}_{r,r}^{-1/2}\hat{\Sigma}_{r,\ell}\hat{\Sigma}_{\ell,\ell}^{-1/2}\gamma_{G_\ell}\right| \le n\sum_{r \in S_0}\sum_{\ell \in S_0^c} \hat{\rho}_{\mathrm{can}}(G_r, G_\ell)\|\gamma_{G_r}\|_2\|\gamma_{G_\ell}\|_2 \\
&= \sum_{r \in S_0}\sum_{\ell \in S_0^c} \hat{\rho}_{\mathrm{can}}(G_r, G_\ell)\|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2\|\mathbf{X}^{(G_\ell)}\beta_{G_\ell}\|_2 \\
&= \sum_{r \in S_0}\sum_{\ell \in S_0^c} \hat{\rho}_{\mathrm{can}}(G_r, G_\ell)\frac{\overline{m}}{\sqrt{m_r m_\ell}}\|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2\sqrt{\frac{m_r}{\overline{m}}}\|\mathbf{X}^{(G_\ell)}\beta_{G_\ell}\|_2\sqrt{\frac{m_\ell}{\overline{m}}} \\
&\le \rho\|\mathbf{X}^{(S_0)}\beta_{S_0}\|_{2,1}\|\mathbf{X}^{(S_0^c)}\beta_{S_0^c}\|_{2,1}.
\end{aligned}$$

Hence, for all β satisfying $\|\beta_{S_0^c}\|_{2,1} \le 3\|\beta_{S_0}\|_{2,1}$, we have

$$|\beta_{S_0}^T (\mathbf{X}^{(S_0)})^T \mathbf{X}^{(S_0^c)}\beta_{S_0^c}| \le 3\rho\|\beta_{S_0}\|_{2,1}^2. \tag{31}$$

Applying the Cauchy–Schwarz inequality and (30) gives

$$\begin{aligned}
\|\beta_{S_0}\|_{2,1}^2 &:= \left(\sum_{r \in S_0}\|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2\sqrt{m_r/\overline{m}}\right)^2 \\
&\le \frac{\sum_{r \in S_0} m_r}{\overline{m}}\sum_{r \in S_0}\|\mathbf{X}^{(G_r)}\beta_{G_r}\|_2^2 \le \frac{\sum_{r \in S_0} m_r}{\overline{m}}\frac{\|\mathbf{X}^{(S_0)}\beta_{S_0}\|_2^2}{\hat{\Lambda}^2_{\min}} \\
&= \frac{s_0\overline{m}_{S_0}}{\overline{m}}\frac{\|\mathbf{X}^{(S_0)}\beta_{S_0}\|_2^2}{\hat{\Lambda}^2_{\min}}.
\end{aligned}$$

Insert this in (31) to get

$$\left|\beta_{S_0}^T (\mathbf{X}^{(S_0)})^T \mathbf{X}^{(S_0^c)}\beta_{S_0^c}\right| \le \rho\frac{3s_0\overline{m}_{S_0}}{\overline{m}\hat{\Lambda}^2_{\min}}\|\mathbf{X}^{(S_0)}\beta_{S_0}\|_2^2.$$

Use the assumption that

$$\rho\frac{3s_0\overline{m}_{S_0}}{\overline{m}\hat{\Lambda}^2_{\min}} < 1,$$

and apply Lemma 6.26 in Bühlmann and van de Geer (2011) to conclude that

$$\|\mathbf{X}^{(S_0)}\beta_{S_0}\|_2^2 \le \left(1-\rho\frac{3s_0\overline{m}_{S_0}}{\overline{m}\hat{\Lambda}^2_{\min}}\right)^{-2}\|\mathbf{X}\beta\|_2^2.$$

Hence,

$$\|\beta_{S_0}\|_{2,1}^2 \le \frac{s_0\overline{m}_{S_0}}{\overline{m}}\frac{\|\mathbf{X}^{(S_0)}\beta_{S_0}\|_2^2}{\hat{\lambda}_{\min}^2} \le \left(1-\rho\frac{3s_0\overline{m}_{S_0}}{\overline{m}\hat{\lambda}_{\min}^2}\right)^{-2}\left(\frac{s_0\overline{m}_{S_0}}{\overline{m}}\right)\frac{\|\mathbf{X}\beta\|_2^2}{\hat{\lambda}_{\min}^2}.$$

This leads to the first lower bound in the statement of the theorem. The second lower bound follows immediately by the incoherence assumption for ρ. Furthermore, it is not difficult to see that $\hat{\lambda}_{\min}^2 \ge (1-|S_{0,\mathrm{Group}}|\rho_{S_{0,\mathrm{Group}}})$, and using the incoherence assumption for $\rho_{S_{0,\mathrm{Group}}}$ leads to strict positivity. □

A.4. Proof of Proposition 4.2

We can invoke the analysis given in Bühlmann and van de Geer (2011, Theorem 6.1). The slight deviations involve: (i) use that $\mathbb{E}[\eta_i|\mathbf{Z}]=0$; (ii) due to the Gaussian assumption $\mathrm{Var}(\eta_i|\mathbf{Z})$ is constant equaling $\mathrm{Var}(\eta_i|\mathbf{Z}) = \mathbb{E}[\mathrm{Var}(\eta_i|\mathbf{Z})] = \mathrm{Var}(\eta_i) = \xi^2 = \sigma^2 + \mathbb{E}[(\mu_X-\mu_Z)^2]$; and (iii) the probability bound in Bühlmann and van de Geer (2011, Lemma 6.2) can be easily obtained for nonstandardized variables when multiplying λ_0 with $\|\hat{\sigma}_{\mathbf{Z}}\|_\infty$. The issues (ii) and (iii) explain the factors appearing in the definition of λ_0. □

A.5. Proof of Proposition 4.3

Because of uncorrelatedness of $\overline{X}^{(r)}$ among $r=1,\ldots,q$, we have

$$\gamma_r^0 = \frac{\mathrm{Cov}(Y,\overline{X}^{(r)})}{\mathrm{Var}(\overline{X}^{(r)})} = |G_r|^{-1}\frac{\sum_{j\in G_r}\mathrm{Cov}(Y,X^{(j)})}{\mathrm{Var}(\overline{X}^{(r)})} = |G_r|\frac{\sum_{j\in G_r}\beta_j^0\mathrm{Var}(X^{(j)}) + \sum_{i\neq j}\beta_i^0\mathrm{Cov}(X^{(i)},X^{(j)})}{\sum_{\ell\in G_r}(\mathrm{Var}(X^{(\ell)}) + \sum_{i\neq\ell}\mathrm{Cov}(X^{(i)},X^{(\ell)}))}.$$

Define

$$w_j = \frac{\mathrm{Var}(X^{(j)}) + \sum_{i\neq j}\mathrm{Cov}(X^{(i)},X^{(j)})}{\sum_{\ell\in G_r}(\mathrm{Var}(X^{(\ell)}) + \sum_{i\neq\ell}\mathrm{Cov}(X^{(i)},X^{(\ell)}))}. \tag{32}$$

Then, $\sum_{j\in G_r} w_j = 1$ and $\gamma_r^0 = \sum_{j\in G_r} w_j\beta_j^0$.

The statement 1 follows immediately from (32). Regarding statement 2, we read off from (32) that $w_j \equiv |G_r|^{-1}$ for all $j\in G_r$ and hence $\gamma_r^0 = \sum_{j\in G_r}\beta_j^0$. □

A.6. Proof of Proposition 4.4

We have

$$\gamma_r^0 = \frac{\mathrm{Cov}(Y,\overline{X}^{(r)}|\{\overline{X}^{(\ell)};\ell\neq r\})}{\mathrm{Var}(\overline{X}^{(r)}|\{\overline{X}^{(\ell)};\ell\neq r\})},$$

since for Gaussian distributions, partial covariances equal conditional covariances (cf. Baba et al., 2004). For the numerator, we have

$$\begin{aligned}
\mathrm{Cov}(Y,\overline{X}^{(r)}|\{\overline{X}^{(\ell)};\ell\neq r\}) &= |G_r|^{-1}\sum_{j\in G_r}\mathrm{Cov}(Y,X^{(j)}|\{\overline{X}^{(\ell)};\ell\neq r\}) \\
&= |G_r|^{-1}\sum_{i,j\in G_r}\beta_i^0\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(\ell)};\ell\neq r\}) \\
&\quad + |G_r|^{-1}\sum_{j\in G_r}\sum_{i\notin G_r}\beta_i^0\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(\ell)};\ell\neq r\}) \\
&= |G_r|^{-1}\sum_{i,j\in G_r}\beta_i^0\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(\ell)};\ell\neq r\}) + \Gamma_r,\ |\Gamma_r| \le \|\beta^0\|_1\nu.
\end{aligned}$$

For the denominator, we have

$$\mathrm{Var}(\overline{X}^{(r)}|\{\overline{X}^{(\ell)};\ell\neq r\}) = |G_r|^{-2}\sum_{i,j\in G_r}\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(\ell)};\ell\neq r\}).$$

Defining

$$w_j = \frac{\sum_{i\in G_r}\mathrm{Cov}(X^{(i)},X^{(j)}|\{\overline{X}^{(\ell)};\ell\neq r\})}{\sum_{i,\ell\in G_r}\mathrm{Cov}(X^{(i)},X^{(\ell)}|\{\overline{X}^{(\ell)};\ell\neq r\})}, \tag{33}$$

we obtain $\gamma_r^0 = |G_r|\sum_{j\in G_r} w_j\beta_j^0 + \Delta_r$ with $|\Delta_r| \le \|\beta^0\|_1\nu/C$. The other statement follows using (33) and as statement 1 in Proposition 4.3. □

A.7. Proof of Proposition 4.5

Write

$$Y = \overline{X}^T\gamma^0 + \eta = U^T\tilde{\beta}^0 + \varepsilon = \overline{X}^T\tilde{\beta}^0 - W\tilde{\beta}^0 + \varepsilon.$$

Therefore,

$$\overline{X}(\tilde{\beta}^0 - \gamma^0) = \eta - \varepsilon + W^T\tilde{\beta}^0.$$

Taking the squares and expectation on both sides,

$$\begin{aligned}(\tilde{\beta}^0 - \gamma^0)^T\mathrm{Cov}(\overline{X})(\tilde{\beta}^0 - \gamma^0) &= \mathbb{E}[(\eta-\varepsilon)^2] + \mathbb{E}|W^T\tilde{\beta}^0|^2 \\ &= B^2 + \mathbb{E}|W^T\tilde{\beta}^0|^2 \le 2\mathbb{E}|W^T\tilde{\beta}^0|^2,\end{aligned}$$

where the last inequality follows from (22). Since $\mathrm{Cov}(\overline{X}) = \mathrm{Cov}(U) + \mathrm{Cov}(W)$, we have that $\lambda^2_{\min}(\mathrm{Cov}(\overline{X})) \ge \lambda^2_{\min}(\mathrm{Cov}(U))$. This completes the proof. □

References

Anderson, T., 1984. An Introduction to Multivariate Statistical Analysis, 2nd ed. Wiley.

Baba, K., Shibata, R., Sibuya, M., 2004. Partial correlation and conditional correlation as measures of conditional independence. Australian & New Zealand Journal of Statistics 46, 657–664.

Balding, D., 2007. A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7, 781–791.

Bickel, P., Ritov, Y., Tsybakov, A., 2009. Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics 37, 1705–1732.

Bondell, H., Reich, B., 2008. Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. Biometrics 64, 115–123.

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

Bühlmann, P., Kalisch, M., Meier, L. High-dimensional statistics with a view towards applications in biology. Annual Review in Statistics, in press.

Carlson, C., Eberle, M., Rieder, M., Yi, Q., Kruglyak, L., Nickerson, D., 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. American Journal of Human Genetics 74, 106–120.

Davidson, K., Szarek, S., 2001. Local operator theory, random matrices and Banach spaces. In: Johnson, W.B., Lindenstrauss, J. (Eds.), Handbook in Banach Spaces, vol. I. . Elsevier, pp. 317–366.

Dettling, M., Bühlmann, P., 2004. Finding predictive gene groups from microarray data. Journal of Multivariate Analysis 90, 106–131.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularized paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22.

Hastie, T., Tibshirani, R., Botstein, D., Brown, P., 2001. Supervised harvesting of expression trees. Genome Biology 2, 1–12.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Chan, W., Botstein, D., Brown, P., 2000. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology 1, 1–21.

Huang, J., Ma, S., Li, H., Zhang, C.-H., 2011. The sparse Laplacian shrinkage estimator for high-dimensional regression. Annals of Statistics 39, 2021–2046.

Kendall, M., 1957. A Course in Multivariate Analysis. Griffin, London.

Meier, L., van de Geer, S., Bühlmann, P., 2009. High-dimensional additive modeling. Annals of Statistics 37, 3779–3821.

Meinshausen, N., 2007. Relaxed Lasso. Computational Statistics & Data Analysis 52, 374–393.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. Annals of Statistics 34, 1436–1462.

Meinshausen, N., Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics 37, 246–270.

Segal, M., Dahlquist, K., Conklin, B., 2003. Regression approaches for microarray data analysis. Journal of Computational Biology 10, 961–980.

She, Y., 2010. Sparse regression with exact clustering. Electronic Journal of Statistics 4, 1055–1096.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group Lasso. Journal of Computational and Graphical Statistics, 22, 231–245.

Sun, T., Zhang, C.-H., 2012. Scaled sparse linear regression. Biometrika 99, 879–898.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tolosi, L., Lengauer, T., 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27, 1986–1994.

van de Geer, S., 2007. The deterministic Lasso. In: JSM Proceedings, 2007, 140. American Statistical Association.

van de Geer, S., Bühlmann, P., 2009. On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics 3, 1360–1392.

van de Geer, S., Lederer, J., 2012. The Lasso, correlated design, and improved oracle inequalities. In: IMS Collections 9 (From Probability to Statistics and Back: High-dimensional models and processes), 303–376.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 69, 49–67.

Zhang, C.-H., Huang, J., 2008. The sparsity and bias of the Lasso selection in high-dimensional linear regression. Annals of Statistics 36, 1567–1594.

Zhao, P., Yu, B., 2006. On model selection consistency of Lasso. Journal of Machine Learning Research 7, 2541–2563.

Zou, H., 2006. The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society Series B 67, 301–320.

Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Annals of Statistics 36, 1509–1566.