

Sparse Empirical Bayes Analysis (SEBA)

Natalia Bochkina & Ya'acov Ritov

January 8, 2010

Abstract

We consider a joint processing of n independent sparse regression problems. Each is based on a sample $(y_{i1}, x_{i1}) \dots, (y_{im}, x_{im})$ of m i.i.d. observations from $y_{i1} = x_{i1}^\top \beta_i + \varepsilon_{i1}$, $y_{i1} \in \mathbb{R}$, $x_{i1} \in \mathbb{R}^p$, $i = 1, \dots, n$, and $\varepsilon_{i1} \sim N(0, \sigma^2)$, say. p is large enough so that the empirical risk minimizer is not consistent. We consider three possible extensions of the lasso estimator to deal with this problem, the lassoes, the group lasso and the RING lasso, each utilizing a different assumption how these problems are related. For each estimator we give a Bayesian interpretation, and we present both persistency analysis and non-asymptotic error bounds based on restricted eigenvalue - type assumptions.

“... and only a star or two set sparsedly in the vault of heaven; and you will find a sight as stimulating as the hoariest summit of the Alps.” R. L. Stevenson

1 Introduction

We consider the model

$$Y_i = X_i^\top \beta_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

or more explicitly

$$y_{ij} = x_{ij}^\top \beta_i + \varepsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m$$

where $\beta_i \in \mathbb{R}^p$, $X_i \in \mathbb{R}^{m \times p}$ is either deterministic fixed design matrix, or a sample of m independent \mathbb{R}^p random vectors. Generally, we think of j indexing replicates (of similar items within the group) and i indexing groups (of replicates). Finally, ε_{ij} , $i = 1, \dots, n$, $j = 1, \dots, m$ are (at least uncorrelated with the x s), but typically assumed to be i.i.d. sub-Gaussian

random variables, independent of the regressors x_{ij} . We can consider this as n partially related regression models, with m i.i.d. observations on the each model. For simplicity, we assume that all variables have expectation 0. The fact that the number of observations does not dependent on i is arbitrary and is assumed only for the sake of notational simplicity.

The standard FDA (functional data analysis) is of this form, when the functions are approximated by their projections on some basis. Here we have n i.i.d. random functions, and each group can be considered as m noisy observations, each one is on the value of these functions at a given value of the argument. Thus,

$$y_{ij} = g_i(z_{ij}) + \varepsilon_{ij}, \quad (2)$$

where $z_{ij} \in [0, 1]$. The model fits the regression setup of (1), if $g(z) = \sum_{\ell=1}^p \beta_{\ell} h_{\ell}(z)$ where h_1, \dots, h_p are in $L_2(0, 1)$, and $x_{ij\ell} = h_{\ell}(z_{ij})$.

This approach is in the spirit of the empirical Bayes approach (or compound decision theory, note however that the term “empirical Bayes” has a few other meanings in the literature), cf, [11, 12, 8]. The empirical Bayes to sparsity was considered before, e.g., [15, 3, 7, 6]. However, in these discussions the compound decision problem was within a single vector, while we consider the compound decision to be between the vectors, where the vectors are the basic units. The beauty of the concept of compound decision, is that we do not have to assume that in reality the units are related. They are considered as related only because our loss function is additive.

One of the standard tools for finding sparse solutions in a large p small m situation is the lasso (Tibshirani [13]), and the methods we consider are its extensions.

We will make use of the following notation. Introduce $l_{p,q}$ norm of a set of vectors z_1, \dots, z_n , not necessarily of the same length, z_{ij} , $i = 1, \dots, n$, $j = 1, \dots, J_i$:

Definition 1.1 $\|z\|_{p,q} = \left[\sum_{i=1}^n \left(\sum_{j \in J_i} |z_{ij}|^p \right)^{q/p} \right]^{1/q}$.

These norms will serve as a penalty on the size of the matrix $\mathcal{B} = (\beta_1, \dots, \beta_n)$. Different norms imply different estimators, each appropriate under different assumptions.

Within the framework of the compound decision theory, we can have different scenarios, and we consider three of them. In Section 2 we investigate the situation when there is no direct relationship between the groups, and the only way the data are combined together is via the selection of the

common penalty. In this case the sparsity pattern of the solution for each group are unrelated. We argue that the alternative formulation of the lasso procedure in terms of $\ell_{2,1}$ (or, more generally, $\ell_{\alpha,1}$) norm which we refer to as “lassoes” can be more natural than the simple lasso, and this is argued from different points of view.

The motivation is as follows. The lasso method can be described in two related ways. Consider the one group version, $y_j = x_j^\top \beta + \varepsilon_j$. The lasso estimator can be defined by

$$\text{Minimize } \sum_{j=1}^m (y_j - x_j^\top \beta)^2 \quad \text{s.t.} \quad \|\beta\|_1 < A.$$

An equivalent definition, using Lagrange multiplier is given by

$$\text{Minimize } \sum_{j=1}^m (y_j - x_j^\top \beta)^2 + \lambda \|\beta\|_1^\alpha,$$

where α can be any arbitrarily chosen positive number. In the literature one can find almost only $\alpha = 1$. One exception is Greenshtein and Ritov [5] where $\alpha = 2$ was found more natural, also it was just a matter of aesthetics. We would argue that $\alpha > 2$ may be more intuitive. Our first algorithm generalizes this representation of the lasso directly to deal with compound model (1).

In the framework of the compound decision problem it is possible to consider the n groups as repeated similar models for p variables, and to choose the variables that are useful for all models. We consider this in Section 3. The relevant variation of the lasso procedure in this case is group lasso introduced by Yuan and Lin [14]:

$$\text{Minimize } \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - x_{ij}^\top \beta_i)^2 + \lambda \|\beta\|_{2,1}. \quad (3)$$

The authors also showed that in this case the sparsity pattern of variables is the same (with probability 1). Non-asymptotic inequalities under restricted eigenvalue type condition for group lasso are given by Lounici et al. [10].

Now, the standard notion of sparsity, as captured by the L_0 norm, or by the standard lasso and group lasso, is basis dependent. Consider the model of (2). If, for example, $g(z) = \mathbf{1}(a < z \leq b)$, then this example is sparse when $h_\ell(z) = \mathbf{1}(z > \ell/p)$. It is not sparse if $h_\ell(z) = (z - \ell/p)^+$. On the other hand, a function g which has a piece-wise constant slope is sparse in

the latter basis, but not in the former, even though, each function can be represented equally well in both bases.

Suppose that there is a sparse representation in some unknown basis, but assumed common to the n groups. The question arises, can we recover the basis corresponding to the sparsest representation? We will argue that this penalty, also known as trace norm or Schatten norm with $p = 1$, aims in finding the rotation that gives the best sparse representation of all vectors instantaneously (Section 4). We refer to this method as the rotation-invariant lasso, or shortly as the RING lasso. This is not surprising as under some conditions, this penalty also solves the minimum rank problem (see Candes and Recht [4] for the noiseless case, and Bach [1] for some asymptotic results). By analogy with the lassoes argument, a higher power of the trace norm as a penalty may be more intuitive to a Bayesian.

For both procedures considered here, the lassoes and the RING lasso, we present the bounds on their persistency as well as non-asymptotic inequalities under restricted eigenvalues type condition. All the proofs are given in the Appendix.

2 The lassoes procedure

The minimal structural relationship we may assume is that the β 's are not related, except that we believe that there is a bound on the average sparsity of the β 's. One possible approach would be to consider the problem as a standard sparse regression problem with nm observations, a single vector of coefficients $\beta = (\beta_1^\top, \dots, \beta_n^\top)^\top$, and a block diagonal design matrix X . This solution imposes very little on the similarity among β_1, \dots, β_n . The lassoes procedure discussed in this section assume that these vectors are similar, at least in their level of sparsity.

2.1 Prediction error minimization

In this paper we adopt an oracle point of view. Our estimator is the empirical minimizer of the risk penalized by the complexity of the solution (i.e., by its ℓ_1 norm). We compare this estimator to the solution of an ‘‘oracle’’ who does the same, but optimizing over the true, unknown to simple human beings, population distribution.

We assume that each vector of β_i , $i = 1, \dots, n$, solves a different problem, and these problems are related only through the joint loss function, which is the sum of the individual losses. To be clearer, we assume that for each $i = 1, \dots, n$, $z_{ij} = (y_{ij}, x_{ij}^\top)^\top$, $j = 1, \dots, m$ are i.i.d., sub-Gaussian

random variables, drawn from a distribution Q_i . Let $z_i = (y_i, x_i^\top)^\top$ be an independent sample from Q_i . For any vector a , let $\tilde{a} = (-1, a^\top)^\top$, and let $\tilde{\Sigma}_i$ be the covariance matrix of z_i and $\mathfrak{S} = (\tilde{\Sigma}_1, \dots, \tilde{\Sigma}_n)$. The goal is to find the matrix $\hat{\mathcal{B}} = (\hat{\beta}_1, \dots, \hat{\beta}_n)$ that minimizes the mean prediction error:

$$L(\mathcal{B}, \mathfrak{S}) = \sum_{i=1}^n \mathbb{E}_{Q_i} (y_i - x_i^\top \beta_i)^2 = \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i. \quad (4)$$

For p small, the natural approach is empirical risk minimization, that is replacing $\tilde{\Sigma}_i$ in (4) by \tilde{S}_i , the empirical covariance matrix of z_i . However, generally speaking, if p is large, empirical risk minimization results in overfitting the data. Greenshtein and Ritov [5] suggested (for the standard $n = 1$) minimization over a restricted set of possible β 's, in particular, to either L_1 or L_0 balls. In fact, their argument is based on the following simple observations

$$|\tilde{\beta}^\top (\tilde{\Sigma}_i - \tilde{S}_i) \tilde{\beta}| \leq \|\tilde{\Sigma}_i - \tilde{S}_i\|_\infty \|\tilde{\beta}\|_1^2$$

and

$$\|\tilde{\Sigma}_i - \tilde{S}_i\|_\infty = \mathcal{O}_p(m^{-1/2} \log p) \quad (5)$$

(see Lemma A.1 in the Appendix for the formal argument.)

This leads to the natural extension of the single vector lasso to the compound decision problem set up, where we penalize by the sum of the *squared* L_1 norms of vectors $\tilde{\beta}_1, \dots, \tilde{\beta}_n$, and obtain the estimator defined by:

$$\begin{aligned} (\tilde{\beta}_1, \dots, \tilde{\beta}_n) &= \arg \min_{\tilde{\beta}_1, \dots, \tilde{\beta}_n} \left\{ m \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{S}_i \tilde{\beta}_i + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \right\} \\ &= \arg \min_{\tilde{\beta}_1, \dots, \tilde{\beta}_n} \left\{ \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - x_{ij}^\top \beta_i)^2 + \lambda_n \|\tilde{\beta}_i\|_1^2 \right\}. \end{aligned} \quad (6)$$

The prediction error of the lassoes estimator can be bounded in the following way. In the statement of the theorem, c_n is the minimal achievable risk, while C_n is the risk achieved by a particular sparse solution.

Theorem 2.1 *Let β_{i0} , $i = 1, \dots, n$ be n arbitrary vectors and let $C_n = n^{-1} \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0}$. Let $c_n = n^{-1} \sum_{i=1}^n \min_{\beta} \tilde{\beta}^\top \tilde{\Sigma}_i \tilde{\beta}$. Then*

$$\sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i \leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \left(\frac{\lambda_n}{m} + \delta_n \right) \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 - \left(\frac{\lambda_n}{m} - \delta_n \right) \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2,$$

where $\delta_n = \max_i \|\tilde{S}_i - \Sigma_i\|_\infty$. If also $\lambda_n/m \rightarrow 0$ and $\lambda_n/(m^{1/2} \log(np)) \rightarrow \infty$, then

$$\sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 = \mathcal{O}_p(mn \frac{C_n - c_n}{\lambda_n}) + (1 + \mathcal{O}(\frac{m^{1/2}}{\lambda_n} \log(np))) \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 \quad (7)$$

and

$$\sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i \leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + (1 + \mathfrak{o}_p(1)) \frac{\lambda_n}{m} \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2.$$

The result is meaningful, although not as strong as may be wished, as long as $C_n - c_n \rightarrow 0$, while $n^{-1} \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 = \mathfrak{o}_p(m^{1/2})$. That is, when there is a relatively sparse approximations to the best regression functions. Here sparse means only that the L_1 norms of vectors is strictly smaller, on the average, than \sqrt{m} . Of course, if the minimizer of $\tilde{\beta}^\top \tilde{\Sigma}_i \tilde{\beta}$ itself is sparse, then by (7) $\tilde{\beta}_1, \dots, \tilde{\beta}_n$ are as sparse as the true minimizers .

Also note, that the prescription that the theorem gives for selecting λ_n , is sharp: choose λ_n as close as possible to $m\delta_n$, or slightly larger than \sqrt{m} .

2.2 A Bayesian perspective

The estimators $\tilde{\beta}_1, \dots, \tilde{\beta}_m$ look as if they are the mode of the a-posteriori distribution of the β_i 's when $y_{ij}|\beta_i \sim N(x_{ij}^\top \beta_i, \sigma^2)$, the β_1, \dots, β_n are a priori independent, and β_i has a prior density proportional to $\exp(-\lambda_n \|\beta_i\|_1^2 / \sigma^2)$. This distribution can be constructed as follows. Suppose $T_i \sim N(0, \lambda_n^{-1} \sigma^2)$. Given T_i , let u_{i1}, \dots, u_{ip} be distributed uniformly on the simplex $\{u_{i\ell} \geq 0, \sum_{\ell=1}^n u_{i\ell} = |T_i|\}$. Let s_{i1}, \dots, s_{ip} be i.i.d. Rademacher random variables (taking values ± 1 with probabilities 0.5), independent of $T_i, u_{i1}, \dots, u_{ip}$. Finally let $\beta_{i\ell} = u_{i\ell} s_{i\ell}$, $\ell = 1, \dots, p$.

However, this Bayesian point of view is not consistent with the conditions of Theorem 2.1. An appropriate prior should express the beliefs on the unknown parameter which are by definition conceptually independent of the amount data to be collected. However, the permitted range of λ_n does not depend on the assumed range of $\|\tilde{\beta}_i\|$, but quite artificially should be in order between $m^{1/2}$ and m . That is, the penalty should be increased with the number of observations on β_i , although in a slower rate than m . In fact, even if we relax what we mean by ‘‘prior’’, the value of λ_n goes in the ‘wrong’ direction. As $m \rightarrow \infty$, one may wish to use weaker a-priori assumptions, and permits T to have a-priori second moment going to infinity, not to 0, as entailed by $\lambda_n \rightarrow 0$.

We would like to consider a more general penalty of the form $\sum_{i=1}^n \|\beta_i\|_1^\alpha$. A power $\alpha \neq 1$ of ℓ_1 norm of β as a penalty introduces a priori dependence between the variables which is not the case for the regular lasso penalty with $\alpha = 1$, where all β_{ij} are a priori independent. As α increases, the sparsity of the different vectors tends to be the same. Note that given the value of λ_n , the n problems are treated independently. The compound decision problem is reduced to picking a common level of penalty. When this choice is data based, the different vectors become dependent. This is the main benefit of this approach—the selection of the regularization is based on all the mn observations.

For a proper Bayesian perspective, we need to consider a prior with much smaller tails than the normal. Suppose for simplicity that $c_n = C_n$ (that is, the “true” regressors are sparse), and $\max_i \|\beta_{i0}\|_1 < \infty$.

Theorem 2.2 *Let β_{i0} be the minimizer of $\tilde{\beta}^\top \Sigma_i \tilde{\beta}$. Suppose $\max_i \|\beta_{i0}\|_1 < \infty$. Consider the estimators:*

$$(\tilde{\beta}_1, \dots, \tilde{\beta}_n) = \arg \min_{\tilde{\beta}_1, \dots, \tilde{\beta}_n} \left\{ m \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{S}_i \tilde{\beta}_i + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^\alpha \right\}$$

for some $\alpha > 2$. Assume that $\lambda_n = \mathcal{O}(m\delta_m) = \mathcal{O}(m^{1/2} \log p)$. Then

$$n^{-1} \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 = \mathcal{O}((m\delta_n/\lambda_n)^{2/(\alpha-2)}),$$

and

$$\sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i \leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \mathcal{O}_p(n(m/\lambda_n)^{2/(\alpha-2)} \delta_n^{\alpha/(\alpha-2)}).$$

Remark 2.1 If the assumption $\lambda_n = \mathcal{O}(m\delta_m)$ does not hold, i.e. if $m\delta_m/\lambda_n = o(1)$, then the error term dominates the penalty and we get similar rates as in Theorem 2.1, i.e.

$$n^{-1} \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 = \mathcal{O}(1),$$

and

$$\sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i \leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \mathcal{O}_p(n\lambda_n/m).$$

Note that we can take in fact $\lambda_n \rightarrow 0$, to accommodate an increasing value of the $\tilde{\beta}_i$'s.

The theorem suggests a simple way to select λ_n based on the data. Note that $n^{-1} \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2$ is a decreasing function of λ . Hence, we can start with a very large value of λ and decrease it until $n^{-1} \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \approx \lambda^{-2/\alpha}$.

2.3 Restricted eigenvalues conditions and non-asymptotic inequalities

Before stating the conditions and the inequalities for the lasso procedure, we introduce some notation and definitions.

For a vector β , let $\mathcal{M}(\beta)$ be the cardinality of its support: $\mathcal{M}(\beta) = \sum_i \mathbf{1}(\beta_i \neq 0)$. Given a matrix $\Delta \in \mathbb{R}^{n \times p}$ and given a set $J = \{J_i\}$, $J_i \subset \{1, \dots, p\}$, we denote $\Delta_J = \{\Delta_{i,j}, i = 1, \dots, n, j \in J_i\}$. By the complement J^c of J we denote the set $\{J_1^c, \dots, J_n^c\}$, i.e. the set of complements of J_i 's. Below, X is $np \times m$ block diagonal design matrix, $X = \text{diag}(X_1, X_2, \dots, X_n)$, and with some abuse of notation, a matrix $\Delta = (\Delta_1, \dots, \Delta_n)$ may be considered as the vector $(\Delta_1^\top, \dots, \Delta_n^\top)^\top$. Finally, recall the notation $\mathcal{B} = (\beta_1, \dots, \beta_n)$

The restricted eigenvalue assumption of Bickel et al. [2] (and Lounici et al. [10]) can be generalized to incorporate unequal subsets J_i s. In the assumption below, the restriction is given in terms of $\ell_{q,1}$ norm, $q \geq 1$.

Assumption $\text{RE}_q(s, c_0, \kappa)$.

$$\kappa = \min \left\{ \frac{\|X^\top \Delta\|_2}{\sqrt{m} \|\Delta_J\|_2} : \max_i |J_i| \leq s, \Delta \in \mathbb{R}^{n \times p} \setminus \{0\}, \|\Delta_{J^c}\|_{q,1} \leq c_0 \|\Delta_J\|_{q,1} \right\} > 0.$$

We apply it with $q = 1$, and in Lounici et al. [10] it was used for $q = 2$. We call it a *restricted eigenvalue assumption* to be consistent with the literature. In fact, as stated it is a definition of κ as the maximal value that satisfies the condition, and the only real assumption is that κ is positive. However, the larger κ is, the more useful the ‘‘assumption’’ is. Discussion of the normalisation by \sqrt{m} can be found in Lounici et al. [10].

For penalty $\lambda \sum_i \|\beta_i\|_1^\alpha$, we have the following inequalities.

Theorem 2.3 *Assume $y_{ij} \sim \mathcal{N}(x_{ij}^\top \beta_i, \sigma^2)$, and let $\hat{\beta}$ be a minimizer of (6), with*

$$\lambda \geq \frac{4A\sigma \sqrt{m \log(np)}}{\alpha \max(B^{\alpha-1}, \hat{B}^{\alpha-1})},$$

where $\alpha \geq 1$ and $A > \sqrt{2}$, $B \geq \max_i \|\beta_i\|_1$ and $\hat{B} \geq \max_i \|\hat{\beta}_i\|_1$, $\max(B, \hat{B}) > 0$ (B may depend on n, m, p , and so can \hat{B}). Suppose that generalized assumption $RE_1(s, 3, \kappa)$ defined above holds, $\sum_{j=1}^m x_{ij\ell}^2 = m$ for all i, ℓ , and $\mathcal{M}(\beta_i) \leq s$ for all i .

Then, with probability at least $1 - (np)^{1-A^2/2}$,

(a) The root means squared prediction error is bounded by:

$$\frac{1}{\sqrt{nm}} \|X^\top(\hat{\mathcal{B}} - \mathcal{B})\|_2 \leq \frac{\sqrt{s}}{\kappa\sqrt{m}} \left[\frac{3\alpha\lambda}{2\sqrt{m}} \max(B^{\alpha-1}, \hat{B}^{\alpha-1}) + 2A\sigma\sqrt{\log(np)} \right],$$

(b) The mean estimation absolute error is bounded by:

$$\frac{1}{n} \|\mathcal{B} - \hat{\mathcal{B}}\|_1 \leq \frac{4s}{m\kappa^2} \left[\frac{3\alpha\lambda}{2} \max(B^{\alpha-1}, \hat{B}^{\alpha-1}) + 2A\sigma\sqrt{m\log(np)} \right],$$

(c) If $|\|\hat{\beta}_i\|_1^{\alpha-1} - b^{\alpha-1}/2| \geq 4\delta/b^{\alpha-1}$ for some $\delta > 0$,

$$\mathcal{M}(\hat{\beta}_i) \leq \|X_i(\beta_i - \hat{\beta}_i)\|_2^2 \frac{m\phi_{i,\max}}{\left(\lambda\alpha\|\hat{\beta}_i\|_1^{\alpha-1}/2 - A\sigma\sqrt{m\log(np)}\right)^2},$$

where $\phi_{i,\max}$ is the maximal eigenvalue of $X_i^\top X_i/m$.

Note that for $\alpha = 1$, if we take $\lambda = 2A\sigma\sqrt{m\log(np)}$, the bounds are of the same order as for the lasso with np -dimensional β (up to a constant of 2, cf. Theorem 7.2 in Bickel et al. [2]). For $\alpha > 1$, we have dependence of the bounds on the ℓ_1 norm of β and $\hat{\beta}$.

We can use bounds on the norm of $\hat{\beta}$ given in Theorem 2.2 to obtain the following results.

Theorem 2.4 Assume $y_{ij} \sim \mathcal{N}(x_{ij}^\top \beta_i, \sigma^2)$, with $\max_i \|\beta_i\|_1 \leq b$ where $b > 0$ can depend on n, m, p . Take some $\eta \in (0, 1)$. Let $\hat{\beta}$ be a minimizer of (6), with

$$\lambda = \frac{4A\sigma}{\alpha b^{\alpha-1}} \sqrt{m\log(np)},$$

$A > \sqrt{2}$, such that $b > c\eta^{1/(2(\alpha-1))}$ for some constant $c > 0$. Also, assume that $C_n - c_n = \mathcal{O}(m\delta_n)$, as defined in Theorem 2.1.

Suppose that generalized assumption $RE_1(s, 3, \kappa)$ defined above holds, $\sum_{j=1}^m x_{ij\ell}^2 = m$ for all i, ℓ , and $\mathcal{M}(\beta_i) \leq s$ for all i .

Then, for some constant $C > 0$, with probability at least $1 - (\eta + (np)^{1-A^2/2})$,

(a) The prediction error can be bounded by:

$$\|X^\top(\hat{\mathcal{B}} - \mathcal{B})\|_2^2 \leq \frac{4A^2\sigma^2 sn \log(np)}{\kappa^2} \left[1 + 3C \left(\frac{b}{\sqrt{\eta}} \right)^{(\alpha-1)/(\alpha-2)} \right]^2,$$

(b) The estimation absolute error is bounded by:

$$\|\mathcal{B} - \hat{\mathcal{B}}\|_1 \leq \frac{2A\sigma sn \sqrt{\log(np)}}{\kappa^2 \sqrt{m}} \left[1 + 3C \left(\frac{b}{\sqrt{\eta}} \right)^{(\alpha-1)/(\alpha-2)} \right].$$

(c) Average sparsity of $\hat{\beta}_i$:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{M}(\hat{\beta}_i) \leq s \frac{4\phi_{\max}}{\kappa^2 \delta^2} \left[1 + 3C \left(\frac{b}{\sqrt{\eta}} \right)^{1+1/(\alpha-2)} \right]^2,$$

where ϕ_{\max} is the largest eigenvalue of $X^\top X/m$.

This theorem also tells us how large ℓ_1 norm of β can be to ensure good bounds on the prediction and estimation errors.

Note that under the Gaussian model and fixed design matrix, assumption $C_n - c_n = \mathcal{O}(m\delta_n)$ is equivalent to $\|\mathcal{B}\|_2^2 \leq Cm\delta_n$.

3 Group LASSO: Bayesian perspective

Group LASSO is defined (see Yuan and Lin [14]) by

$$(\hat{\beta}_1, \dots, \hat{\beta}_n) = \arg \min \left[\sum_{i=1}^n \sum_{j=1}^m (y_{ij} - x_{ij}^\top \beta_i)^2 + \lambda \sum_{\ell=1}^p \left\{ \sum_{i=1}^n \beta_{i\ell}^2 \right\}^{1/2} \right] \quad (8)$$

Note that $(\hat{\beta}_1, \dots, \hat{\beta}_n)$ are defined as the minimum point of a strictly convex function, and hence they can be found by equating the gradient of this function to 0.

Recall the notation $\mathcal{B} = (\beta_1, \dots, \beta_n) = (\mathbf{b}_1^\top, \dots, \mathbf{b}_p^\top)^\top$. Note that (8) is equivalent to the mode of the a-posteriori distribution when given \mathcal{B} , Y_{ij} , $i = 1, \dots, n$, $j = 1, \dots, m$, are all independent, $y_{ij} \mid \mathcal{B} \sim \mathcal{N}(x_{ij}^\top \beta_i, \sigma^2)$, and a-priori, $\mathbf{b}_1, \dots, \mathbf{b}_p$, are i.i.d.,

$$f_{\mathbf{b}}(\mathbf{b}_\ell) \propto \exp\{-\tilde{\lambda} \|\mathbf{b}_\ell\|_2\}, \quad \ell = 1, \dots, p,$$

where $\tilde{\lambda} = \lambda/(2\sigma^2)$. We consider now some property of this prior. For each ℓ , b_ℓ have a spherically symmetric distribution. In particular they are uncorrelated and have mean 0. However, they are not independent. Change of variables to a polar system where

$$\begin{aligned} R_\ell &= \|\mathbf{b}_\ell\|_2 \\ \beta_{\ell i} &= R w_{\ell i}, \quad w_\ell \in \mathbb{S}^{n-1}, \end{aligned}$$

where \mathbb{S}^{n-1} is the sphere in \mathbb{R}^n . Then, clearly,

$$f(R_\ell, w_\ell) = C_{n,\lambda} R_\ell^{n-1} e^{-\tilde{\lambda} R_\ell}, \quad R_\ell > 0, \quad (9)$$

where $C_{n,\lambda} = \tilde{\lambda}^n \Gamma(n/2) / (2\Gamma(n)\pi^{n/2})$. Thus, R_ℓ, w_ℓ are independent $R_\ell \sim \Gamma(n, \tilde{\lambda})$, and w_ℓ is uniform over the unit sphere.

The conditional distribution of one of the coordinates of \mathbf{b}_ℓ , say the first, given the rest has the form

$$f(\mathbf{b}_{\ell 1} | \mathbf{b}_{\ell 2}, \dots, \mathbf{b}_{\ell n}, \sum_{i=2}^n \mathbf{b}_{\ell i}^2 = \rho^2) \propto e^{-\tilde{\lambda} \rho \sqrt{1 + \mathbf{b}_{\ell 1}^2 / \rho^2}}$$

which for small $\mathbf{b}_{\ell 1} / \rho$ looks like the normal density with mean 0 and variance $\rho / \tilde{\lambda}$, while for large $\mathbf{b}_{\ell 1} / \rho$ behaves like the exponential distribution with mean $\tilde{\lambda}^{-1}$.

The sparsity property of the prior comes from the linear component of log-density of R . If $\tilde{\lambda}$ is large and the Y s are small, this component dominates the log-a-posteriori distribution and hence the maximum will be at 0.

Fix now $\ell \in \{1, \dots, p\}$, and consider the estimating equation for \mathbf{b}_ℓ — the ℓ components of the β 's. Fix the rest of the parameters and let $\tilde{Y}_{ij\ell}^{\mathcal{B}} = y_{ij} - \sum_{k \neq \ell} \beta_{ik} x_{ijk}$. Then $\hat{\mathbf{b}}_{\ell i}$, $i = 1, \dots, n$, satisfy

$$\begin{aligned} 0 &= - \sum_{j=1}^m x_{ij\ell} (\tilde{Y}_{ij\ell}^{\mathcal{B}} - \hat{\mathbf{b}}_{\ell i} x_{ij\ell}) + \frac{\lambda \hat{\mathbf{b}}_{\ell i}}{\sqrt{\sum_k \hat{\mathbf{b}}_{\ell k}^2}}, \quad i = 1, \dots, n \\ &= - \sum_{j=1}^m x_{ij\ell} (\tilde{Y}_{ij\ell}^{\mathcal{B}} - \hat{\mathbf{b}}_{\ell i} x_{ij\ell}) + \lambda_\ell^* \hat{\mathbf{b}}_{\ell i}, \quad \text{say.} \end{aligned}$$

Hence

$$\hat{\mathbf{b}}_{\ell i} = \frac{\sum_{j=1}^m x_{ij\ell} \tilde{Y}_{ij\ell}^{\mathcal{B}}}{\lambda_\ell^* + \sum_{j=1}^m x_{ij\ell}^2}. \quad (10)$$

The estimator has an intuitive appeal. It is the least square estimator of $\mathbf{b}_{\ell i}$, $\sum_{j=1}^m x_{ij\ell} \tilde{Y}_{ij\ell}^{\mathcal{B}} / \sum_{j=1}^m x_{ij\ell}^2$, pulled to 0. It is pulled less to zero as the variance of $\mathbf{b}_{\ell 1}, \dots, \mathbf{b}_{\ell n}$ increases (and λ_{ℓ}^* is getting smaller), and as the variance of the LS estimator is lower (i.e., when $\sum_{j=1}^m x_{ij\ell}^2$ is larger).

If the design is well balanced, $\sum_{j=1}^m x_{ij\ell}^2 \equiv m$, then we can characterize the solution as follows. For a fixed ℓ , $\hat{\mathbf{b}}_{\ell 1}, \dots, \hat{\mathbf{b}}_{\ell n}$ are the least square solution shrunk toward 0 by the same amount, which depends only on the estimated variance of $\hat{\mathbf{b}}_{\ell 1}, \dots, \hat{\mathbf{b}}_{\ell n}$. In the extreme case, $\hat{\mathbf{b}}_{\ell 1} = \dots = \hat{\mathbf{b}}_{\ell n} = 0$, otherwise (assuming the error distribution is continuous) they are shrunken toward 0, but are different from 0.

We can use (10) to solve for λ_{ℓ}^*

$$\left(\frac{\lambda}{\lambda_{\ell}^*}\right)^2 = \|\hat{\mathbf{b}}_{\ell}\|_2^2 = \sum_{i=1}^n \left(\frac{\sum_{j=1}^m x_{ij\ell} \tilde{Y}_{ij\ell}^{\mathcal{B}}}{\lambda_{\ell}^* + \sum_{j=1}^m x_{ij\ell}^2} \right)^2.$$

Hence λ_{ℓ}^* is the solution of

$$\lambda^2 = \sum_{i=1}^n \left(\frac{\lambda_{\ell}^* \sum_{j=1}^m x_{ij\ell} \tilde{Y}_{ij\ell}^{\mathcal{B}}}{\lambda_{\ell}^* + \sum_{j=1}^m x_{ij\ell}^2} \right)^2. \quad (11)$$

Note that the RHS is monotone increasing, so (11) has at most a unique solution. It has no solution if at the limit $\lambda_{\ell}^* \rightarrow \infty$, the RHS is still less than λ^2 . That is if

$$\lambda^2 > \sum_{i=1}^n \left(\sum_{j=1}^m x_{ij\ell} \tilde{Y}_{ij\ell}^{\mathcal{B}} \right)^2$$

then $\hat{\mathbf{b}}_{\ell} = 0$. In particular if

$$\lambda^2 > \sum_{i=1}^n \left(\sum_{j=1}^m x_{ij\ell} Y_{ij\ell} \right)^2, \quad \ell = 1, \dots, p$$

Then all the random effect vectors are 0. In the balanced case the RHS is $\mathcal{O}_p(mn \log(p))$. By (9), this means that if we want that the estimator will be 0 if the underlined true parameters are 0, then the prior should prescribe that \mathbf{b}_{ℓ} has norm which is $\mathfrak{o}(m^{-1})$. This conclusion is supported by the recommended value of λ given, e.g. in [10].

Non-asymptotic inequalities and prediction properties of the group lasso estimators under restricted eigenvalues conditions are given in [10].

4 The RING lasso

The rotation invariant group (RING) lasso is suggested as a natural extension of the group lasso to the situation where the proper sparse description of the regression function within a given basis is not known in advance. For example, when we prefer to leave it a-priori open whether the function should be described in terms of the standard Haar wavelet basis, a collection of interval indicators, or a collection of step functions. All these three span the same linear space, but the true functions may be sparse in only one of them.

4.1 Definition

Let $A = \sum c_i x_i x_i^\top$, be a positive semi-definite matrix, where x_1, x_2, \dots is an orthonormal basis of eigenvectors. Then, we define $A^\gamma = \sum c_i^\gamma x_i x_i^\top$. We consider now as penalty the function

$$|||\mathcal{B}|||_1 = \text{trace} \left\{ \left(\sum_{i=1}^n \beta_i \beta_i^\top \right)^{1/2} \right\},$$

where $\mathcal{B} = (\beta_1, \dots, \beta_n) = (\mathbf{b}_1^\top, \dots, \mathbf{b}_p^\top)^\top$. This is also known as trace norm or Schatten norm with $p = 1$. Note that $|||\mathcal{B}|||_1 = \sum c_i^{1/2}$ where c_1, \dots, c_p are the eigenvalues of $\mathcal{B}\mathcal{B}^\top = \sum_{i=1}^n \beta_i \beta_i^\top$ (including multiplicities), i.e. this is the ℓ_1 norm on the singular values of \mathcal{B} . $|||\mathcal{B}|||_1$ is a convex function of \mathcal{B} .

In this section we study the estimator defined by

$$\hat{\mathcal{B}} = \arg \min_{\mathcal{B} \in \mathbb{R}^{p \times n}} \left\{ \sum_{i=1}^n (y_{ij} - x_{ij}^\top \beta_i)^2 + \lambda |||\mathcal{B}|||_1 \right\} \quad (12)$$

We refer to this problem as RING (Rotation INvariant Group) lasso.

The lassoes penalty considered primary the columns of \mathcal{B} . The main focus of the group lasso was the rows. Penalty $|||\mathcal{B}|||_1$ is symmetric in its treatment of the rows and columns since $\mathfrak{S}\mathcal{B} = \mathfrak{S}\mathcal{B}^\top$, where $\mathfrak{S}A$ denotes the spectrum of A . Moreover, the penalty is invariant to the rotation of the matrix \mathcal{B} . In fact, $|||\mathcal{B}|||_1 = |||T\mathcal{B}U|||_1$, where T and U are $n \times n$ and $p \times p$ rotation matrices:

$$(T\mathcal{B}U)^\top (T\mathcal{B}U) = U^\top \mathcal{B}^\top \mathcal{B} U$$

and the RHS have the same eigenvalues as $\mathcal{B}^\top \mathcal{B} = \sum \beta_i \beta_i^\top$.

The rotation-invariant penalty aims at finding a basis in which β_1, \dots, β_n have the same pattern of sparsity. This is meaningless if n is small — any function is well approximated by the span of the basis is sparse in under the right rotation. However, we will argue that this can be done when n is large.

The following lemma describes a relationship between group lasso and RING lasso.

Lemma 4.1

- (i) $\|\mathcal{B}\|_{2,1} \geq \inf_{U \in \mathcal{U}} \|U\mathcal{B}\|_{2,1} = \|\mathcal{B}\|_1$, where \mathcal{U} is the set of all unitary matrices.
- (ii) There is a unitary matrix U , which may depend on the data, such that if X_1, \dots, X_n are rotated by U^\top , then the solution of the RING lasso (12) is the solution of the group lasso in this basis.

4.2 The estimator

Let $\mathcal{B} = \sum_{\xi=1}^{p \wedge n} \alpha_\xi \beta_\xi^* \mathbf{b}_\xi^{*\top}$ be the singular value decomposition, or the PCA, of \mathcal{B} : $\beta_1^*, \dots, \beta_p^*$ and $\mathbf{b}_1^*, \dots, \mathbf{b}_n^*$ are orthonormal sub-bases of \mathbb{R}^p and \mathbb{R}^n respectively, $\alpha_1 \geq \alpha_2 \geq \dots$, and $\mathcal{B}\mathcal{B}^\top \beta_\xi^* = \alpha_\xi^2 \beta_\xi^*$, $\mathcal{B}^\top \mathcal{B} \mathbf{b}_\xi^* = \alpha_\xi^2 \mathbf{b}_\xi^*$, $\xi = 1, \dots, p \wedge n$. Let $T = \sum_{\xi=1}^{p \wedge n} e_\xi \beta_\xi^{*\top}$ (clearly, $TT^\top = I$). Consider the parametrization of the problem in the rotated coordinates, $\tilde{x}_{ij} = Tx_{ij}$ and $\tilde{\beta}_i = T\beta_i$. Then geometrically the regression problem is invariant: $x_{ij}^\top \beta_i = \tilde{x}_{ik}^\top \tilde{\beta}_i$, and $\|\mathcal{B}\|_1 = \|\tilde{\mathcal{B}}\|_{2,1}$, up to a modified regression matrix.

The representation $\hat{\mathcal{B}} = \sum_{\xi=1}^s \alpha_\xi \beta_\xi^* \mathbf{b}_\xi^{*\top}$ shows that the difficulty of the problem is the difficulty of estimating $s(n+p)$ parameters with nm observations. Thus it is feasible as long as $s/m \rightarrow 0$ and $sp/nm \rightarrow 0$.

We have

Theorem 4.2 *Suppose $p < n$. Then the solution of the RING lasso is given by $\sum_{\xi=1}^s \beta_\xi^* \mathbf{b}_\xi^{*\top}$, $s = s_\lambda \leq p$, and $s_\lambda \searrow 0$ as $\lambda \rightarrow \infty$. If $s = p$ then the gradient of the target function is given in a matrix form by*

$$-2R + \lambda(\hat{\mathcal{B}}\hat{\mathcal{B}}^\top)^{-1/2}\hat{\mathcal{B}}$$

where

$$R = \left(X_1^\top(Y_1 - X_1\hat{\beta}_1), \dots, X_n^\top(Y_n - X_n\hat{\beta}_n) \right).$$

And hence

$$\hat{\beta}_i = (X_i^\top X_i + \frac{\lambda}{2}(\hat{\mathcal{B}}\hat{\mathcal{B}}^\top)^{-1/2})^{-1} X_i^\top Y_i.$$

That is, the solution of a ridge regression with adaptive weight.

More generally, let $\hat{\mathcal{B}} = \sum_{\xi=1}^s \alpha_\xi \beta_\xi^* \mathbf{b}_\xi^\top$, $s < p$, where $\beta_1^*, \dots, \beta_p^*$ is an orthonormal base of \mathbb{R}^p . Then the solution satisfies

$$\begin{aligned} \beta_\xi^{*\top} R &= \frac{\lambda}{2} \beta_\xi^{*\top} (\hat{\mathcal{B}}\hat{\mathcal{B}}^\top)^{+1/2} \hat{\mathcal{B}}, \quad \xi \leq s \\ |\beta_\xi^{*\top} R \mathbf{b}_\xi^*| &\leq \frac{\lambda}{2}, \quad s < \xi \leq p. \end{aligned}$$

where for any positive semi-definite matrix A , $A^{+1/2}$ is the Moore-Penrose generalized inverse of $A^{1/2}$.

Roughly speaking the following can be concluded from the theorem. Suppose the data were generated by a sparse model (in *some* basis). Consider the problem in the transformed basis, and let S be the set of non-zero coefficients of the true model. Suppose that the design matrix is of full rank within the sparse model: $X_i^\top X_i = \mathcal{O}(m)$, and that λ is chosen such that $\lambda \gg \sqrt{nm \log(np)}$. Then the coefficients corresponding to S satisfy

$$\hat{\beta}_{Si} = (X_i^\top X_i + \frac{\lambda}{2}(\hat{\mathcal{B}}_S \hat{\mathcal{B}}_S^\top)^{1/2})^{-1} X_i^\top Y_i.$$

Since it is expected that $\lambda(\mathcal{B}_S \mathcal{B}_S^\top)^{1/2}$ is only slightly larger than $\mathcal{O}(m \log(np))$, it is completely dominated by $X_i^\top X_i$, and the estimator of this part of the model is consistent. On the other hand, the rows of R corresponding to coefficient not in the true model are only due to noise and hence each of them is $\mathcal{O}(\sqrt{nm})$. The factor of $\log(np)$ ensures that their maximal norm will be below $\lambda/2$, and the estimator is consistent.

4.3 Bayesian perspectives

We consider now the penalty for β_k for a fixed k . Let $A = n^{-1} \sum_{k \neq i} \beta_k \beta_k^\top$, and write the spectral value decomposition $n^{-1} \sum_{k=1}^n \beta_k \beta_k^\top = \sum c_j x_j x_j^\top$ where $\{x_j\}$ is an orthonormal basis of eigenvectors. Using Taylor expansion for not too big β_i , we get

$$\text{trace}((nA + \beta_i \beta_i^\top)^{1/2}) \approx \sqrt{n} \text{trace}(A^{1/2}) + \sum_{j=1}^p \frac{x_j^\top \beta_i \beta_i^\top x_j}{2c_j^{1/2}}$$

$$\begin{aligned}
&= \sqrt{n} \operatorname{trace}(A^{1/2}) + \frac{1}{2} \beta_i^\top \left(\sum c_j^{-1/2} x_j x_j^\top \right) \beta_i \\
&= \sqrt{n} \operatorname{trace}(A^{1/2}) + \frac{1}{2} \beta_i^\top A^{-1/2} \beta_i
\end{aligned}$$

So, this like β_i has a prior of $\mathcal{N}(0, n\sigma^2/\lambda A^{1/2})$. Note that the prior is only related to the estimated variance of β , and A appears with the power of $1/2$. Now A is not really the estimated variance of β , only the variance of the estimates, hence it should be inflated, and the square root takes care of that. Finally, note that eventually, if β_i is very large relative to nA , then the penalty become $\|\beta\|$, so the ‘‘prior’’ becomes essentially normal, but with exponential tails.

A better way to look on the penalty from a Bayesian perspective is to consider it as prior on the $n \times p$ matrix $\mathcal{B} = (\beta_1, \dots, \beta_n)$. Recall that the penalty is invariant to the rotation of the matrix \mathcal{B} . In fact, $\|\mathcal{B}\|_1 = \|\|T\mathcal{B}U\|_1$, where T and U are $n \times n$ and $p \times p$ rotation matrices. Now, this means that if $\mathbf{b}_1, \dots, \mathbf{b}_p$ are orthonormal set of eigenvectors of $\mathcal{B}^\top \mathcal{B}$ and $\gamma_{ij} = \mathbf{b}_j^\top \beta_i$ — the PCA of β_1, \dots, β_n , then $\|\mathcal{B}\|_1 = \sum_{j=1}^p \left(\sum_{i=1}^n \gamma_{ij}^2 \right)^{1/2}$ — the RING lasso penalty in terms of the principal components. The ‘‘prior’’ is then proportional to $e^{-\lambda \sum_{j=1}^p \|\gamma_{\cdot j}\|^2}$. which is as if to obtain a random \mathcal{B} from the prior the following procedure should be followed:

1. Sample r_1, \dots, r_p independently from $\Gamma(n, \lambda)$ distribution.
2. For each $j = 1, \dots, p$ sample $\gamma_{1j}, \dots, \gamma_{nj}$ independently and uniformly on the sphere with radius r_j .
3. Sample an orthonormal base χ_1, \dots, χ_p ‘‘uniformly’’.
4. Construct $\beta_i = \sum_{j=1}^p \gamma_{ij} \chi_j$.

4.4 Inequalities under an RE condition

The assumption on the design matrix X needs to be modified to account for the search over rotations, in the following way.

Assumption RE2(s, c_0, κ). For some integer s such that $1 \leq s \leq p$, and a positive number c_0 the following condition holds:

$$\begin{aligned}
\kappa = \min \{ & \frac{\|X^\top \Delta\|_2}{\sqrt{m} \|P_V \Delta\|_2} : V \text{ is a linear subspace of } \mathbb{R}^p, \dim(V) \leq s, \\
& \Delta \in \mathbb{R}^{p \times n} \setminus \{0\}, \|(I - P_V) \Delta\|_1 \leq c_0 \|P_V \Delta\|_1 \} > 0,
\end{aligned}$$

where P_V is the projection on linear subspace V .

If we restrict the subspaces V to be of the form $V = \bigoplus_{k=1}^r \langle e_{i_k} \rangle$, $r \leq s$ and $\langle e_i \rangle$ is the linear subspace generated by the standard basis vector e_i , and change the Schatten norm to $\ell_{2,1}$ norm, then we obtain the restricted eigen value assumption $\text{RE}_2(s, c_0, \kappa)$ of Lounici et al. [10].

Theorem 4.3 *Let $y_{ij} \sim \mathcal{N}(f_{ij}, \sigma^2)$ independent, $f_{ij} = x_{ij}^\top \beta_i$, $x_{ij} \in \mathbb{R}^p$, $\beta_i \in \mathbb{R}^p$, $i = 1, \dots, n$, $j = 1, \dots, m$, $p \geq 2$. Assume that $\sum_{j=1}^m x_{ij\ell}^2 = m$ for all i, ℓ . Let assumption $\text{RE}_2(s, 3, \kappa)$ be satisfied for $X = (x_{ij})$, where $s = \text{rank}(\mathcal{B})$. Consider the RING lasso estimator $\hat{f}_{ij} = X_{ij}^\top \hat{\beta}_i$ where $\hat{\mathcal{B}}$ is defined by (12) with*

$$\lambda = 4\sigma\sqrt{(A+1)mnp}, \quad \text{for some } A > 1.$$

Then, for large n or p , with probability at least $1 - e^{-Anp/8}$,

$$\begin{aligned} \frac{1}{mn} \|X^\top(\mathcal{B} - \hat{\mathcal{B}})\|_2^2 &\leq \frac{64(A+1)\sigma^2 sp}{\kappa^2 m}; \\ \frac{1}{n} \|\|\mathcal{B} - \hat{\mathcal{B}}\|_1 &\leq \frac{32\sigma\sqrt{1+A} s \sqrt{p}}{\kappa^2 \sqrt{mn}}, \\ \text{rank}(\hat{\mathcal{B}}) &\leq s \frac{64\phi_{\max}}{\kappa^2}, \end{aligned}$$

where ϕ_{\max} is the maximal eigenvalue of $X^\top X/m$.

Thus we have bounds similar to those of group lasso as a function of the threshold λ , with s being the rank of \mathcal{B} rather than its sparsity. However, for RING lasso we need a larger threshold compared to that of the group lasso ($\lambda_{GL} = 4\sigma\sqrt{mn} \left(1 + \frac{A \log p}{\sqrt{n}}\right)^{1/2}$, Lounici et al. [10]).

4.5 Persistence

We discuss now the persistence of the RING lasso estimators (see Section A.1 for definition and a general result).

We focus on the sets which are related to the trace norm which defines the RING lasso estimator:

$$B_{n,p} = \{\mathcal{B} \in \mathbb{R}^{n \times p} : \|\|\mathcal{B}\|_1 \leq b(n, p)\}.$$

Theorem 4.4 *Assume that $n > 1$. For any $F \in \mathcal{F}_{n,p}^m(V)$, $\beta \in B_{n,p}$ and*

$$\hat{\beta}^{(m,n,p)} = \arg \min_{\beta \in B_{n,p}} L_F(\beta),$$

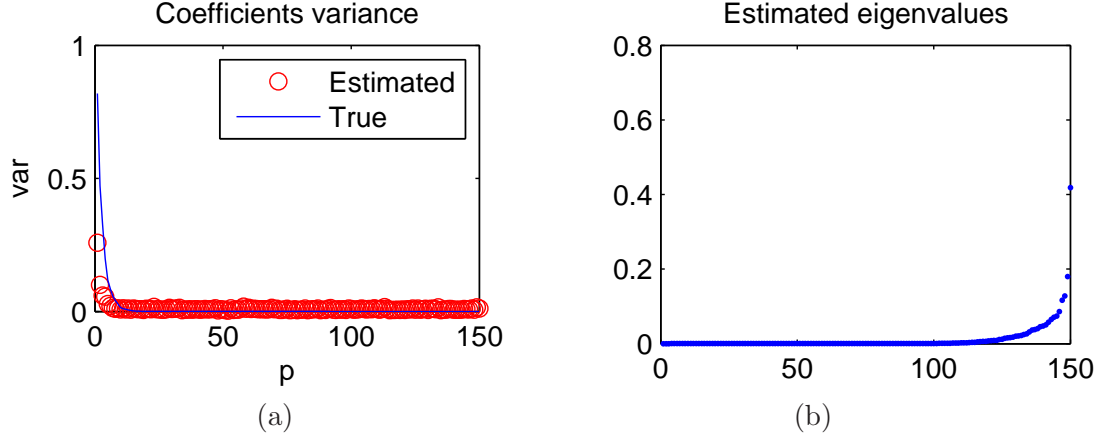


Figure 1: Component variances and eigenvalues, $m = 25$, $n = 150$

we have

$$L_F(\hat{\beta}) - \min_{\beta \in B_{n,p}} L_F(\beta) \leq \left(\frac{1}{m} + \frac{pb^2}{nm} \right) \left(16eV \frac{\log(np)}{m\eta} \right)^{1/2}$$

with probability at least $1 - \eta$, for any $\eta \in (0, 1)$.

Thus, for η sufficiently small, the conditions $\log(np) \leq c_p m^3 \eta$ and $b \leq c_b \sqrt{nm/p}$, for some $c_b, c_p > 0$, imply that with sufficiently high probability, the estimator is persistent. Roughly speaking, b is the number of components in the SVD of \mathcal{B} (the rank of \mathcal{B} , $\mathcal{M}(\beta)$ after the proper rotation), and if $m \gg \log n$, then what is needed is that this number will be strictly less $n^{1/2} m^{3/4} p^{-1/2}$. That is, if the true model is sparse, p can be almost as large as $m^{3/2} n^{1/2}$.

4.6 Algorithm and small simulation study

A simple algorithm is the following:

1. Initiate some small value of $\hat{\beta}_1, \dots, \hat{\beta}_n$. Let $A = \sum_{j=1}^n \hat{\beta}_j \hat{\beta}_j^T$. Fix $\gamma \in (0, 1]$, $\varepsilon > 0$, k , and $c > 1$.
2. For $i = 1, \dots, n$:
 - (a) Compute $\delta_i = (X_i^T X_i + \lambda A^{-1/2})^{-1} X_i^T (y_i - X_i \hat{\beta}_i)$.
 - (b) Update $A \leftarrow A - \hat{\beta}_i \hat{\beta}_i^T$; $\hat{\beta}_i \leftarrow \hat{\beta}_i + \gamma \delta_i$; $A \leftarrow A + \hat{\beta}_i \hat{\beta}_i^T$;

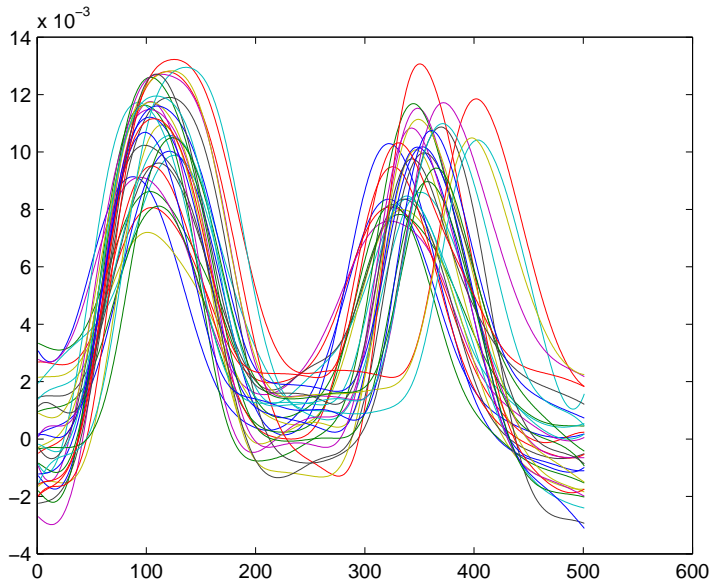


Figure 2: Lower lip position while repeating 32 times ‘Say bob again’

3. if $\sum_{j=1}^p \mathbf{1}(n^{-1} \sum_{i=1}^n \hat{\beta}_{ij}^2 > \varepsilon) > k$ update $\lambda \leftarrow \lambda c$ otherwise $\lambda \leftarrow \lambda/c$.
4. Return to step 2 unless there is no real change of coefficients.

To fasten the computation, the SVD was computed only every 10 values of i .

As a simulation we applied the above algorithm to the following simulated data. We generated random $\beta_1, \dots, \beta_{150} \in \mathbb{R}^{150}$ such that all coordinates are independent, and $\beta_{ij} \sim \mathcal{N}(0, e^{-2j/5})$. All $X_{ij\ell}$ are i.i.d. $\mathcal{N}(0, 1)$, and $y_{ij} = x_{ij}^\top \beta_i + \varepsilon_{ij}$, where ε_{ij} are all i.i.d. $\mathcal{N}(0, 1)$. The true R^2 obtained was approximately 0.73. The number of replicates per value of β , m , varied between 5 to 300. We consider two measures of estimation error:

$$L_{\text{par}} = \frac{\sum_{i=1}^n \|\hat{\beta}_i - \beta_i\|_\infty}{\sum_{i=1}^n \|\beta_i\|_\infty}$$

$$L_{\text{pre}} = \frac{\sum_{i=1}^n \|X_j(\hat{\beta}_i - \beta_i)\|_\infty}{\sum_{i=1}^n \|X_i \beta_i\|_\infty}$$

The algorithm stopped after 30–50 iterations. Figure is a graphical presentation of a typical result. A summary is given in Table 1. Note that m

has a critical impact on the estimation problem. However, with as little as 5 observations per R^{150} vector of parameter we obtain a significant reduction in the prediction error.

m	L_{par}	L_{pre}
5	0.9530 (0.0075)	0.7349 (0.0375)
25	0.7085 (0.0289)	0.7364 (0.0238)
300	0.2470 (0.0080)	0.5207 (0.0179)

Table 1: The estimation and prediction error as function of the number of observations per vector of parameters Means (and SDK).

The technique is natural for functional data analysis. We used the data LipPos. The data is described by Ramsay and Silverman and can be found in <http://www.stats.ox.ac.uk/~silverma/fdacasbook/lipemg.html>. The original data is given in Figure 2. However we added noise to the data as can be seen in Figure 3. The lip position is measured at $m = 501$ time points, with $n = 32$ repetitions.

As the matrix X we considered the union of 6 cubic spline bases with, respectively, 5, 10, 20, 100, 200, and 500 knots (i.e., $p = 841$, and X_i does not depend on i). A Gaussian noise with $\sigma = 0.001$ was added to Y . The result of the analysis is given in Figure 3. Figure 4 presents the projection of the mean path on the first eigen-vectors of $\sum_{i=1}^n \hat{\beta}_i \hat{\beta}_i^T$.

The final example we consider is somewhat arbitrary. The data, taken from StatLib, is of the daily wind speeds for 1961-1978 at 12 synoptic meteorological stations in the Republic of Ireland. As the Y variable we considered one of the stations (station BIR). As explanatory variables we considered the 11 other station of the same day, plus all 12 stations 70 days back (with the constant we have altogether 852 explanatory variables). The analysis was stratified by month. For simplicity, only the first 28 days of the month were taken, and the first year, 1961, served only for explanatory purpose. The last year was served only for testing purpose, so, the training set was for 16 years ($n = 12$, $m = 448$, and $p = 852$). In Figure 5 we give the 2nd moments of the coefficients and the scatter plot of predictions vs. true value of the last year.

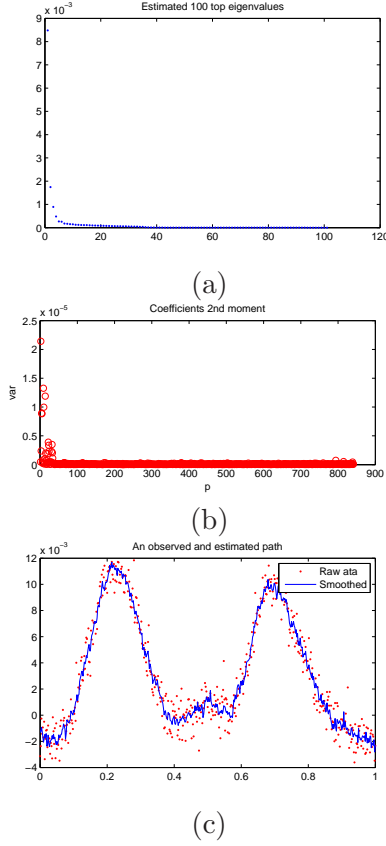


Figure 3: Eigenvalue, coefficient variance and typical observed and smooth path.

A Appendix

A.1 General persistence result.

A sequence of estimators $\hat{\beta}^{(m,n,p)}$ is persistent with respect to a set of distributions $\mathcal{F}_{n,p}^m$ for $\beta \in B_{n,p}$, if for any $F_{m,n,p} \in \mathcal{F}_{n,p}^m$,

$$L_{F_{m,n,p}}\left(\hat{\beta}^{(m,n,p)}\right) - L_{F_{m,n,p}}\left(\beta_{F_{m,n,p}}^*\right) \xrightarrow{P} 0,$$

where $L_F(\beta) = (nm)^{-1} E_F \sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - X_{ij}^\top \beta_i)^2$, $F_{m,n,p}$ is the empirical distribution function of $n \times (p+1)$ matrix Z , $Z_i = (Y_i, X_{i1}, \dots, X_{ip})$, $i = 1, \dots, n$, observed m times. Here $\beta_{F_{m,n,p}}^* = \arg \min_{\beta \in B_{n,p}} L_{F_{m,n,p}}(\beta)$, and

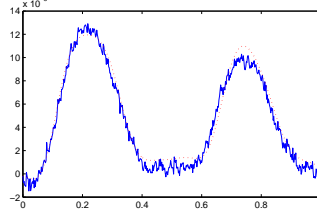


Figure 4: Projection of the estimated mean path on the 2 first eigen-vectors of $\sum_{i=1}^n \hat{\beta}_i \hat{\beta}_i^T$ and the true mean path.

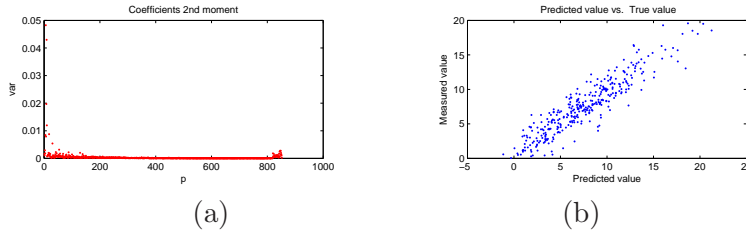


Figure 5: Coefficient 2nd moment and prediction vs.true value of the test year.

$\mathcal{F}_{n,p}^m$ stands for a collection of distributions of m observations of vectors $Z_i = (Y_i, X_{i1}, \dots, X_{ip})$, $i = 1, \dots, n$.

Assumption F. Under the distributions of random variables Z in $\mathcal{F}_{n,p}$, $\xi_{ilk} = Z_{i\ell} Z_{ik}$ satisfy $E(\max_{i=1, \dots, n} \max_{\ell, k=1, \dots, p+1} \xi_{ilk}^2) < V$. Denote this set of distributions by $\mathcal{F}_{n,p}(V)$.

This assumption is similar to one of the assumptions of Greenshtein and Ritov (2004). It is satisfied if, for instance, the distribution of $Z_{i\ell}$ has finite support and the variance of $Z_{i\ell} Z_{ik}$ is finite.

Lemma A.1 Let $F \in \mathcal{F}_{n,p}(V)$, and denote $\Sigma_i = (\sigma_{ijk})$ and $\hat{\Sigma}_i = (\hat{\sigma}_{ik\ell})$, with $\sigma_{ijk} = E_F Z_{ij} Z_{ik}$ and $\hat{\sigma}_{ik\ell} = m^{-1} \sum_{j=1}^m Z_{ik}^{(j)} Z_{i\ell}^{(j)}$, where $Z = (Z_{i\ell}^{(j)})$ is a sample from F^m , $i = 1, \dots, n$, $j = 1, \dots, m$, $\ell = 1, \dots, p$.

Let $\hat{\beta}$ be the estimator minimising $\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - X_{ij}^T \beta_i)^2$ subject to $\beta \in B$ where B is some subset of $\mathbb{R}^{n \times p}$.

Then, for any $\eta \in (0, 1)$,

$$(a) \max_{i=1, \dots, n} \|\Sigma_i - \hat{\Sigma}_i\|_{\infty} \leq \sqrt{\frac{2eV \log(n(p+1)^2)}{m\eta}},$$

$$(b) |L_F(\beta) - L_{\hat{F}}(\beta)| \leq \frac{1}{nm} \sqrt{\frac{2eV \log(n(p+1)^2)}{m\eta}} \left(n + \sum_{i=1}^n \|\beta_i\|_1^2 \right)$$

with probability at least $1 - \eta$.

Proof. Follows that of Theorem 1 in Greenshtein and Ritov (2004).

a) Let $\hat{\sigma}_{ik\ell} = \sigma_{ik\ell} + \epsilon_{ik\ell}$, $E_i = (\epsilon_{ik\ell})$. Then, under Assumption F and by Nemirovsky's inequality (see e.g. Lounici et al [10]),

$$\begin{aligned} P(\max_i \|\Sigma_i - \hat{\Sigma}_i\|_\infty > A) &\leq \frac{1}{A^2} E(\max_i \|\Sigma_i - \hat{\Sigma}_i\|_\infty^2) \\ &\leq \frac{2e \log(n(p+1)^2)}{mA^2} E(\max_{i=1, \dots, n} \max_{j, k=1, \dots, p+1} (Z_{ij} Z_{ik} - E(Z_{ij} Z_{ik}))^2) \\ &\leq \frac{2eV \log(n(p+1)^2)}{mA^2}. \end{aligned}$$

Taking $A = \sqrt{\frac{2eV \log(n(p+1)^2)}{m\eta}}$ proves the first part of the lemma.

b) By the definition of $\hat{\beta}$ and β_F^* ,

$$L_F(\hat{\beta}) - L_F(\beta_F^*) \geq 0, \quad L_{\hat{F}}(\hat{\beta}) - L_{\hat{F}}(\beta_F^*) \leq 0.$$

Hence,

$$\begin{aligned} 0 &\leq L_F(\hat{\beta}) - L_F(\beta_F^*) = L_F(\hat{\beta}) - L_{\hat{F}}(\hat{\beta}) \\ &\quad + L_{\hat{F}}(\hat{\beta}) - L_F(\hat{\beta}) + L_F(\hat{\beta}) - L_F(\beta_F^*) \\ &\leq 2 \sup_{\beta \in B_{n,p}} |L_F(\beta) - L_{\hat{F}}(\beta)|. \end{aligned}$$

Denote $\delta_i^\top = (-1, \beta_{i,1}, \dots, \beta_{i,p})$, then

$$L_F(\beta) = \frac{1}{nm} \sum_{i=1}^n \delta_i^\top \Sigma_{F,i} \delta_i,$$

where $\Sigma_{F,i} = (\sigma_{ijk})$ and $\sigma_{ijk} = E_F Z_{ij} Z_{ik}$. For the empirical distribution function \hat{F}_{mn} determined by a sample $Z_{il}^{(j)}$, $i = 1, \dots, n$, $j = 1, \dots, m$, $\ell = 1, \dots, p$, $\Sigma_{\hat{F},i} = (\hat{\sigma}_{ik\ell})$ and $\hat{\sigma}_{ik\ell} = \frac{1}{m} \sum_{j=1}^m Z_{ik}^{(j)} Z_{il}^{(j)}$.

Introduce matrix $\hat{\mathcal{E}}$ with $\hat{\mathcal{E}}_{j\ell} = A$. Hence, with probability at least $1 - \eta$,

$$\begin{aligned}
|L_F(\beta) - L_{\hat{F}}(\beta)| &= \left| \frac{1}{nm} \sum_{i=1}^n \delta_i^\top (\Sigma_{F,i} - \Sigma_{\hat{F}}, i) \delta_i \right| \\
&\leq \frac{1}{nm} \sum_{i=1}^n |\delta_i|^\top \hat{\mathcal{E}} |\delta_i| \\
&= \frac{1}{nm} \sqrt{\frac{2eV \log(n(p+1)^2)}{m\eta}} (n + \sum_{i=1}^n \|\beta_i\|_1^2).
\end{aligned}$$

□

A.2 Proofs of Section 2

Proof of Theorem 2.1. Note that by the definition of $\tilde{\beta}_i$ and (5).

$$\begin{aligned}
mnc_n + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 &\leq m \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq m \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{S}_i \tilde{\beta}_i + (\lambda_n + m\delta_n) \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq m \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{S}_i \tilde{\beta}_{i0} + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq m \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + (\lambda_n + m\delta_n) \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&= mnC_n + (\lambda_n + m\delta_n) \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2.
\end{aligned} \tag{13}$$

Comparing the LHS with the RHS of (13), noting that $m\delta_n \ll \lambda_n$:

$$\sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \leq mn \frac{C_n - c_n}{\lambda_n - m\delta_n} + \frac{\lambda_n + m\delta_n}{\lambda_n - m\delta_n} \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2.$$

By (5) and (6):

$$\begin{aligned}
\sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i &\leq \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{S}_i \tilde{\beta}_i + \delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{S}_i \tilde{\beta}_{i0} + \frac{\lambda_n}{m} \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 - \frac{\lambda_n}{m} \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 + \delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \left(\frac{\lambda_n}{m} + \delta_n\right) \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 - \left(\frac{\lambda_n}{m} - \delta_n\right) \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \left(\frac{\lambda_n}{m} + \delta_n\right) \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2.
\end{aligned} \tag{14}$$

The result follows. \square

Proof of Theorem 2.2. The proof is similar to the proof of Theorem 2.1. Similar to (13) we obtain:

$$\begin{aligned}
mnc_n + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^\alpha &\leq m \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^\alpha \\
&\leq m \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{S}_i \tilde{\beta}_i + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^\alpha + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq m \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{S}_i \tilde{\beta}_{i0} + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^\alpha + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&\leq m \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^\alpha + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\
&= mnc_n + \lambda_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^\alpha + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2.
\end{aligned} \tag{15}$$

That is,

$$\begin{aligned}
\sum_{i=1}^n (\lambda_n \|\tilde{\beta}_i\|_1^\alpha - m\delta_n \|\tilde{\beta}_i\|_1^2) &\leq \lambda_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^\alpha + m\delta_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 \\
&= \mathcal{O}(mn\delta_n).
\end{aligned} \tag{16}$$

It is easy to see that the maximum of $\sum_{i=1}^n \|\tilde{\beta}_i\|_1^2$ subject to the constraint (16) is achieved when $\|\tilde{\beta}_1\|_1^2 = \dots = \|\tilde{\beta}_n\|_1^2$. That is when $\|\tilde{\beta}_i\|_1^2$ solves $\lambda_n u^\alpha - m\delta_n u^2 = \mathcal{O}(m\delta_n)$. As $\lambda_n = \mathcal{O}(m\delta_m)$, the solution satisfies $u = \mathcal{O}(m\delta_n/\lambda_n)^{1/(\alpha-2)}$.

Hence we can conclude from (16)

$$\sum_{i=1}^n \|\tilde{\beta}_i\|_2^2 = \mathcal{O}(n(m\delta_n/\lambda_n)^{2/(\alpha-2)})$$

We now proceed similar to (14)

$$\begin{aligned} \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i &\leq \sum_{i=1}^n \tilde{\beta}_i^\top \tilde{S}_i \tilde{\beta}_i + \delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\ &\leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{S}_i \tilde{\beta}_{i0} + \frac{\lambda_n}{m} \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^\alpha - \frac{\lambda_n}{m} \sum_{i=1}^n \|\tilde{\beta}_i\|_1^\alpha + \delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\ &\leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \frac{\lambda_n}{m} \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^\alpha + \delta_n \sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 + \delta_n \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \\ &\leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \mathcal{O}_p(n(m/\lambda_n)^{2/(\alpha-2)} \delta_n^{\alpha/(\alpha-2)}), \end{aligned}$$

since $\lambda_n = \mathcal{O}(m\delta_m)$. □

Proof of Remark 2.1. If $m\delta_m/\lambda = o(1)$, then, following the proof of Theorem 2.2, the solution maximising $\sum_{i=1}^n \|\tilde{\beta}_i\|_1^2$ subject to the constraint (16) satisfies $\|\tilde{\beta}_i\|_1 = \mathcal{O}(1)$, and hence we have

$$\sum_{i=1}^n \tilde{\beta}_i^\top \tilde{\Sigma}_i \tilde{\beta}_i \leq \sum_{i=1}^n \tilde{\beta}_{i0}^\top \tilde{\Sigma}_i \tilde{\beta}_{i0} + \mathcal{O}_p(n\lambda_n/m + n\delta_n).$$

□

Proof of Theorem 2.3. The proof follows that of Lemma 3.1 in Lounici et al. [10].

We start with (a) and (b). Since $\hat{\beta}$ minimizes (6), then, $\forall \beta$

$$\sum_{i=1}^n \|Y_i - X_i^\top \hat{\beta}_i\|_2^2 + \lambda \sum_{i=1}^n \|\hat{\beta}_i\|_1^\alpha \leq \sum_{i=1}^n \|Y_i - X_i^\top \beta_i\|_2^2 + \lambda \sum_{i=1}^n \|\beta_i\|_1^\alpha,$$

and hence, for $Y_i = X_i^\top \beta_i + \varepsilon_i$,

$$\sum_{i=1}^n \|X_i^\top (\hat{\beta}_i - \beta_i)\|_2^2 \leq \sum_{i=1}^n \left[2\varepsilon_i^\top X_i^\top (\beta_i - \hat{\beta}_i) + \lambda(\|\beta_i\|_1^\alpha - \|\hat{\beta}_i\|_1^\alpha) \right].$$

Denote $V_{i\ell} = \sum_{j=1}^m x_{ij\ell} \varepsilon_{ij} \sim \mathcal{N}(0, m\sigma^2)$, and introduce event $\mathcal{A}_i = \bigcap_{\ell=1}^p \{|V_{i\ell}| \leq \mu\}$, for some $\mu > 0$. Then

$$\begin{aligned} P(\mathcal{A}_i^c) &\leq \sum_{\ell=1}^p P(|V_{i\ell}| > \mu) \\ &= \sum_{\ell=1}^p 2 \left[1 - \Phi \left\{ \mu / (\sigma \sqrt{m}) \right\} \right] \\ &\leq p \exp \left\{ -\mu^2 / (2m\sigma^2) \right\}. \end{aligned}$$

For $\mathcal{A} = \bigcap_{i=1}^n \mathcal{A}_i$, due to independence,

$$P(\mathcal{A}^c) = \sum_{i=1}^n P(\mathcal{A}_i^c) \leq pn \exp \left\{ -\mu^2 / (2m\sigma^2) \right\}.$$

Thus, if μ is large enough, $P(\mathcal{A}^c)$ is small, e.g., for $\mu = \sigma A (m \log(np))^{1/2}$, $A > \sqrt{2}$, we have $P(\mathcal{A}^c) \leq (np)^{1-A^2/2}$.

On event \mathcal{A} , for some $\nu > 0$,

$$\begin{aligned} &\sum_{i=1}^n \left[\|X_i(\hat{\beta}_i - \beta_i)\|_2^2 + \nu \|\beta_i - \hat{\beta}_i\|_1 \right] \\ &\leq \sum_{i=1}^n \left[2\mu \|\beta_i - \hat{\beta}_i\|_1 + \lambda(\|\beta_i\|_1^2 - \|\hat{\beta}_i\|_1^2) + \nu \|\beta_i - \hat{\beta}_i\|_1 \right] \\ &= \sum_{i=1}^n \sum_{j=1}^m \left[\alpha \lambda \max(\|\beta_i\|_1^{\alpha-1}, \|\hat{\beta}_i\|_1^{\alpha-1}) (|\beta_{ij}| - |\hat{\beta}_{ij}|) + (\nu + 2\mu) |\beta_{ij} - \hat{\beta}_{ij}| \right] \\ &\leq \sum_{i=1}^n \sum_{j=1}^m \left[\alpha \lambda \max(B^{\alpha-1}, \hat{B}^{\alpha-1}) (|\beta_{ij}| - |\hat{\beta}_{ij}|) + (\nu + 2\mu) |\beta_{ij} - \hat{\beta}_{ij}| \right], \end{aligned}$$

due to inequality $|x^\alpha - y^\alpha| \leq \alpha |x - y| \max(|x|^{\alpha-1}, |y|^{\alpha-1})$ which holds for $\alpha \geq 1$ and any x and y . To simplify the notation, denote $\mathcal{C} = \alpha \max(B^{\alpha-1}, \hat{B}^{\alpha-1})$.

Denote $J_i = J(\beta_i) = \{j : \beta_{ij} \neq 0\}$, $\mathcal{M}(\beta_i) = |J(\beta_i)|$. For each i and $j \in J(\beta_i)$, the expression in square brackets is bounded above by

$$[\lambda \mathcal{C} + \nu + 2\mu] |\beta_{ij} - \hat{\beta}_{ij}|,$$

and for $j \in J^c(\beta)$, the expression in square brackets is bounded above by 0, as long as $\nu + 2\mu \leq \lambda\mathcal{C}$:

$$-\lambda\mathcal{C}|\hat{\beta}_{ij}| + (\nu + 2\mu)|\hat{\beta}_{ij}| \leq 0.$$

This condition is satisfied if $\nu + 2\mu \leq \lambda\mathcal{C}$.

Hence, on \mathcal{A} , for $\nu + 2\mu \leq \lambda\mathcal{C}$,

$$\sum_{i=1}^n \left[\|X_i^\top(\hat{\beta}_i - \beta_i)\|_2^2 + \nu\|\beta_i - \hat{\beta}_i\|_1 \right] \leq \sum_{i=1}^n [\lambda\mathcal{C} + 2\mu + \nu] \|(\beta_i - \hat{\beta}_i)_{J_i}\|_1.$$

This implies that

$$\sum_{i=1}^n \|X_i(\hat{\beta}_i - \beta_i)\|_2^2 \leq [\lambda\mathcal{C} + \nu + 2\mu] \|(\beta - \hat{\beta})_J\|_1,$$

as well as that

$$\|\beta - \hat{\beta}\|_1 \leq \left[1 + \frac{2\mu}{\nu} + \frac{\lambda}{\nu}\mathcal{C} \right] \|(\beta - \hat{\beta})_J\|_1.$$

Take $\nu = \lambda\mathcal{C}/2$, hence we need to assume that $2\mu \leq \lambda\mathcal{C}/2$:

$$\begin{aligned} \sum_{i=1}^n \|X_i^\top(\hat{\beta}_i - \beta_i)\|_2^2 &\leq \left[\frac{3\lambda}{2}\mathcal{C} + 2\mu \right] \|(\beta - \hat{\beta})_J\|_1, \\ \|\beta - \hat{\beta}\|_1 &\leq \left[3 + \frac{4\mu}{\lambda\mathcal{C}} \right] \|(\beta - \hat{\beta})_J\|_1 \leq 4\|(\beta - \hat{\beta})_J\|_1. \end{aligned} \tag{17}$$

which implies

$$\|(\beta - \hat{\beta})_{J^c}\|_1 \leq 3\|(\beta - \hat{\beta})_J\|_1.$$

Due to the generalized restricted eigenvalue assumption $\text{RE}_1(s, 3, \kappa)$, $\|X^\top(\beta - \hat{\beta})\|_2 \geq \kappa\sqrt{m}\|(\beta - \hat{\beta})_J\|_2$, and hence, using (17),

$$\begin{aligned} \|X^\top(\hat{\beta} - \beta)\|_2^2 &\leq \left[\frac{3\lambda}{2}\mathcal{C} + 2\mu \right] \sqrt{n\mathcal{M}(\beta)} \|(\hat{\beta} - \beta)_J\|_2 \\ &\leq \left[\frac{3\lambda}{2}\mathcal{C} + 2\mu \right] \frac{\sqrt{n\mathcal{M}(\beta)}}{\kappa\sqrt{m}} \|X^\top(\hat{\beta} - \beta)\|_2, \end{aligned}$$

where $\mathcal{M}(\beta) = \max_i \mathcal{M}(\beta_i)$, implying that

$$\|X^\top(\hat{\beta} - \beta)\|_2 \leq \left[\frac{3\lambda}{2}\mathcal{C} + 2\mu \right] \frac{\sqrt{n\mathcal{M}(\beta)}}{\kappa\sqrt{m}}$$

$$= \frac{\sqrt{n\mathcal{M}(\beta)}}{\kappa\sqrt{m}} \left[\frac{3\lambda}{2}\mathcal{C} + 2A\sigma\sqrt{m\log(np)} \right].$$

Also,

$$\begin{aligned} \|\beta - \hat{\beta}\|_1 &\leq 4\|(\beta - \hat{\beta})_J\|_1 \leq 4\frac{\sqrt{n\mathcal{M}(\beta)}}{\sqrt{m\kappa}}\|X^\top(\beta - \hat{\beta})\|_2 \\ &\leq \frac{4n\mathcal{M}(\beta)}{m\kappa^2} \left[\frac{3\lambda}{2}\mathcal{C} + 2A\sigma\sqrt{m\log(np)} \right]. \end{aligned}$$

Hence, a) and b) of the theorem are proved.

(c) For $i, \ell: \hat{\beta}_{i\ell} \neq 0$, we have

$$2X_{i,\ell}(Y_i - X_i^\top \hat{\beta}_i) = \lambda\alpha \text{sgn}(\hat{\beta}_{i\ell})\|\hat{\beta}_i\|_1^{\alpha-1},$$

Hence,

$$\begin{aligned} \sum_{\ell: \hat{\beta}_{i\ell} \neq 0} \|X_{i,\ell}X_i^\top(\beta_i - \hat{\beta}_i)\|_2^2 &\geq \sum_{\ell: \hat{\beta}_{i\ell} \neq 0} \left(\|X_{i,\ell}(Y_i - X_i^\top \hat{\beta}_i)\|_2 - \|X_{i,\ell}(Y_i - X_i^\top \beta_i)\|_2 \right)^2 \\ &\geq \sum_{\ell: \hat{\beta}_{i\ell} \neq 0} \left(\alpha\lambda\|\hat{\beta}_i\|_1^{\alpha-1}/2 - \mu \right)^2 \\ &= \mathcal{M}(\hat{\beta}_i)(\alpha\lambda\|\hat{\beta}_i\|_1^{\alpha-1}/2 - \mu)^2. \end{aligned}$$

Thus,

$$\mathcal{M}(\hat{\beta}_i) \leq \|X_i(\beta_i - \hat{\beta}_i)\|_2^2 \frac{m\phi_{i,\max}}{\left(\lambda\alpha\|\hat{\beta}_i\|_1^{\alpha-1}/2 - \mu\right)^2}.$$

Theorem is proved. \square

Proof of Theorem 2.4. To satisfy the conditions of Theorem 2.3, we can take $B = b$ and $\lambda = \frac{4A\sigma}{\alpha b^{\alpha-1}}\sqrt{m\log(np)}$.

Thus, by Lemma A.1,

$$\frac{\lambda}{m\delta_n} = \frac{4A\sigma}{\alpha b^{\alpha-1}} \sqrt{\frac{\log(np)}{m}} \sqrt{\frac{m\eta}{2eV\log(n(p+1)^2)}} = C \frac{\sqrt{\eta}}{\alpha b^{\alpha-1}} \leq C_1,$$

hence assumption $\lambda = \mathcal{O}(m\delta_n)$ of Theorem 2.2 is satisfied.

Hence, from the proof of Theorem 2.3, it follows that

$$\|\hat{\beta}_i\|_1 = \mathcal{O}\left(\left(m\delta_n/\lambda_n\right)^{1/(\alpha-2)}\right) = \mathcal{O}\left(\left(\frac{b^{\alpha-1}}{\sqrt{\eta}}\right)^{1/(\alpha-2)}\right).$$

Hence, we can take $B = b$ and $\hat{B} = C \left(\frac{b^{\alpha-1}}{\sqrt{\eta}} \right)^{1/(\alpha-2)}$ for some $C > 0$, and apply Theorem 2.3. Then $\max(1, \hat{B}/B)$ is bounded by

$$\max \left[1, C \frac{b^{(\alpha-1)/(\alpha-2)-1}}{\eta^{1/(2(\alpha-2))}} \right] = \max \left[1, C \frac{b^{1/(\alpha-2)}}{\eta^{1/(2(\alpha-2))}} \right] = \left(\frac{Cb}{\sqrt{\eta}} \right)^{1/(\alpha-2)},$$

since $\frac{Cb}{\sqrt{\eta}} \geq C_2 \frac{\eta^{1/(2(\alpha-1))}}{\sqrt{\eta}} \geq C_2 \eta^{-(\alpha-2)/(2(\alpha-1))}$ is large for small η .
Hence,

$$\begin{aligned} & \frac{3\alpha\lambda}{2\sqrt{m}} \max(B^{\alpha-1}, \hat{B}^{\alpha-1}) + 2A\sigma\sqrt{\log(np)} \\ & \leq 6AC\sigma\sqrt{\log(np)} \frac{b^{(\alpha-1)/(\alpha-2)}}{\eta^{(\alpha-1)/(2(\alpha-2))}} + 2A\sigma\sqrt{\log(np)} \\ & = 2A\sigma\sqrt{\log(np)} \left[3C \left(\frac{b}{\sqrt{\eta}} \right)^{(\alpha-1)/(\alpha-2)} + 1 \right], \end{aligned}$$

and, applying Theorem 2.3, we obtain (a) and (b).

c) Apply c) in Theorem 2.3, summing over $i \in \mathcal{I}$:

$$\begin{aligned} \sum_{i \in \mathcal{I}} \mathcal{M}(\hat{\beta}_i) & \leq \|X^\top(\beta - \hat{\beta})\|_2^2 \frac{m\phi_{\max}}{(\mu\delta)^2} \\ & \leq \frac{4sn\phi_{\max}}{\kappa^2\delta^2} \left[1 + 3C \left(\frac{b}{\sqrt{\eta}} \right)^{(\alpha-1)/(\alpha-2)} \right]^2. \end{aligned}$$

□

A.3 Proofs of Section 4

Proof of Lemma 4.1. Let $\mathcal{B} = \sum_{\xi=1}^k \alpha_\xi \beta_\xi^* \mathbf{b}_\xi^{*\top}$ be the spectral decomposition of \mathcal{B} , where $\beta_1^*, \dots, \beta_k^*$ are orthonormal \mathbb{R}^p vectors, $\mathbf{b}_1^*, \dots, \mathbf{b}_k^*$ are orthonormal \mathbb{R}^n vectors, $\alpha_1, \dots, \alpha_k \geq 0$, and $k = \min\{p, n\}$. Clearly $\|\mathcal{B}\|_1 = \sum_{\xi=1}^k \alpha_\xi$. Let $U = \sum_{\xi=1}^k e_\xi \beta_\xi^{*\top}$ where e_1, \dots, e_p is the natural basis of \mathbb{R}^n . Then

$$\|UB\|_{2,1} = \left\| \sum_{\xi=1}^k \alpha_\xi e_\xi \beta_\xi^{*\top} \right\|_{2,1} = \sum_{\xi=1}^k \alpha_\xi = \|\mathcal{B}\|_1.$$

Let $\mathcal{B} = \sum_{\xi=1}^k e_{\xi} \mathbf{b}_{\xi}^{\top}$ where $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k$ are orthogonal, and let U be a unitary matrix. Then by Schwarz inequality

$$\begin{aligned}
\|\mathcal{B}\|_{2,1} &= \sum_{j=1}^p \|\mathbf{b}_j\| \\
&= \sum_{i=1}^p \sum_{j=1}^p U_{ij}^2 \|\mathbf{b}_j\| && \text{since } \sum_{i=1}^p U_{ij}^2 = 1 \\
&\leq \sum_{i=1}^p \sqrt{\sum_{j=1}^p U_{ij}^2 \|\mathbf{b}_j\|^2} \sqrt{\sum_{j=1}^p U_{ij}^2} && \text{by Schwarz inequality} \\
&= \sum_{i=1}^p \sqrt{\sum_{j=1}^p U_{ij}^2 \|\mathbf{b}_j\|^2} && \text{since } \sum_{j=1}^p U_{ij}^2 = 1 \\
&= \|U\mathcal{B}\|_{2,1}
\end{aligned}$$

which completes the proof of the (i).

Now, consider the U defined as above for the solution of (12). Let \tilde{X}_i be the design matrices $\tilde{\mathcal{B}}$ be the solution expressed in this basis. By the first part of the lemma $\|\tilde{\mathcal{B}}\|_1 = \|\tilde{\mathcal{B}}\|_{2,1}$. Suppose there is a matrix $\mathcal{B} \neq \tilde{\mathcal{B}}$ which minimizes the group lasso penalty. Hence

$$\begin{aligned}
\sum_{i=1}^n \|Y_i - \tilde{X}_i \beta_i\|^2 + \lambda \|\mathcal{B}\|_1 &\leq \sum_{i=1}^n \|Y_i - \tilde{X}_i \beta_i\|^2 + \lambda \|\mathcal{B}\|_{2,1} \\
&< \sum_{i=1}^n \|Y_i - \tilde{X}_i \tilde{\beta}_i\|^2 + \lambda \|\tilde{\mathcal{B}}\|_{2,1} \\
&= \sum_{i=1}^n \|Y_i - \tilde{X}_i \tilde{\beta}_i\|^2 + \lambda \|\tilde{\mathcal{B}}\|_1,
\end{aligned}$$

contradiction since $\tilde{\mathcal{B}}$ minimized (12). Part (ii) is proved. \square

Proof of Theorem 4.2 . Let $A = \sum_{i=1}^n \hat{\beta}_i \hat{\beta}_i^{\top} = \hat{\mathcal{B}} \hat{\mathcal{B}}^{\top}$ be of rank $s \leq p < n$, and hence the spectral decomposition of $\hat{\mathcal{B}}$ can be written as $\hat{\mathcal{B}} = \sum_{\xi=1}^s \alpha_{\xi} \beta_{\xi}^* \mathbf{b}_{\xi}^{*\top}$, where $\beta_1^*, \dots, \beta_s^* \in \mathbb{R}^p$ are orthonormal, and so are $\mathbf{b}_1^*, \dots, \mathbf{b}_s^* \in \mathbb{R}^n$. Hence, the rotation U leading to a sparse representation $U\hat{\mathcal{B}}$ (with s non-zero rows) is given by $U = \sum_{\xi=1}^s e_{\xi} \beta_{\xi}^{*\top}$, where e_1, \dots, e_p is the natural basis of \mathbb{R}^p . Another way to write the rotation matrix is $U = (\beta_1^{*\top}, \dots, \beta_s^{*\top}, \mathbf{0}^{\top}, \dots, \mathbf{0}^{\top})^{\top}$. Denote by U_S the non-zero $s \times p$ -dimensional submatrix $(\beta_1^{*\top}, \dots, \beta_s^{*\top})^{\top}$.

Let $A(t) = A + t(\tilde{\beta}\hat{\beta}_i^\top + \hat{\beta}_i\tilde{\beta}^\top) + t^2\tilde{\beta}\tilde{\beta}^\top$ for some fixed i , with $\tilde{\beta} \in \text{span}\{\hat{\beta}_1, \dots, \hat{\beta}_n\} = \text{span}\{\hat{\beta}_1^*, \dots, \hat{\beta}_s^*\}$.

If $(x_k(t), c_k(t))$ is an eigen-pair of $A(t)$, then taking the derivative of $x_i^\top x_i = 1$ yields $x_i^\top \dot{x}_i = 0$, and trivially, since x_i is an eigenvector, also $x_i^\top A \dot{x}_i = 0$. Here and the first and second derivative, respectively, according to t . Also, we have

$$\begin{aligned} x_k(t) &= x_k + tu_k + o(t) \\ c_k(t) &= c_k + t\nu_k + o(t) \end{aligned}$$

and

$$\left(A + t(\tilde{\beta}\hat{\beta}_i^\top + \hat{\beta}_i\tilde{\beta}^\top)\right)(x_k + tu_k) = (c_k + t\nu_k)(x_k + tu_k) + o(t),$$

where $u_k \perp x_k$.

Equating the $\mathcal{O}(t)$ terms obtain

$$Au_k + (\tilde{\beta}\hat{\beta}_i^\top + \hat{\beta}_i\tilde{\beta}^\top)x_k = c_k u_k + \nu_k x_k.$$

Take now the inner product of both sides with x_k to obtain that

$$\nu_k = 2(\tilde{\beta}^\top x_k)(x_k^\top \hat{\beta}_i). \quad (18)$$

Note that the null space of $A(t)$ does not depend on t . Hence, if we call $\psi(\mathcal{B}) = \|\mathcal{B}\|_1$,

$$\begin{aligned} \frac{\partial}{\partial t} \psi(A(t))|_{t=0} &= \sum_{c_k > 0} \frac{\partial}{\partial t} c_k^{1/2}(t)|_{t=0} \\ &= \frac{1}{2} \sum_{c_k > 0} \frac{\nu_k}{c_k^{1/2}} \\ &= \tilde{\beta}^\top \sum_{c_k > 0} c_k^{-1/2} x_k x_k^\top \hat{\beta}_i \\ &= \tilde{\beta}^\top A^{+1/2} \hat{\beta}_i = \tilde{\beta}^\top (\hat{\mathcal{B}}\hat{\mathcal{B}}^\top)^{+1/2} \hat{\beta}_i \\ &= \tilde{\beta}^\top U_S^\top (U_S \hat{\mathcal{B}}\hat{\mathcal{B}}^\top U_S^\top)^{-1/2} U_S \hat{\beta}_i, \end{aligned}$$

where $A^{+1/2}$ is the generalized inverse of $A^{1/2}$.

Taking, therefore, the derivative of the target function with respect to $\hat{\beta}_i$ in the directions of $\tilde{\beta} \in \text{span}\{\hat{\beta}_1, \dots, \hat{\beta}_n\}$ (e.g., in the directions $\tilde{\beta} = \beta_\xi^*$, $\xi = 1, \dots, s$) gives

$$0 = (\beta_\xi^*)^\top (-2X_i^\top (Y_i - X_i \hat{\beta}_i) + \lambda(\hat{\mathcal{B}}\hat{\mathcal{B}}^\top)^{+1/2} \hat{\beta}_i), \quad \text{or, equivalently,}$$

$$\mathbf{0} = U_S(2X_i^\top(Y_i - X_i\hat{\beta}_i) - \lambda(\hat{\mathcal{B}}\hat{\mathcal{B}}^\top)^{+1/2}\hat{\beta}_i).$$

Let $R = (r_1, \dots, r_p)^\top$ be the matrix of projected residuals:

$$R_{\ell i} = \sum_{j=1}^m x_{ij\ell}(y_{ij} - x_{ij}^\top\hat{\beta}_i), \quad \ell = 1, \dots, p; \quad i = 1, \dots, n.$$

Then

$$U_S R = \frac{\lambda}{2} U_S (\hat{\mathcal{B}}\hat{\mathcal{B}}^\top)^{+1/2} \hat{\mathcal{B}}.$$

Consider again the general expansion $\hat{\mathcal{B}} = \sum_{\xi=1}^{p \wedge n} \alpha_\xi \beta_\xi^* \mathbf{b}_\xi^{*\top}$. Then $\|\hat{\mathcal{B}}\|_1 = \sum_{\xi=1}^{p \wedge n} |\alpha_\xi|$. Taking the derivative of the sum of squares part of the target function with respect to α_ξ we get

$$\sum_{i=1}^n \mathbf{b}_{\xi i}^* \beta_\xi^{*\top} X_i^\top (Y_i - X_i \hat{\beta}_i) = \beta_\xi^{*\top} R \mathbf{b}_\xi^*.$$

Considering the sub-gradient of the target function we obtain that $|\beta_\xi^{*\top} R \mathbf{b}_\xi^*| \leq \lambda/2$, and $\alpha_\xi = 0$ in case of strict inequality. \square

Proof of Theorem 4.3. (a) and (b) Similarly to the proof of Theorem 2.3, we have

$$\|Y - X^\top \hat{\mathcal{B}}\|_2^2 = \|Y - X^\top \mathcal{B}\|_2^2 + 2 \sum_{ij} \varepsilon_{ij} x_{ij}^\top (\beta_i - \hat{\beta}_i).$$

The last term can be bounded with high probability. Introduce matrix M with independent columns $M_i = X_i \varepsilon_i \sim \mathcal{N}_p(\mathbf{0}, m\sigma^2 I_p)$, $i = 1, \dots, n$, since $\sum_j x_{ij}^2 = m$. Denote q -Schatten norm by $\|\cdot\|_q$. Using the Cauchy-Swartz inequality and the equivalence between ℓ_2 (Frobenius) and Schatten with $q = 2$ norms, we obtain:

$$\begin{aligned} \left| \sum_{ij} \varepsilon_{ij} x_{ij}^\top (\beta_i - \hat{\beta}_i) \right| &= \left| \sum_{i\ell} M_{i\ell} (\beta_{i\ell} - \hat{\beta}_{i\ell}) \right| \leq \|\mathcal{B} - \hat{\mathcal{B}}\|_2 \|M\|_2 = \|\mathcal{B} - \hat{\mathcal{B}}\|_2 \|M\|_2 \\ &\leq \|\mathcal{B} - \hat{\mathcal{B}}\|_1 \|M\|_2. \end{aligned}$$

Now, $\|M\|_2^2 \sim m\sigma^2 \chi_{np}^2$ hence it can be bounded by $B^2 = m\sigma^2(np + c)$ (Lemma A.1, Lounici et al. [10]) with probability at least $1 - \exp(-\frac{1}{8} \min(c, c^2/(np)))$. Denote this event by \mathcal{A} . Hence, we need to choose c such that $c/\sqrt{np} \rightarrow \infty$.

For example, we can take $c = Anp$ with $A > 1$, then $B = \sigma\sqrt{(1+A)mnp}$, and, since $\min(Anp, A^2np) = Anp$, the probability is at least $1 - e^{-Anp/2}$.

Denote by V the subspace of \mathbb{R}^p corresponding to the union of subspaces where the eigenvalues of $\mathcal{B}\mathcal{B}^\top$ are non-zero, and by P_V the projection on that space. Then, $\mathbb{R}^p = V \oplus V^c$ and $\dim(V) = \text{rank}(\mathcal{B}) \leq s$.

Hence, adding $\lambda_2\|\mathcal{B} - \hat{\mathcal{B}}\|_1$ to both sides, we have that on \mathcal{A} ,

$$\begin{aligned} \|X^\top(\mathcal{B} - \hat{\mathcal{B}})\|_2^2 + \lambda_2\|\mathcal{B} - \hat{\mathcal{B}}\|_1 &\leq \lambda\|\mathcal{B}\|_1 - \lambda\|\hat{\mathcal{B}}\|_1 + (2B + \lambda_2)\|\mathcal{B} - \hat{\mathcal{B}}\|_1 \\ &\leq \lambda\|P_V\mathcal{B}\|_1 - \lambda\text{trace}(P_V|\hat{\mathcal{B}}| + (I - P_V)|\hat{\mathcal{B}}|) \\ &\quad + (2B + \lambda_2)\|P_V(\mathcal{B} - \hat{\mathcal{B}})\|_1 \\ &\quad + (2B + \lambda_2)\|(I - P_V)(\mathcal{B} - \hat{\mathcal{B}})\|_1 \\ &\leq \lambda\text{trace}(|P_V\mathcal{B}|) - \lambda\text{trace}(P_V|\hat{\mathcal{B}}|) + (2B + \lambda_2)\text{trace}(|P_V(\mathcal{B} - \hat{\mathcal{B}})|) \\ &\quad + (2B + \lambda_2)\text{trace}(|(I - P_V)\hat{\mathcal{B}}|) - \lambda\text{trace}((I - P_V)|\hat{\mathcal{B}}|) \\ &\leq (\lambda + 2B + \lambda_2)\text{trace}(|P_V(\mathcal{B} - \hat{\mathcal{B}})|), \end{aligned}$$

if $\lambda \geq 2B + \lambda_2$, since $\text{trace}(|P_V\hat{\mathcal{B}}|) = \text{trace}(|P_V|\hat{\mathcal{B}}|) = \text{trace}(P_V|\hat{\mathcal{B}}|)$. Here $|A| = (AA^\top)^{1/2}$. We can take, e.g. $\lambda_2 = 2B = \lambda/2$, implying that $\lambda = 4\sigma\sqrt{(1+A)mnp}$.

Hence, we have that $\frac{\lambda}{2}\|\mathcal{B} - \hat{\mathcal{B}}\|_1 \leq 2\lambda\|P_V(\mathcal{B} - \hat{\mathcal{B}})\|_1$, i.e. $\|(I - P_V)(\mathcal{B} - \hat{\mathcal{B}})\|_1 \leq 3\lambda\|P_V(\mathcal{B} - \hat{\mathcal{B}})\|_1$. Thus, applying RE2($s, 3, \kappa$), $\text{rank}(\mathcal{B}) \leq s$, we have that

$$\begin{aligned} \|X^\top(\beta - \hat{\beta})\|_2^2 &\leq 2\lambda\|P_V(\mathcal{B} - \hat{\mathcal{B}})\|_1 \leq 2\lambda\sqrt{s}\|P_V(\mathcal{B} - \hat{\mathcal{B}})\|_2 \\ &= 2\lambda\sqrt{s}\|P_V(\mathcal{B} - \hat{\mathcal{B}})\|_2 \leq \frac{2\lambda\sqrt{s}}{\kappa\sqrt{m}}\|X^\top(\beta - \hat{\beta})\|_2 \end{aligned}$$

hence

$$\|X^\top(\beta - \hat{\beta})\|_2 \leq \frac{2\lambda\sqrt{s}}{\kappa\sqrt{m}}.$$

Using this and the RE2 assumption,

$$\|\mathcal{B} - \hat{\mathcal{B}}\|_1 \leq 4\|P_V(\mathcal{B} - \hat{\mathcal{B}})\|_1 \leq \frac{4\sqrt{s}}{\kappa\sqrt{m}}\|X^\top(\beta - \hat{\beta})\|_2 \leq \frac{8\lambda s}{\kappa^2 m}.$$

Substituting the value of λ , we obtain the results.

(c) Since $\hat{\gamma}_i = \hat{U}\hat{\beta}_i$ are the solution of group lasso problem with design matrices $\tilde{X}_i = \hat{U}X_i$, for $\ell \in J(\hat{\gamma})$: $\|\hat{\gamma}_{\cdot\ell}\|_2 \neq 0$, $\hat{\gamma}_{i\ell}$ satisfies the following equations;

$$2\tilde{X}_{i\cdot\ell}^\top(Y_i - X_i\hat{\beta}_i) = \lambda \frac{\hat{\gamma}_{i\ell}}{\|\hat{\gamma}_{\cdot\ell}\|_2}$$

(see also Theorem 4.2).

Hence,

$$\sum_{i=1}^n \left(\tilde{X}_{i\ell}^\top (Y_i - X_i \hat{\beta}_i) \right)^2 = \frac{\lambda^2}{4}.$$

On one hand, for $\ell \in J(\hat{\gamma})$,

$$\begin{aligned} \left[\sum_{i=1}^n \left(\tilde{X}_{i\ell} X_i^\top (\hat{\beta}_i - \beta_i) \right)^2 \right]^{1/2} &\geq \left[\sum_{i=1}^n \left(\tilde{X}_{i\ell} (Y_i - X_i^\top \hat{\beta}_i) \right)^2 \right]^{1/2} \\ &\quad - \left[\sum_{i=1}^n \left(\tilde{X}_{i\ell} (Y_i - X_i^\top \beta_i) \right)^2 \right]^{1/2} \\ &= \frac{\lambda}{2} - \left(\sum_{i=1}^n (U_\ell X_i \varepsilon_i)^2 \right)^{1/2}. \end{aligned}$$

On event \mathcal{A} ,

$$\sum_{i=1}^n (U_\ell X_i \varepsilon_i)^2 = \sum_{i=1}^n (U_\ell M_i)^2 \leq \sum_{i=1}^n \|U_\ell\|_2^2 \|M_i\|^2 = \|M\|_2^2 \leq B^2 = (\lambda/4)^2.$$

Summing over $\ell \in J(\hat{\gamma})$, we have

$$\sum_{\ell \in J(\hat{\gamma})} \sum_{i=1}^n \left(\tilde{X}_{i\ell} X_i^\top (\hat{\beta}_i - \beta_i) \right)^2 \geq \mathcal{M}(\hat{\gamma}) \left(\frac{\lambda}{2} - \frac{\lambda}{4} \right)^2 = \mathcal{M}(\hat{\gamma}) \frac{\lambda^2}{16}.$$

On the other hand,

$$\begin{aligned} \sum_{\ell=1}^s \sum_{i=1}^n \left(\tilde{X}_{i\ell} X_i^\top (\hat{\beta}_i - \beta_i) \right)^2 &\leq \sum_{i=1}^n \|\tilde{X}_i X_i^\top (\hat{\beta}_i - \beta_i)\|_2^2 = \sum_{i=1}^n \|X_i X_i^\top (\hat{\beta}_i - \beta_i)\|_2^2 \\ &\leq m \phi_{\max} \|X^\top (\hat{\mathcal{B}} - \mathcal{B})\|_2^2. \end{aligned}$$

Since $\text{rank}(\hat{\mathcal{B}}) = \mathcal{M}(\hat{\gamma})$,

$$\text{rank}(\hat{\mathcal{B}}) \leq \frac{m \phi_{\max} \|X^\top (\hat{\mathcal{B}} - \mathcal{B})\|_2^2}{(\lambda/4)^2} = \frac{16m \phi_{\max}}{\lambda^2} \frac{4\lambda^2 s}{m\kappa^2} = s \frac{64\phi_{\max}}{\kappa^2}.$$

□

Proof. of Theorem 4.4.

Using Lemma A.1, with probability at least $1 - \eta$,

$$|L_F(\beta) - L_{\hat{F}}(\beta)| \leq \frac{1}{nm} \sqrt{\frac{4eV \log(np)}{m\eta}} \left(n + \sum_{i=1}^n \|\beta_i\|_1^2 \right),$$

since $n > 1$. Note that if $n = 1$, it is sufficient to replace p by $p + 1$ under the logarithm.

In our case, the estimators are in set $B_{n,p}$. If $\sum_{i=1}^n \beta_i \beta_i^\top = U^\top \Lambda U$ is the spectral decomposition, and $\gamma_i = U \beta_i$, $\Lambda_{kk} = \|\gamma_{\cdot k}\|_2^2$, $\gamma_{\cdot k}$ are orthogonal, hence

$$\text{trace} \left\{ \sum_{i=1}^n \beta_i \beta_i^\top \right\}^{1/2} = \sum_{k=1}^p \|\gamma_{\cdot k}\|_2.$$

Thus, we need to bound $\sum_{i=1}^n \|\beta_i\|_1^2$ in terms of $\sum_{k=1}^p \|\gamma_{\cdot k}\|_2$.

$$\begin{aligned} \sum_{i=1}^n \|\beta_i\|_1^2 &\leq \sum_{i=1}^n M(\beta_i) \|\beta_i\|_2^2 \\ &= \max_i M(\beta_i) \sum_{i=1}^n \|\gamma_i\|_2^2 \\ &= \max_i M(\beta_i) \sum_{\ell=1}^p \|\gamma_{\cdot \ell}\|_2^2 \\ &\leq 2 \max_i M(\beta_i) \left(\sum_{\ell=1}^p \|\gamma_{\cdot \ell}\|_2 \right)^2 \\ &\leq \max_i M(\beta_i) b^2, \end{aligned}$$

since $\sum_{\ell=1}^p \|\gamma_{\cdot \ell}\|_2 \leq b$.

Hence, with probability at least $1 - \eta$,

$$\sup_{F \in \mathcal{F}} P_F \left(L_F(\hat{\beta}) - L_F(\beta_F^*) \right) \leq 2 \left(\frac{1}{m} + \frac{\max_i M(\beta_i) b^2}{nm} \right) \sqrt{\frac{4eV \log(np)}{m\eta}}.$$

Note that we can use p instead of $\max_i M(\beta_i)$. The theorem is proved. \square

References

- [1] F. Bach. Consistency of trace norm minimization. *The Journal of Machine Learning Research*, 2008.
- [2] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.
- [3] L.D. Brown and E. Greenshtein. Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Annals of Statistics*, 37:1685–1704, 2009.
- [4] E. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, to appear, 2009.
- [5] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, 10:971–988, 2004.
- [6] E. Greenshtein and Y. Ritov. Asymptotic efficiency of simple decisions for the compound decision problem. *The 3rd Lehmann Symposium, IMS Lecture-Notes Monograph series. J. Rojo, editor*, 1:xxx=xxx, 2008.
- [7] Eitan Greenshtein, Junyong Park, and Ya’acov Ritov. Estimating the mean of high valued observations in high dimensions. *Journal of Statistical Theory and Practice*, 2:407–418, 2008.
- [8] Zhang C. H. Compound decision theory and empirical bayes methods. *Annals of Statistics*, 31:379–390, 2003.
- [9] I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29:295–327, 2001.
- [10] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468*, 2009.
- [11] H. Robbins. Asymptotically subminimax solutions of compound decision problems. *Proc. Second Berkeley Symp. Math. Statist. Probab.*, 1:131–148, 1951.
- [12] H. Robbins. An empirical bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1:157–163, 1956.
- [13] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267–288, 1996.

- [14] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2006.
- [15] C. H. Zhang. General empirical bayes wavelet methods and exactly adaptive minimax estimation. *Annals of Statistics*, 33:54–100, 2005.

