IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 235–244
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL723

The Lasso with within group structure

Sara van de Geer[1]

Seminar for Statistics, ETH Zurich

Abstract: We study the group Lasso, where the number of groups is very large, and the sizes of the groups is large as well. We assume there is within group structure, in the sense that the ordering of the variables within groups in some loose sense expresses their relevance. We propose a within group weighting of the variables, and show that with this structure, the group Lasso satisfies a sparsity oracle inequality.

1. Introduction

We study a procedure for regression with group structure, in the linear model

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

Here, $\mathbf{Y}$ is an n-vector of observations, and $\mathbf{X}$ a $(n \times M)$-matrix of co-variables. Moreover, ϵ is a vector of noise, which, for simplicity, we assume to be $\mathcal{N}(0, I)$-distributed. We consider the high-dimensional case, where $M \gg n$, and in fact, where there are p groups of co-variables, each of size T (i. e., $M = pT$), where both p and T can be large. We rewrite the model as

$$\mathbf{Y} = \sum_{j=1}^{p} \mathbf{X}_j \beta_j^0 + \epsilon,$$

where $\mathbf{X}_j = \{\mathbf{X}_{j,t}\}_{t=1}^{T}$ is an $(n \times T)$-matrix and $\beta_j = (\beta_{j,1}, \cdots, \beta_{j,T})^T$ is a vector in $\mathbb{R}^T$. To simplify the exposition, we consider the case where $T \leq n$ and where the Gram matrix within groups is normalized, i. e., $\mathbf{X}_j^T \mathbf{X}_j / n = I$ for all j. The number of groups p can be very large.

The group Lasso was introduced by Yuan and Lin [10]. With large T (say $T = n$), a standard group Lasso will generally not have good prediction properties, even when p is small (say $p = 1$). Therefore, one needs to impose a certain structure within groups. Such an approach has been considered by Meier et al. [4], Ravikumar et al. [5], and Koltchinskii and Yuan [3].

In this paper, we use a similar approach as in Meier et al. [4], but now with a very simple description of structure. This will greatly simplify the theory, i. e., we

[1] Seminar for Statistics, ETH Zurich, Rämistrasse 101, 8092 Zürich, e-mail: geer@stat.math.ethz.ch

AMS 2000 subject classifications: Primary 62G08, 60K35; secondary 62J07.
Keywords and phrases: group Lasso, oracle, sparsity.

need no high-level entropy or concentration of measure arguments. Moreover, it will provide more insight into the required “compatibility condition” (see van de Geer [7] and van de Geer and Bühlmann [8]) or “restricted eigenvalue condition” (see Bickel et all. Bickel et al. [1], Koltchinskii [2]). We remark that the papers Ravikumar et al. [5], and Koltchinskii and Yuan [3] use a fundamentally different penalty. The first puts certain coefficients a priori to zero, whereas the second uses a single penalization instead of the double penalization considered here.

We stress that the present paper is of theoretical nature, giving simplifications of the arguments in Meier et al. [4]. For practical applications and motivations, we refer to the above mentioned papers Meier et al. [4], Ravikumar et al. [5], and Koltchinskii and Yuan [3].

We assume that for all j, there is an ordering in the variables of group j: the larger t, the less important variable $\mathbf{X}_{j,t}$ is likely to be. Given positive weights $\{w_t\}_{t=1}^T$ (which for simplicity we assume to be the same for all groups j), satisfying $0 < w_1 \le \cdots \le w_T$, we express the (lack of) structure in group j with the weighted sum

$$\|W\beta_j\|_2^2 := \sum_{t=1}^T w_t^2 \beta_{j,t}^2, \;\; \beta_j \in \mathbb{R}^p .$$

Examples of weights w_t and of the interpretation of $\|W\beta_j\|_2$ are given in Section 2. The structured group Lasso estimator is defined as

$$\hat\beta := \arg_{\beta\in\mathbb{R}^{pT}} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^p \|\beta_j\|_2 + \lambda\mu \sum_{j=1}^p \|W\beta_j\|_2 \right\},$$

where λ and μ are tuning parameters. Note that the penalty involves two terms proportional to ℓ_2-norms. Penalties proportional to *squared* ℓ_2-norms (as in ridge regression) will in the high-dimensional case generally lead to inconsistent estimators. Note also that when $T = 1$, the above estimator reduces to the standard Lasso as considered by e. g. Tibshirani [6].

We show in this paper that $\hat\beta$ satisfies a sparsity oracle inequality (see Theorem 6.1). This essentially means that the prediction error of the estimator is almost as good as in the case where it is known beforehand which groups are relevant.

The paper is organized as follows. Section 2 gives a typical example for the choice of the weights. In Section 3, we describe how we deal with the noise term. Section 4 discusses approximating quadratic forms $\beta^T\hat\Sigma\beta$, where $\hat\Sigma = \mathbf{X}^T\mathbf{X}/n$ is the Gram matrix. The reason for doing so is that we need a certain amount of identifiability of the parameters, expressed in terms of the compatibility condition of Section 5. The compatibility condition is an extension of the restricted eigenvalue condition of Bickel et al. [1] (see also van de Geer and Bühlmann [8] for a comparison of conditions). It holds for non-singular matrices Σ, and the singular matrix $\hat\Sigma$ inherits this if Σ and $\hat\Sigma$ are close enough. One may for example think of Σ as a “population” version of $\hat\Sigma$. Section 5 presents the details for the present context. Our main result, a sparsity oracle inequality, can then be found in Section 6. The result is given in a non-asymptotic form. A brief discussion of its implications for a typical case is given in Section 7, using orders-of-magnitude to clear up the picture. All proofs are deferred to Section 8.

2. The amount of structure

Let

$$R^2(t) := \sum_{s>t} \frac{1}{w_s^2},\ t = 1, \ldots, T,$$

and let $T_0 \in \{1, \ldots, T\}$ be the smallest value such that

$$T_0 \ge R(T_0)\sqrt{n}.$$

Take $T_0 = T$ if such a value does not exist. We call T_0 the *hidden truncation level.* The faster the w_j increase, the smaller T_0 will be, and the more structure we have within groups. The choice of T_0 is in a sense inspired by a bias-variance trade off.

An extreme case. Suppose we know beforehand that all variables $\mathbf{X}_{j,t}$ with $t \ge 2$ are irrelevant. We then take $w_j = \infty$ for all $j \ge 2$, and we get that $R(t) \equiv 0$. In that case, $T_0 = 1$.

A typical case. Suppose that T is large, and that for some $m > \frac{1}{2}$,

$$w_t = t^m.$$

This may for example correspond to having the basis functions of the Sobolev space of m times differentiable functions as variables. Then $\|W\beta_j\|_2$ can be thought of as a Sobolev norm. For t large, $R(t) \asymp t^{-(2m-1)/2}$, and we find $T_0 \asymp n^{\frac{1}{2m+1}}$, and $T_0/n \asymp n^{-\frac{2m}{2m+1}}$.

We will throughout take the tuning parameters such that $\lambda \ge \sqrt{T_0/n}$ and $\lambda\mu \ge T_0/n$.

3. Handling the noise

It turns out that the noisy part of the problem can be handled by appropriately bounding, for all β, the sample correlations $\epsilon^T\mathbf{X}\beta/n$. We note that

$$\epsilon^T\mathbf{X}\beta/n = \epsilon^T \sum_{j=1}^p \mathbf{X}_j\beta_j/n = \frac{1}{\sqrt{n}} \sum_{j=1}^p V_j^T \beta_j,$$

with $V_j^T := \epsilon^T\mathbf{X}_j/\sqrt{n}$, $j = 1, \ldots, p$. Write

$$\chi_j^2 := \sum_{t=1}^{T_0} V_{j,t}^2.$$

Lemma 3.1. *For all β, it holds that*

$$|\epsilon^T\mathbf{X}\beta|/n \le \left(\max_{1\le j\le p} \sqrt{\frac{\chi_j^2}{T_0}}\right) \sqrt{\frac{T_0}{n}} \sum_{j=1}^p \|\beta_j\|_2 + \left(\max_{1\le j\le p} \|V_j\|_\infty\right) \frac{T_0}{n} \sum_{j=1}^p \|W\beta_j\|_2.$$

The idea of penalization is to prevent a complex model from overfitting i. e., to reduce the estimation error. In our setup the estimation error is due to the noise

ϵ, through the term $\epsilon^T \mathbf{X}\beta/n$. The above lemma will be invoked to show that the penalty

$$\lambda \sum_{j=1}^p \|\beta_j\|_2 + \lambda\mu \sum_{j=1}^p \|W\beta_j\|_2$$

will overrule the noise, provided we choose the tuning parameters λ and μ large enough.

We now derive bounds for the χ_j and $\|V_j\|_\infty$. Note that, for each j, the $\{V_{j,t}\}$ are i.i.d. $\mathcal{N}(0,1)$-distributed, and hence that χ_j^2 is chi-square distributed with T_0 degrees of freedom. Our bounds are based on the following expressions (see Lemma 3.2). Let, for $x > 0$,

$$\nu_0^2 := \nu_0^2(x) = (2x + 2\log(pT)),$$

and

$$\xi_0^2 := \xi_0^2(x) = 1 + \sqrt{\frac{4x + 4\log p}{T_0}} + \frac{4x + 4\log p}{T_0}.$$

Define the set

$$\mathcal{T} := \left\{ \max_{1\le j\le p} \chi_j^2/T_0 \le \xi_0^2, \ \max_{1\le j\le p} \|V_j\|_\infty \le \nu_0 \right\}.$$

Lemma 3.2. *It holds that*

$$\mathbb{P}(\mathcal{T}) \ge 1 - 3\exp[-x].$$

By Lemma 3.1, on $\mathcal{T}$,

$$|\epsilon^T \mathbf{X}\beta|/n \le \xi_0 \sqrt{\frac{T_0}{n}} \sum_{j=1}^p \|\beta_j\|_2 + \nu_0 \frac{T_0}{n} \sum_{j=1}^p \|W\beta_j\|_2.$$

With these result in mind, we will choose $\lambda \ge 8\xi_0\sqrt{T_0/n}$ and $\lambda\mu \ge 8\nu_0 T_0/n$ (the constant 8 is chosen for explicitness).

4. Comparing quadratic forms

Recall that the (sample) Gram matrix is

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X}/n.$$

As $M = pT$ is larger than n, it is clear that $\hat{\Sigma}$ is singular. To deal with this, we will approximate $\hat{\Sigma}$ by a matrix Σ, which potentially is non-singular. For example, when the rows of $\mathbf{X}$ are normalized versions of n i.i.d. random vectors, the matrix Σ could be the population variant of $\mathbf{X}^T\mathbf{X}/n$. We let Σ_j be the $(T \times T)$-submatrix of Σ corresponding to the variables in the j^{th} group (as $\hat{\Sigma}_j := \mathbf{X}_j^T \mathbf{X}_j/n = I$, we typically take $\Sigma_j = I$ as well). We write, for general Σ,

$$\|\beta\|_\Sigma^2 := \beta^T \Sigma \beta, \ \|\beta_j\|_{\Sigma_j}^2 := \beta_j^T \Sigma_j \beta_j, \ j = 1, \ldots, p.$$

Define

$$\text{pen}_1(\beta) := \lambda \sum_j \|\beta_j\|_2, \ \text{pen}_2(\beta) := \lambda\mu \sum_j \|W\beta_j\|_2,$$

and

$$\mathrm{pen}(\beta) := \mathrm{pen}_1(\beta) + \mathrm{pen}_2(\beta).$$

Let

$$\|\hat{\Sigma} - \Sigma\|_\infty := \max_{j,k} |\hat{\Sigma}_{j,k} - \Sigma_{j,k}|.$$

Lemma 4.1. *For all* β

$$\big| \|\beta\|_\Sigma^2 - \|\beta\|_{\hat{\Sigma}}^2 \big| \leq n \|\hat{\Sigma} - \Sigma\|_\infty \mathrm{pen}^2(\beta).$$

5. The compatibility condition

For an index set $S \subset \{1, \dots, p\}$, we let

$$\beta_{j,S} = \beta_j \mathrm{l}\{j \in S\}.$$

Define the set of restrictions

$$\mathcal{R}(S) := \left\{ \beta : \ \mathrm{pen}_1(\beta_{S^c}) + \mathrm{pen}_2(\beta) \leq 3\mathrm{pen}_1(\beta_S) \right\}.$$

Definition *The* structured group Lasso compatibility condition *holds for the set* S*, with constant* $\phi(S) > 0$*, if for all* $\beta \in \mathcal{R}(S)$ *it holds that*

$$\left(\sum_{j \in S} \|\beta_j\|_{\Sigma_j} \right)^2 \leq |S| \|\beta\|_\Sigma^2 / \phi^2(S).$$

This condition is a generalization of the compatibility condition of van de Geer [7] to the case $T > 1$, which is in turn a slightly more general condition than the restricted eigenvalue condition of Bickel et al. [1]. A comparison can be found in van de Geer and Bühlmann [8].

Note that the above condition depends on the choice of Σ. Note also that the compatibility holds if the matrix

$$\begin{pmatrix} \Sigma_1^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_p^{-1/2} \end{pmatrix} \Sigma \begin{pmatrix} \Sigma_1^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_p^{-1/2} \end{pmatrix}$$

is non-singular. One can then take $\phi^2(S)$ as the smallest eigenvalue of this matrix.

The next lemma shows that the structured grouped Lasso compatibility condition implies an analogous compatibility condition with Σ replaced by $\hat{\Sigma}$, provided $|S|$ is sufficiently small (depending on $\|\hat{\Sigma} - \Sigma\|_\infty$). This will be used in the sparsity oracle inequality of the next section.

Let

$$\mathcal{S}(\Sigma) := \left\{ S : \ \frac{64 n \lambda^2 \|\hat{\Sigma} - \Sigma\|_\infty |S|}{\phi^2(S)} \leq \frac{1}{2} \right\}.$$

Lemma 5.1. *For all* $S \in \mathcal{S}(\Sigma)$ *and all* $\beta \in \mathcal{R}(S)$

$$\mathrm{pen}_1^2(\beta_S) \leq 4\lambda^2 |S| \|\beta\|_{\hat{\Sigma}}^2 / \phi^2(S).$$

6. A sparsity oracle inequality

Theorem 6.1. *Consider the structured group Lasso*

$$\hat{\beta} := \arg\min_{\beta} \biggl\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \mathrm{pen}(\beta), \biggr\}$$

where

$$\mathrm{pen}(\beta) := \mathrm{pen}_1(\beta) + \mathrm{pen}_2(\beta),$$

and where

$$\mathrm{pen}_1(\beta) := \lambda \sum_{j=1}^p \|\beta_j\|_2, \ \mathrm{pen}_2(\beta) := \lambda\mu \sum_{j=1}^p \|W\beta_j\|_2,$$

with

$$\lambda \ge 8\xi_0\sqrt{T_0/n}, \ \lambda\mu \ge 8\nu_0 T_0/n,$$

with ξ_0 and ν_0 given in Section 3. Let also $\mathcal{T}$ be as in Section 3. Then $\mathbb{P}(\mathcal{T}) \ge 1 - 3\exp[-x]$. On $\mathcal{T}$, we have for all $S \in \mathcal{S}(\Sigma)$ (with $\mathcal{S}(\Sigma)$, as given in Section 5, the small enough index sets), and all β_S,

$$\|\hat{\beta} - \beta^0\|_{\hat{\Sigma}}^2 + \mathrm{pen}(\hat{\beta} - \beta_S) \le 4 \biggl\{ \frac{4\lambda^2|S|}{\phi^2(S)} + \|\beta_S - \beta^0\|_{\hat{\Sigma}}^2 + 2\mathrm{pen}_2(\beta_S) \biggr\}.$$

The above theorem gives a bound for the prediction error

$$\|\hat{\beta} - \beta^0\|_{\hat{\Sigma}}^2 = \|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n.$$

In addition, it bounds the ℓ_1/ℓ_2 estimation error

$$\sum_{j=1}^p \|\hat{\beta}_j - \beta_{S_*}\|_2,$$

where β_{S_*} can be taken as the "oracle" minimizing the right hand side, i. e.,

$$\beta_{S_*} := \arg \min_{\beta_S:\ S \in \mathcal{S}(\Sigma)} \biggl\{ \frac{4\lambda^2|S|}{\phi^2(S)} + \|\beta_S - \beta^0\|_{\hat{\Sigma}}^2 + 2\mathrm{pen}_2(\beta_S) \biggr\}.$$

Thirdly, it bounds the estimated "smoothness"

$$\sum_{j=1}^p \|W\hat{\beta}_j\|_2.$$

7. A typical case

Let $S_0 := \{j : \ \|\beta_j^0\|_2 = 0\}$ be the active set of β^0.

— Suppose that β^0 itself is sparse, in fact that $S_0 \in \mathcal{S}(\Sigma)$.

— Let $T = n$, $w_t = t^m$ $(m > 1/2)$, and $p \ge n$.

— Assume moreover that $\|W\beta_j^0\|_2 \le 1$.

We may choose $\lambda \asymp \sqrt{\log(p)T_0/n}$, and (invoking $\log p/T_0 = O(\log(p))$) $\lambda\mu \asymp \log(p)T_0/n$. Recall moreover that (with this particular choice of weights), $T_0/n \asymp n^{-\frac{2m}{2m+1}}$. Taking $\beta_S = \beta^0$ in Theorem 6.1. now yields

$$\|\hat{\beta} - \beta^0\|_{\hat{\Sigma}}^2 + \text{pen}(\hat{\beta} - \beta^0) = O\biggl(n^{-\frac{2m}{2m+1}} \log(p) \frac{|S_0|}{\phi^2(S_0)} \biggr).$$

In other words, the rate of convergence is roughly the same as in the case where S_0 is known beforehand. The price paid is a logarithmic term and a possibly very small constant $\phi(S_0)$.

Let us now have a closer look at the requirement $S_0 \in \mathcal{S}(\Sigma)$. Recall that the compatibility constant depends on Σ, say $\phi(S) := \phi_\Sigma(S)$. The assumption $S_0 \in \mathcal{S}(\Sigma)$ is a means to get a hold on $\phi_{\hat{\Sigma}}(S)$. A typical case (say the case where the rows of $\mathbf{X}$ are normalized versions of n i.i.d. sub-Gaussian random vectors, and Σ is the population Gram matrix) is

$$\|\hat{\Sigma} - \Sigma\|_\infty \asymp \sqrt{\frac{\log(p)}{n}}.$$

We then require that $|S_0|/\phi_\Sigma^2(S_0)$ is sufficiently small, say

$$\frac{|S_0|}{\phi_\Sigma^2(S_0)} = o\biggl(\frac{n^{\frac{2m-1}{2(2m+1)}}}{\log^{3/2}(p)} \biggr).$$

8. Proofs

Proof of Lemma 3.1. We have

$$\begin{aligned} |\epsilon^T \mathbf{X}\beta|/n &\le \sum_{j=1}^p |V_j^T \beta_j|/\sqrt{n} \\ &\le \sum_{j=1}^p \sqrt{\frac{\chi_j^2}{T_0}} \sqrt{\frac{T_0}{n}} \|\beta_j\|_2 + \sum_{j=1}^p \|V_j\|_\infty \frac{R(T_0)}{\sqrt{n}} \|W\beta_j\|_2 \\ &\le \biggl(\max_{1\le j\le p} \sqrt{\frac{\chi_j^2}{T_0}} \biggr) \sqrt{\frac{T_0}{n}} \sum_{j=1}^p \|\beta_j\|_2 + \biggl(\max_{1\le j\le p} \|V_j\|_\infty \biggr) \frac{R(T_0)}{\sqrt{n}} \sum_{j=1}^p \|W\beta_j\|_2. \end{aligned}$$

The choice of T_0 guarantees that $R(T_0)/\sqrt{n} \le T_0/n$. □

Proof of Lemma 3.2. As $V_{j,t}$ is $\mathcal{N}(0,1)$-distributed, it follows from the union bound that

$$\begin{aligned} \mathbb{P}\biggl(\max_{1\le j\le p} \max_{1\le t\le T} |V_j| > \sqrt{2x + 2\log(pT)} \biggr) &\le 2pT \exp\Bigl[-(x + \log(pT)) \Bigr] \\ &= 2\exp\left[-x\right]. \end{aligned}$$

Furthermore, by the inequality of Wallace [9], for all $a > 0$,

$$\mathbb{P}\Bigl(\chi_j^2 \ge T(1+a) \Bigr) \le \exp\biggl[-\frac{T_0}{2}\Bigl(a - \log(1+a) \Bigr) \biggr].$$

We now use that

$$a - \log(1+a) \geq \frac{a^2}{2(1+a)}.$$

This gives

$$\mathbb{P}(\chi_j^2 \geq T_0(1+a)) \leq \exp\left[-\frac{T_0}{4}\left(\frac{a^2}{1+a} \right) \right].$$

Insert

$$a = \sqrt{\frac{4x}{T_0}} + \frac{4x}{T_0}.$$

Then

$$\frac{a^2}{1+a} \geq \frac{4x}{T_0},$$

so

$$\mathbb{P}\left(\chi_j^2 \geq T_0 \left(1 + \sqrt{\frac{4x}{T_0}} + \frac{4x}{T_0} \right) \right) \leq \exp[-x].$$

Finally, apply the union bound to arrive at

$$\mathbb{P}\left(\max_{1 \leq j \leq p} \chi_j^2 / T_0 \geq \xi_0^2 \right) \leq \exp[-x].$$

□

Proof of Lemma 4.1.

$$|\beta^T \hat{\Sigma} \beta - \beta^T \Sigma \beta| \leq \|\hat{\Sigma} - \Sigma\|_\infty \|\beta\|_1^2,$$

and

$$\|\beta_j\|_1 \leq \sqrt{T_0} \|\beta_j\|_2 + R(T_0)\|W\beta_j\|_2,$$

Hence

$$\|\beta\|_1 = \sum_{j=1}^p \|\beta_j\|_1 \leq \sum_{j=1}^p \left\{ \sqrt{T_0}\|\beta_j\|_2 + T_0/\sqrt{n}\|W\beta_j\|_2 \right\},$$

where we use $R(T_0) \leq T_0/\sqrt{n}$. Finally, invoke $\sqrt{T_0/n} \leq \lambda$ and $T_0/n \leq \lambda\mu$. □

Proof of Lemma 5.1. Let β be some vector in $\mathcal{R}(S)$. Then

$$\text{pen}(\beta_S) = \text{pen}_1(\beta_S) + \text{pen}_2(\beta_S) \leq 4\text{pen}_1(\beta_S),$$

and

$$\text{pen}(\beta) = \text{pen}_1(\beta_S) + \text{pen}_1(\beta_{S^c}) + \text{pen}(\beta) \leq 4\text{pen}_1(\beta_S).$$

Define

$$\eta^2 := n\lambda^2 \|\hat{\Sigma} - \Sigma\|_\infty |S| / \phi^2(S).$$

Then, since $\phi(S) \leq 1$, and $|S| \geq 1$,

$$\lambda^2 \|\beta_j\|_2^2 = \lambda^2 \|\beta_j\|_{\hat{\Sigma}_j}^2 \leq \lambda^2 \|\beta_j\|_{\Sigma_j}^2 + \eta^2(\lambda\|\beta_j\|_2 + \lambda\mu\|W\beta_j\|_2)^2.$$

It follows that

$$\text{pen}_1(\beta_S) = \lambda \sum_{j=1}^p \|\beta_j\|_2 \leq \lambda \sum_{j \in S} \|\beta_j\|_{\Sigma_j} + \eta\text{pen}(\beta_S)$$

$$\leq \sqrt{|S|}\lambda\|\beta\|_\Sigma/\phi(S) + 4\eta\text{pen}_1(\beta_S)$$

$$\leq \sqrt{|S|}\left(\frac{\lambda\|\beta\|_{\hat\Sigma}^2 + \phi(S)\eta\text{pen}(\beta)/\sqrt{|S|}}{\phi(S)}\right) + 4\eta\text{pen}_1(\beta_S)$$

$$\leq \lambda\sqrt{|S|}\|\beta\|_{\hat\Sigma} + 8\eta\text{pen}_1(\beta_S).$$

The assumption

$$8\eta \leq \frac{1}{2}$$

gives

$$\text{pen}_1(\beta_S) \leq 2\lambda\sqrt{|S|}\|\beta\|_{\hat\Sigma}/\phi(S).$$

□

Proof of Theorem 6.1. Throughout, we assume we are on $\mathcal{T}$.

We have for all β,

$$\|\hat\beta - \beta^0\|_{\hat\Sigma}^2 + \text{pen}(\hat\beta) \leq 2\epsilon^T\mathbf{X}(\hat\beta - \beta)/n + \text{pen}(\beta) + \|\beta - \beta^0\|_{\hat\Sigma}^2$$

$$\leq \frac{1}{4}\text{pen}(\hat\beta - \beta) + \text{pen}(\beta) + \|\beta - \beta^0\|_{\hat\Sigma}^2.$$

It follows that for all S and for $\beta = \beta_S$,

$$\|\hat\beta - \beta^0\|_{\hat\Sigma}^2 + \frac{3}{4}\text{pen}_1(\hat\beta_{S^c}) + \frac{3}{4}\text{pen}_2(\hat\beta - \beta_S)$$

$$\leq \frac{5}{4}\text{pen}_1(\hat\beta_S - \beta_S) + 2\text{pen}_2(\beta_S) + \|\beta_S - \beta^0\|_{\hat\Sigma}^2.$$

Case i)

If

$$\text{pen}_1(\hat\beta_S - \beta_S) \geq \|\hat\beta - \beta^0\|_{\hat\Sigma}^2 + 2\text{pen}_2(\beta_S),$$

we get

$$4\|\hat\beta - \beta^0\|_{\hat\Sigma}^2 + 3\text{pen}_1(\hat\beta_{S^c}) + 3\text{pen}_2(\hat\beta - \beta_S) \leq 9\text{pen}_1(\hat\beta_S - \beta_S). \tag{8.1}$$

So we then have $\hat\beta - \beta_S \in \mathcal{R}(S)$. We therefore can apply Lemma 5.1, to find that when $S \in \mathcal{S}(\Sigma)$, from (8.1),

$$4\|\hat\beta - \beta^0\|_{\hat\Sigma}^2 + 3\text{pen}(\hat\beta - \beta_S) \leq 12\text{pen}_1(\hat\beta_S - \beta_S)$$

$$\leq 24\lambda\sqrt{|S|}\|\hat\beta - \beta_S\|_{\hat\Sigma}/\phi(S) \leq 3\|\hat\beta - \beta_S\|_{\hat\Sigma}^2 + \frac{16\lambda^2|S|}{\phi^2(S)}.$$

Hence

$$\|\hat\beta - \beta^0\|_{\hat\Sigma}^2 + 3\text{pen}(\hat\beta - \beta_S) \leq \frac{16\lambda^2|S|}{\phi^2(S)},$$

so also

$$\|\hat\beta - \beta^0\|_{\hat\Sigma}^2 + \text{pen}(\hat\beta - \beta_S) \leq \frac{16\lambda^2|S|}{\phi^2(S)}.$$

Case ii)

If

$$\text{pen}_1(\hat{\beta}_S - \beta_S) < \|\beta - \beta^0\|_{\hat{\Sigma}}^2 + 2\text{pen}_2(\beta_S),$$

we obtain

$$4\|\hat{\beta} - \beta^0\|_{\hat{\Sigma}}^2 + 3\text{pen}_1(\hat{\beta}_{S^c}) + 3\text{pen}_2(\hat{\beta} - \beta_S) \leq 9\|\beta_S - \beta^0\|_{\hat{\Sigma}}^2 + 18\text{pen}_2(\beta_S),$$

and hence

$$4\|\hat{\beta} - \beta^0\|_{\hat{\Sigma}}^2 + 3\text{pen}(\hat{\beta} - \beta_S) \leq 12\|\beta - \beta^0\|_{\hat{\Sigma}}^2 + 24\text{pen}_2(\beta_S).$$

This gives

$$\|\hat{\beta} - \beta^0\|_{\hat{\Sigma}}^2 + \text{pen}(\hat{\beta} - \beta_S) \leq 4\|\beta - \beta^0\|_{\hat{\Sigma}}^2 + 8\text{pen}_2(\beta_S).$$

□

References

[1] Bickel, J., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

[2] Koltchinskii, V. (2009). Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **45** 7–57.

[3] Koltchinskii, V. and Yuan, M. (2008). Sparse recovery in large ensembles of kernel machines. In *Conference on Learning Theory, COLT* 29–238.

[4] Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics* **37** 3779–3821.

[5] Ravikumar, P., Liu, H., Lafferty, J., and Wasserman, L. (2008). SpAM: sparse additive models. *Advances in neural information processing systems* **20** 1201–1208.

[6] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 1 267–288.

[7] van de Geer, S. (2007). The deterministic Lasso. In *JSM proceedings,* (see also http://stat.ethz.ch/research/research_reports)*/2007/140.* Amer. Statist. Assoc.

[8] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* 1360–1392.

[9] Wallace, D. L. (1959). Bounds for normal approximations of student's t and the chi-square distributions. *Ann. Math. Statist.* **30** 1121–1130.

[10] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal Royal Statistical Society Series B* **68** 1 49.