Bayesian Perspectives on Sparse Empirical Bayes Analysis (SEBA)

Natalia Bochkina & Ya'acov Ritov

February 27, 2010

Abstract

We consider a joint processing of n independent similar 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 $y_{i1} = x_{i1}^\mathsf{T}\beta_i + \varepsilon_{i1}$, $y_{i1} \in \mathbb{R}$, $x_{i1} \in \mathbb{R}^p$, and $\varepsilon_{i1} \sim N(0, \sigma^2)$, say. The dimension p is large enough so that the empirical risk minimizer is not feasible. We consider, from a Bayesian point of view, three possible extensions of the lasso. Each of the three estimators, the lassoes, the group lasso, and the RING lasso, utilizes different assumptions on the relation between the n vectors $\beta_1, \dots, \beta_n$.

"... 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

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

or in a standard vector form

$$Y_i = X_i^\mathsf{T}\beta_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $\beta_i \in \mathbb{R}^p$. The matrix $X_i \in \mathbb{R}^{m\times p}$ is either deterministic fixed design matrix, or a sample of m independent $\mathbb{R}^p$ random vectors. Finally, ε_{ij}, $i = 1, \dots, n,\ j = 1, \dots, m$ are (at least uncorrelated with the xs), 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. Let $\mathcal{B}$ be the matrix $(\beta_1, \ldots, \beta_n)$.

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,

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

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

This approach is in the spirit of the empirical Bayes (compound decision) approach. Note however that the term "empirical Bayes" has a few other meanings in the literature), cf, [9, 10, 7]. The empirical Bayes approach to sparsity was considered before, e.g., [13, 3, 4, 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 [11]), and the methods we consider are possible extensions.

We will make use of the following notation, introducing the $l_{p,q}$ norm of matrices and sets z of vectors:

Definition 1.1 *For a matrix A, $\|A\|_{p,q} = \left(\sum_i \left(\sum_j A_{ij}^p\right)^{q/p}\right)^{1/q}$. If $z_1, \ldots, z_n$, is a collection of vectors, not necessarily of the same length, z_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, J_i$, then $||\{z_1, \ldots, z_n\}||_{p,q} = \left[\sum_{i=1}^n \left(\sum_{j \in J_i} |z_{ij}|^p\right)^{q/p}\right]^{1/q}$.*

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

Within the framework of the compound decision theory, we can have different scenarios. The first one is that the n groups are considered as repeated similar models for p variables, and the aim is to choose the variables

that are useful for all models. The relevant variation of the lasso procedure in this case is the group lasso introduced by Yuan and Lin [12]:

$$\hat{\mathcal{B}} = \arg\min_{\mathcal{B}} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - x_{ij}^{\mathsf{T}} \beta_i)^2 + \lambda \|\mathcal{B}^{\mathsf{T}}\|_{2,1}. \tag{3}$$

Yuan and Lin 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. [8].

Another possible scenario is where 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:

$$\hat{\mathcal{B}} = \arg\min_{\mathcal{B}} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - x_{ij}^{\mathsf{T}} \beta_i)^2 + \lambda \|\mathcal{B}\|_{1,\alpha}, \tag{4}$$

which we refer to as "lassoes" can be more natural than the simple lasso. The standard choice is $\alpha = 1$, but we believe that $\alpha > 4$ is, in fact, more consistent with a Bayesian point of view.

If we compare (4) to (3) we can see the difference between the prior assumptions of the lassoes and the group lasso. The basic elements of the lassoes are the β_i's vector, and we assume a priori that each of them is sparse. On the other hand, the basic elements of the group lasso are the variables, and we assume a priori that most of them do not contribute significantly to any of the regression equation.

We shall also consider a third situation where there is a sparse representation in some unknown basis, but assumed common to the n groups. The standard notion of sparsity, as captured by the ℓ_0 norm, or by the standard lasso, the lassoes, and the group lasso, is basis dependent. 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.

The rotation invariant group (RING) lasso was 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 ([2]). The corresponding penalty is the trace norm (or Schatten norm with $p = 1$) of the matrix $\mathcal{B}$, which finds the rotation that gives the best sparse representation of all vectors instantaneously.

The aim is to discuss the Bayesian interpretation of the three lasso extensions to the compound decision problem setting. Since the lassoes method, to our knowledge, has not been considered previously, we also present some theoretical results for it such as sparsity oracle inequalities and the persistency analysis.

The chapter is organised as follows. In Section 2 we introduce the lassoes method, discuss the Bayesian perspective, perform the persistency analysis and give the sparsity oracle inequalities. Section 3 is devoted to a Bayesian perspective on group lasso and Section 4 - to a Bayesian perspective on RING lasso. All the proofs are given in the Appendix.

2 The lassoes procedure

2.1 Persistency and Bayesian interpretation

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 $\beta = (\beta_1^\mathsf{T}, \dots, \beta_n^\mathsf{T})^\mathsf{T}$, and a block diagonal design matrix X. This solution, which corresponds to the solution of (4) with $\alpha = 1$, imposes very little on the similarity among $\beta_1, \dots, \beta_n$. The lassoes procedure discussed in this section assumes that these vectors are similar, at least in their level of sparsity.

We assume that each vector of β_i, $i = 1, \dots, n$, solves a different problem, and these problems are related only through the common penalty in the joint loss function, which is the sum of the individual losses, see (4).

We want to introduce some notation. We assume that for each $i = 1, \dots, n$, $z_{ij} = (y_{ij}, x_{ij}^\mathsf{T})^\mathsf{T}$, $j = 1, \dots, m$ are i.i.d., sub-Gaussian random variables, drawn from a distribution Q_i. Let $z_i = (y_i, x_i^\mathsf{T})^\mathsf{T}$ be an independent sample from Q_i. For any vector a, let $\tilde{a} = (-1, a^\mathsf{T})^\mathsf{T}$, 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:

$$L(\mathcal{B}, \mathfrak{S}) = \sum_{i=1}^{n} \mathrm{E}_{Q_i}(y_i - x_i^\mathsf{T}\beta_i)^2 = \sum_{i=1}^{n} \tilde{\beta}_i^\mathsf{T} \tilde{\Sigma}_i \tilde{\beta}_i. \tag{5}$$

For p small, the natural approach is empirical risk minimization, that is replacing $\tilde{\Sigma}_i$ in (5) 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

$$\begin{aligned} \left|\tilde{\beta}^{\mathsf{T}}(\tilde{\Sigma}_i - \tilde{S}_i)\tilde{\beta}\right| &\le \|\tilde{\Sigma}_i - \tilde{S}_i\|_\infty \|\tilde{\beta}\|_1^2 \\ \text{and} \qquad & \qquad (6) \\ \delta_m \equiv \|\tilde{\Sigma}_i - \tilde{S}_i\|_\infty &= \mathcal{O}_p(m^{-1/2}\log p), \end{aligned}$$

where $\|A\|_\infty = \max_{i,j}|A_{ij}|$.

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 $\tilde{\beta}_1, \ldots, \tilde{\beta}_n$, and obtain the estimator defined by:

$$\begin{aligned} (\tilde{\hat{\beta}}_i, \ldots, \tilde{\hat{\beta}}_n) &= \underset{\tilde{\beta}_1,\ldots,\tilde{\beta}_n}{\arg\min}\Big\{ m \sum_{i=1}^n \tilde{\beta}_i^{\mathsf{T}} \tilde{S}_i \tilde{\beta}_i + \lambda_m \sum_{i=1}^n \|\tilde{\beta}_i\|_1^2 \Big\} \\ &= \underset{\tilde{\beta}_1,\ldots,\tilde{\beta}_n}{\arg\min} \sum_{i=1}^n \Big\{ \sum_{j=1}^m (y_{ij} - x_{ij}^{\mathsf{T}}\beta_i)^2 + \lambda_m \|\tilde{\beta}_i\|_1^2 \Big\}. \end{aligned} \qquad (7)$$

Note that when $n = 1$, the fact that we penalized by the squared ℓ_1 norm, and not by the ℓ_1 norm itself does not make a difference. To be more exact, if $n = 1$, for any λ_m in (7), there is λ'_m such that the least square with penalty $\lambda'_m\|\tilde{\beta}_1\|_1$ yields the same value as (7).

Also, (7) may seem as, and numerically it certainly is, n separate problems, each involving a specific β_i. The problems are related however, because the penalty function is the same for all. They are tied, therefore, by the chosen value of λ_n, whether this is done *a-priori*, or by solving a single constraint maximization problem, or if λ_m is selected *a-posteriori* by a method like cross validation.

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}, $i = 1, \ldots, n$ be n arbitrary vectors and let $C_n = n^{-1}\sum_{i=1}^n \tilde{\beta}_{i0}^{\mathsf{T}}\tilde{\Sigma}_i\tilde{\beta}_{i0}$. Let $c_n = n^{-1}\sum_{i=1}^n \min_\beta \tilde{\beta}^{\mathsf{T}}\tilde{\Sigma}_i\tilde{\beta}$. Then*

$$\sum_{i=1}^n \tilde{\hat{\beta}}_i^{\mathsf{T}}\tilde{\Sigma}_i\tilde{\hat{\beta}}_i \le \sum_{i=1}^n \tilde{\beta}_{i0}^{\mathsf{T}}\tilde{\Sigma}_i\tilde{\beta}_{i0} + (\frac{\lambda_m}{m} + \delta_m)\sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 - (\frac{\lambda_m}{m} - \delta_m)\sum_{i=1}^n \|\tilde{\hat{\beta}}_i\|_1^2,$$

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

$$\sum_{i=1}^n \|\tilde{\hat{\beta}}_i\|_1^2 = \mathcal{O}_p\big(mn\frac{C_n - c_n}{\lambda_m}\big) + \big(1 + \mathcal{O}(\frac{m^{1/2}}{\lambda_m}\log(np))\big)\sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 \qquad (8)$$

and

$$\sum_{i=1}^n \tilde{\hat{\beta}}_i^\mathsf{T}\tilde{\Sigma}_i\tilde{\hat{\beta}}_i \le \sum_{i=1}^n \tilde{\beta}_{i0}^\mathsf{T}\tilde{\Sigma}_i\tilde{\beta}_{i0} + \big(1 + \mathcal{o}_p(1)\big)\frac{\lambda_m}{m}\sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2.$$

The result is meaningful, although not as strong as may be wished, as long as $C_n - c_n \to 0$, while $n^{-1}\sum_{i=1}^n \|\tilde{\beta}_{i0}\|_1^2 = \mathcal{o}_p(m^{1/2})$. 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 $\tilde{\beta}^\mathsf{T}\tilde{\Sigma}_i\tilde{\beta}$ itself is sparse, then by (8) $\tilde{\hat{\beta}}_1, \ldots, \tilde{\hat{\beta}}_n$ are as sparse as the true minimizers.

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

The estimators $\tilde{\hat{\beta}}_1, \ldots, \tilde{\hat{\beta}}_m$ look as if they are the mode of the a-posteriori distribution of the β_i's when $y_{ij}|\beta_i \sim N(x_{ij}^\mathsf{T}\beta_i, \sigma^2)$, the $\beta_1, \ldots, \beta_n$ are a priori independent, and β_i has a prior density proportional to $\exp(-\lambda_m\|\tilde{\beta}_i\|_1^2/\sigma^2)$. This distribution can be constructed as follows. Suppose $T_i \sim N(0, \lambda_m^{-1}\sigma^2)$. Given T_i, let $u_{i1}, \ldots, u_{ip}$ be distributed uniformly on the simplex $\{u_{i\ell} \ge 0, \sum_{\ell=1}^n u_{i\ell} = |T_i|\}$. Let $s_{i1}, \ldots, s_{ip}$ be i.i.d. Rademacher random variables (taking values ± 1 with probabilities 0.5), independent of $T_i, u_{i1}, \ldots, u_{ip}$. Finally let $\beta_{i\ell} = u_{i\ell}s_{i\ell}$, $\ell = 1, \ldots, p$.

However, this Bayesian point of view is not consistent with the above suggested value of λ_m. 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_m 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 each β_i, although at a slower rate than m. In fact, even if we relax what we mean by "prior", the value of λ_m goes in the 'wrong' direction. As $m \to \infty$, one may wish to use weaker a-priori assumptions, and allow T above to have a-priori second moment going to infinity, not to 0, as entailed by $\lambda_m \to 0$.

The Bayesian inconsistency does not come from the asymptotic setup, and it does not come from considering a more and more complex model. It

was presented as asymptotic in $m \to \infty$, because it is clear that asymptotically we get the wrong results, but the phenomena occurs along the way and not only in the final asymptotic destination. The parameter λ should be much larger than we believe *a priori* that $\|\tilde{\beta}_i\|_1^{-1}$ should be. If λ is chosen such that the prior distribution have the level of sparsity we believe in, than the posteriori distribution would not be sparse at all! To obtain a sparse solution, we should pose a prior which predicts an almost zero vector β. Also, the problem does not follow from increasing the dimension, because the asymptotic is in m and not in p, the latter is very large along the process. We could start the "asymptotic" discussion with m_0 observations per $\beta_i \in \mathbb{R}^{p_0}$, p_0 almost exponential in m_0. Then we could keep p constant, while increasing m. We would get the inconsistency much before m will be $\mathcal{O}(p_0)$. Finally, the Bayesian inconsistency is not because the real dimension, the number of non-zero entries of β_i, is increasing. In fact, the inconsistency appears when this number is kept of the same order, and the prior predicts increasingly sparse vectors (but not fast enough). In short, the problem is that the considered prior distribution cannot compete with the likelihood when the dimension of the observations is large (note, just 'large', not 'asymptotically large').

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—the price for a single non-sparse vector is higher as α increases. Note that given the value of λ_m, 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 $\tilde{\beta}^\mathsf{T}\Sigma_i\tilde{\beta}$. Suppose $\max_i \|\beta_{i0}\|_1 < \infty$. Consider the estimators:*

$$(\tilde{\hat{\beta}}_i, \ldots, \tilde{\hat{\beta}}_n) = \underset{\tilde{\beta}_1,\ldots,\tilde{\beta}_n}{\arg\min}\Big\{ m \sum_{i=1}^n \tilde{\beta}_i^\mathsf{T} \tilde{S}_i \tilde{\beta}_i + \lambda_m \sum_{i=1}^n \|\tilde{\beta}_i\|_1^\alpha \Big\}$$

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

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

and

$$\sum_{i=1}^{n}\tilde{\tilde{\beta}}_i^{\mathsf{T}}\tilde{\Sigma}_i\tilde{\tilde{\beta}}_i \le \sum_{i=1}^{n}\tilde{\beta}_{i0}^{\mathsf{T}}\tilde{\Sigma}_i\tilde{\beta}_{i0} + \mathcal{O}_p(n(m/\lambda_m)^{2/(\alpha-2)}\delta_m^{\alpha/(\alpha-2)}).$$

Note that, if we, the Bayesians, believe that $n^{-1}\sum_{i=1}^{n}\|\tilde{\tilde{\beta}}_i\|_1^2 \xrightarrow{\mathrm{p}} c$, then λ_m should also converge to a constant. Theorem 2.2 implies that the estimator is persistent if $m^2\delta_n^\alpha \to 0$, or $\alpha > 4$. That is, the prior should have a very short tails. In fact, if the prior's tails are short enough, we can accommodate an increasing value of the $\tilde{\tilde{\beta}}_i$'s by taking $\lambda_m \to 0$.

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

We want to conclude on another role of the parameter α. The parameter λ_m controls the average $n^{-1}\sum_{i=1}^{n}\|\tilde{\beta}_i\|_1^\alpha$. When $\alpha = 1$, we are relatively flexible and allow some $\|\tilde{\beta}_i\|_1$ to be very large, as long as other are small. If α is larger, the penalty for $\|\tilde{\beta}_i\|_1$ much larger than the average becomes too large, and the solution tends to be with all $\|\tilde{\beta}_i\|_1$ being of the same order.

2.2 Restricted eigenvalues condition and oracle inequalities

The above discussion was based on the persistent type of argument. The results are relatively weak, but in return the conditions are very weak. For completeness we give much stronger results based on much stronger conditions. We show that the needed coefficient of the penalty function remains the same, and therefore the Bayesian discussion did not depend on the results presented above. Before stating the conditions and the resulted inequalities 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, $X = \mathrm{diag}(X_1, X_2, \dots, X_n)$, and with some abuse of notation, a matrix $\Delta = (\Delta_1,\dots,\Delta_n)$ may be

considered as the vector $(\Delta_1^\mathsf{T}, \ldots, \Delta_n^\mathsf{T})^\mathsf{T}$. Finally, recall the notation $\mathcal{B} = (\beta_1, \ldots, \beta_n)$

The restricted eigenvalue assumption of Bickel et al. [1] (and Lounici et al. [8]) can be generalized to incorporate unequal subsets J_is. In the assumption below, the restriction is given in terms of $\ell_{q,1}$ norm, $q \geqslant 1$.

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

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

We apply it with $q = 1$, and in Lounici et al. [8] 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. [8].

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

Theorem 2.3 *Assume $y_{ij} \sim \mathcal{N}(x_{ij}^\mathsf{T}\beta_i, \sigma^2)$, and let $\hat{\beta}$ be a minimizer of (7), with*

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

where $\alpha \geqslant 1$ and $A > \sqrt{2}$, $B \geqslant \max_i ||\beta_i||_1$ and $\hat{B} \geqslant \max_i ||\hat{\beta}_i||_1$, $\max(B, \hat{B}) > 0$ (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 $\mathcal{M}(\beta_i) \leqslant s$ for all i.

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

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

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

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

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

(c)

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

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

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. [1]). 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 $y_{ij} \sim \mathcal{N}(x_{ij}^\mathsf{T}\beta_i, \sigma^2)$, with $\max_i \|\beta_i\|_1 \leqslant b$ where $b > 0$ can depend on n, m, p. Take some $\eta \in (0,1)$. Let $\hat{\beta}$ be a minimizer of (7), 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 $\mathcal{M}(\beta_i) \leqslant s$ for all i.

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

(a) The prediction error can be bounded by:

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

(b) The estimation absolute error is bounded by:

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

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

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

where $\phi_{\max}$ is the largest eigenvalue of $X^\mathsf{T}X/m$.

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 $||\mathcal{B}||_2^2 \leqslant Cm\delta_n$.

3 Group lasso: Bayesian perspective

Write $\mathcal{B} = (\beta_1, \ldots, \beta_n) = (\mathfrak{b}_1^\mathsf{T}, \ldots, \mathfrak{b}_p^\mathsf{T})^\mathsf{T}$. Th group lasso is defined (see Yuan and Lin [12]) by

$$\hat{\mathcal{B}} = \arg\min \left[\sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - x_{ij}^\mathsf{T} \beta_i)^2 + \lambda \sum_{\ell=1}^{p} \|\mathfrak{b}_\ell\|_2 \right] \tag{9}$$

Note that $(\hat{\beta}_1, \ldots, \hat{\beta}_n)$ 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.

Note that (9) is equivalent to the mode of the a-posteriori distribution when given $\mathcal{B}$, Y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, are all independent, $y_{ij} \mid \mathcal{B} \sim \mathcal{N}(x_{ij}^\mathsf{T} \beta_i, \sigma^2)$, and a-priori, $\mathfrak{b}_1, \ldots, \mathfrak{b}_p$, are i.i.d.,

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

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 its components are uncorrelated and have mean 0. However, they are not independent. Change of variables to a polar system where

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

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

$$f(R_\ell, w_\ell) = C_{n,\lambda} R_\ell^{n-1} e^{-\tilde{\lambda} R_\ell}, \qquad R_\ell > 0, \tag{10}$$

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 $\mathfrak{b}_\ell$, say the first, given the rest has the form

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

which for small $\mathfrak{b}_{\ell 1}/\rho$ looks like the normal density with mean 0 and variance $\rho/\tilde{\lambda}$, while for large $\mathfrak{b}_{\ell 1}/\rho$ 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 Ys are small, this component dominates the log-a-posteriori distribution and hence the maximum will be at 0.

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

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

Hence

$$\hat{\mathfrak{b}}_{\ell i} = \frac{\sum_{j=1}^{m} x_{ij\ell} \tilde{Y}^{\mathcal{B}}_{ij\ell}}{\lambda^*_\ell + \sum_{j=1}^{m} x^2_{ij\ell}}. \tag{11}$$

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

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

We can use (11) to solve for λ^*_ℓ

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

Hence λ^*_ℓ is the solution of

$$\lambda^2 = \sum_{i=1}^{n} \left(\frac{\lambda^*_\ell \sum_{j=1}^{m} x_{ij\ell} \tilde{Y}^{\mathcal{B}}_{ij\ell}}{\lambda^*_\ell + \sum_{j=1}^{m} x^2_{ij\ell}} \right)^2. \tag{12}$$

Note that the RHS is monotone increasing, so (12) has at most a unique solution. It has no solution if at the limit $\lambda_\ell^* \to \infty$, the RHS is still less than λ^2. That is if

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

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

$$\lambda^2 > \sum_{i=1}^{n} \Big(\sum_{j=1}^{m} x_{ij\ell} Y_{ij\ell} \Big)^2, \qquad \ell = 1, \ldots, p$$

Then all the random effect vectors are 0. In the balanced case the RHS is $\mathcal{O}_p(mn \log(p))$. By (10), 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 $\mathfrak{b}_\ell$ has norm which is $o(m^{-1})$. This conclusion is supported by the recommended value of λ given, e.g. in [8].

4 RING lasso: Bayesian perspective

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

$$|||\mathcal{B}|||_1 = \mathrm{trace}\Big\{ \big(\sum_{i=1}^{n} \beta_i \beta_i^\mathsf{T} \big)^{1/2} \Big\},$$

where $\mathcal{B} = (\beta_1, \ldots, \beta_n) = (\mathfrak{b}_1^\mathsf{T}, \ldots, \mathfrak{b}_p^\mathsf{T})^\mathsf{T}$. This is also known as trace norm or Schatten norm with $p = 1$. Note that $|||\mathcal{B}|||_1 = \sum c_i^{1/2}$ where $c_1, \ldots, c_p$ are the eigenvalues of $\mathcal{B}\mathcal{B}^\mathsf{T} = \sum_{i=1}^n \beta_i \beta_i^\mathsf{T}$ (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

$$\hat{\mathcal{B}} = \underset{\mathcal{B} \in \mathbb{R}^{p \times n}}{\arg\min} \{ \sum_{i=1}^{n} (y_{ij} - x_{ij}^\mathsf{T} \beta_i)^2 + \lambda |||\mathcal{B}|||_1. \} \tag{13}$$

We refer to this problem as RING (Rotation INvariant Group) lasso. See [2] for more details.

We consider now the penalty for β_k for a fixed k. Let $A = n^{-1} \sum_{k \neq i} \beta_k \beta_k^\mathsf{T}$, and write the spectral value decomposition $n^{-1} \sum_{k=1}^n \beta_k \beta_k^\mathsf{T} = \sum c_j x_j x_j^\mathsf{T}$

where $\{x_j\}$ is an orthonormal basis of eigenvectors. Using Taylor expansion for not too big β_i, we get

$$\begin{aligned}\operatorname{trace}\big((nA+\beta_i\beta_i^\mathsf{T})^{1/2}\big) &\approx \sqrt{n}\operatorname{trace}(A^{1/2}) + \sum_{j=1}^{p}\frac{x_j^\mathsf{T}\beta_i\beta_i^\mathsf{T}x_j}{2c_j^{1/2}} \\ &= \sqrt{n}\operatorname{trace}(A^{1/2}) + \frac{1}{2}\beta_i^\mathsf{T}\big(\sum c_j^{-1/2}x_jx_j^\mathsf{T}\big)\beta_i \\ &= \sqrt{n}\operatorname{trace}(A^{1/2}) + \frac{1}{2}\beta_i^\mathsf{T}A^{-1/2}\beta_i\end{aligned}$$

Hence the estimator is if 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" looks like normal for the center of the distribution and has 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 $\mathcal{B} = (\beta_1, \ldots, \beta_n)$. 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 $\mathfrak{b}_1, \ldots, \mathfrak{b}_p$ are orthonormal set of eigenvectors of $\mathcal{B}^\mathsf{T}\mathcal{B}$ and $\gamma_{ij} = \mathfrak{b}_j^\mathsf{T}\beta_i$ — the PCA of $\beta_1, \ldots, \beta_n$, then $|||\mathcal{B}|||_1 = \sum_{j=1}^{p}\big(\sum_{i=1}^{n}\gamma_{ij}^2\big)^{1/2}$ — the RING lasso penalty in terms of the principal components. The "prior" is then proportional to $e^{-\tilde{\lambda}\sum_{j=1}^{p}\|\gamma_{\cdot j}\|_2}$ where $\tilde{\lambda} = \lambda/(2\sigma^2)$. Namely, we can obtain a random $\mathcal{B}$ from the prior by the following procedure:

1. Sample $r_1, \ldots, r_p$ independently from $\Gamma(n, \tilde{\lambda})$ distribution.
2. For each $j = 1, \ldots, p$ sample $\gamma_{1j}, \ldots, \gamma_{nj}$ independently and uniformly on the sphere with radius r_j.
3. Sample an orthonormal base $\chi_1, \ldots, \chi_p$ "uniformly".
4. Construct $\beta_i = \sum_{j=1}^{p}\gamma_{ij}\chi_j$.

A Appendix

Proof of Theorem 2.1. Note that by the definition of $\tilde{\hat{\beta}}_i$ and (6).

$$
\begin{aligned}
&mnc_n + \lambda_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad \le m \sum_{i=1}^{n} \tilde{\hat{\beta}}_i^{\mathsf{T}} \tilde{\Sigma}_i \tilde{\hat{\beta}}_i + \lambda_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad \le m \sum_{i=1}^{n} \tilde{\hat{\beta}}_i^{\mathsf{T}} \tilde{S}_i \tilde{\hat{\beta}}_i + (\lambda_m + m\delta_m) \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad \le m \sum_{i=1}^{n} \tilde{\beta}_{i0}^{\mathsf{T}} \tilde{S}_i \tilde{\beta}_{i0} + \lambda_m \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2 + m\delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad \le m \sum_{i=1}^{n} \tilde{\beta}_{i0}^{\mathsf{T}} \tilde{\Sigma}_i \tilde{\beta}_{i0} + (\lambda_m + m\delta_m) \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2 + m\delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad = mnC_n + (\lambda_m + m\delta_m) \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2 + m\delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2.
\end{aligned}
\tag{14}
$$

Comparing the LHS with the RHS of (14), noting that $m\delta_m \ll \lambda_m$:

$$
\sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \le mn \frac{C_n - c_n}{\lambda_m - m\delta_m} + \frac{\lambda_m + m\delta_m}{\lambda_m - m\delta_m} \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2.
$$

By (6) and (7):

$$
\begin{aligned}
\sum_{i=1}^{n} \tilde{\hat{\beta}}_i^{\mathsf{T}} \tilde{\Sigma}_i \tilde{\hat{\beta}}_i &\le \sum_{i=1}^{n} \tilde{\hat{\beta}}_i^{\mathsf{T}} \tilde{S}_i \tilde{\hat{\beta}}_i + \delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\le \sum_{i=1}^{n} \tilde{\beta}_{i0}^{\mathsf{T}} \tilde{S}_i \tilde{\beta}_{i0} + \frac{\lambda_m}{m} \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2 - \frac{\lambda_m}{m} \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 + \delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\le \sum_{i=1}^{n} \tilde{\beta}_{i0}^{\mathsf{T}} \tilde{\Sigma}_i \tilde{\beta}_{i0} + (\frac{\lambda_m}{m} + \delta_m) \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2 - (\frac{\lambda_m}{m} - \delta_m) \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\le \sum_{i=1}^{n} \tilde{\beta}_{i0}^{\mathsf{T}} \tilde{\Sigma}_i \tilde{\beta}_{i0} + (\frac{\lambda_m}{m} + \delta_m) \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2.
\end{aligned}
\tag{15}
$$

The result follows. □

Proof of Theorem 2.2. The proof is similar to the proof of Theorem 2.1. Similar to (14) we obtain:

$$\begin{aligned}
&mnc_n + \lambda_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^\alpha \\
&\quad \le m \sum_{i=1}^{n} \tilde{\hat{\beta}}_i^\mathsf{T} \tilde{\Sigma}_i \tilde{\hat{\beta}}_i + \lambda_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^\alpha \\
&\quad \le m \sum_{i=1}^{n} \tilde{\hat{\beta}}_i^\mathsf{T} \tilde{S}_i \tilde{\hat{\beta}}_i + \lambda_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^\alpha + m\delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad \le m \sum_{i=1}^{n} \tilde{\beta}_{i0}^\mathsf{T} \tilde{S}_i \tilde{\beta}_{i0} + \lambda_m \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^\alpha + m\delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad \le m \sum_{i=1}^{n} \tilde{\beta}_{i0}^\mathsf{T} \tilde{\Sigma}_i \tilde{\beta}_{i0} + \lambda_m \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^\alpha + m\delta_m \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2 + m\delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2 \\
&\quad = mnc_n + \lambda_m \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^\alpha + m\delta_m \sum_{i=1}^{n} \|\tilde{\beta}_{i0}\|_1^2 + m\delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2.
\end{aligned}$$

That is,

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

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

Hence we can conclude from (16)

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

We now proceed similar to (15)

$$\sum_{i=1}^{n} \tilde{\hat{\beta}}_i^\mathsf{T} \tilde{\Sigma}_i \tilde{\hat{\beta}}_i \le \sum_{i=1}^{n} \tilde{\hat{\beta}}_i^\mathsf{T} \tilde{S}_i \tilde{\beta}_i + \delta_m \sum_{i=1}^{n} \|\tilde{\hat{\beta}}_i\|_1^2$$

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

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

□

Proof of Theorem 2.3. The proof follows that of Lemma 3.1 in Lounici et al. [8].

We start with (a) and (b). Since $\hat{\beta}$ minimizes (7), then, $\forall \beta$

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

and hence, for $Y_i = X_i^{\mathsf{T}} \beta_i + \varepsilon_i$,

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

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

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

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

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

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

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

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

due to inequality $|x^\alpha-y^\alpha|\leqslant\alpha|x-y|\max(|x|^{\alpha-1},|y|^{\alpha-1})$ which holds for $\alpha\geqslant 1$ 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 $\nu+2\mu\leqslant\lambda\mathcal{C}$:

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

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

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

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

This implies that

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

as well as that

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

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

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

which implies

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

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

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

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

$$\begin{aligned}||X^{\mathsf{T}}(\hat{\beta} - \beta)||_2 &\leqslant \left[\frac{3\lambda}{2}\mathcal{C} + 2\mu\right] \frac{\sqrt{n\mathcal{M}(\beta)}}{\kappa\sqrt{m}} \\ &= \frac{\sqrt{n\mathcal{M}(\beta)}}{\kappa\sqrt{m}} \left[\frac{3\lambda}{2}\mathcal{C} + 2A\sigma\sqrt{m\log(np)}\right].\end{aligned}$$

Also,

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

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

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

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

By the triangle inequality,

$$\sum_{\ell:\,\hat{\beta}_{i\ell}\neq 0} ||X_{i\cdot\ell}X_i^{\mathsf{T}}(\beta_i - \hat{\beta}_i)||_2^2 \geqslant \sum_{\ell:\,\hat{\beta}_{i\ell}\neq 0} \Big(||X_{i\cdot\ell}(Y_i - X_i^{\mathsf{T}}\hat{\beta}_i)||_2 - ||X_{i\cdot\ell}(Y_i - X_i^{\mathsf{T}}\beta_i)||_2\Big)^2$$

$$\geq \sum_{\ell:\hat{\beta}_{i\ell}\neq 0} \left(\alpha\,\lambda||\hat{\beta}_i||_1^{\alpha-1}/2-\mu\right)^2$$
$$= \mathcal{M}(\hat{\beta}_i)(\alpha\,\lambda||\hat{\beta}_i||_1^{\alpha-1}/2-\mu)^2.$$

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. □

Proof of Theorem 2.4. To satisfy the conditions of Theorem 2.3, we can take $B=b$ and $\lambda=\frac{4A\sigma}{\alpha b^{\alpha-1}}\sqrt{m\log(np)}$. By Lemma A.1 in Bochkina & Ritov ([2]),

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

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

From the proof of Theorem 2.3, it follows that

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

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

Thus,

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

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

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

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

□

References

[1] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.

[2] N. Bochkina and Y. Ritov. Sparse empirical bayes analysis. *http://arxiv.org/abs/0911.5482*, 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. Greenshtein, J. Park, and Y. Ritov. Estimating the mean of high valued observations in high dimensions. *Journal of Statistical Theory and Practice*, 2:407–418, 2008.

[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] Zhang C. H. Compound decision theory and empirical bayes methods. *Annals of Statistics*, 31:379–390, 2003.

[8] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *COLT*, pages 73–82, 2009.

[9] H. Robbins. Asymptotically subminimax solutions of compound decision problems. *Proc. Second Berkeley Symp. Math. Statist. Probab.*, 1:131–148, 1951.

[10] H. Robbins. An empirical bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1:157–163, 1956.

[11] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58:267–288, 1996.

[12] 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.

[13] C. H. Zhang. General empirical bayes wavelet methods and exactly adaptive minimax estimation. *Annals of Statistics*, 33:54–100, 2005.