Electronic Journal of Statistics
Vol. 0 (0000)
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

On the Asymptotic Properties of The Group Lasso Estimator for Linear Models

Yuval Nardi[*] **and Alessandro Rinaldo**[†]

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA
e-mail: yuval@stat.cmu.edu; arinaldo@stat.cmu.edu

Abstract: We establish estimation and model selection consistency, prediction and estimation bounds and persistence for the group-lasso estimator and model selector proposed by Yuan and Lin (2006) for least squares problems when the covariates have a natural grouping structure. We consider the case of a fixed-dimensional parameter space with increasing sample size and the double asymptotic scenario where the model complexity changes with the sample size.

AMS 2000 subject classifications: Primary 62J05; secondary 62F12.
Keywords and phrases: Least Squares, Sparsity, Group-Lasso, Model Selection, Oracle Inequalities, Persistence.

1. Introduction

In recent years there has been a rapidly growing interest in penalized least squares problems via ℓ_1 regularization, especially in high dimensional settings where the model complexity is comparable or even larger than the sample size. The lasso, originally put forward by Tibshirani (1996) for linear regression models, is a regularization procedure in which the penalty for model complexity is the ℓ_1 norm of the estimated coefficients. It has the crucial advantages of being a convex problem, thus computationally feasible even when the number of predictor is larger than the sample size, and of producing solutions that are sparse, i.e. containing zero components. These two key properties make the lasso simultaneously a shrinkage estimation and a model selection procedure that is viable in high-dimensional problems where traditional model selection criteria are not feasible. Furthermore, the lasso has been shown to have optimal theoretical properties: model selection, or sign consistency, or sparsistency (see, e.g., Meinshausen and Bühlmann, 2006; Wainwright, 2006, 2007; Zhao and Yu, 2006), consistency and oracle properties (see, e.g. Meinshausen and Yu, 2006; Bickel et al., 2007; Bunea et al., 2007a,b; Zhang, 2007; Koltchinskii, 2005; Zhang

[*] supported in part by NSF grant EIA0131884 and Army contract DAAD19-02-1-3-0389.
[†] supported in part by NSF grant DMS-0631589 and a grant from the Pennsylvania Department of Health through the Commonwealth Universal Research Enhancement Program.

and Huang, 2007), and persistence (Greenshtein and Ritov, 2006; Greenshtein, 2006).

Researches have also devised few extensions of the lasso that are suited to deal with regression problems in which the explanatory variables are grouped or are organized in a hierarchal manner and, at the same time, exhibit similar computational ease and the shrinkage properties of the lasso. We mention, in particular, the group-lasso procedure by Yuan and Lin (2006) and its extension by Kim et al. (2006), the elastic net regularization by Zou and Hastie (2005), the hierarchical lasso by Zhou and Zhu (2007), regularization methods based on ℓ_∞ penalty by Gilbert et al. (2005) and the very general CAP penalties by Zhao et al. (2007). Most of these procedures essentially comprise a penalty for the model complexity that results from a composition of the ℓ_1 norm with some other norm computed over each group of parameters, thus exhibiting a behavior that, at the group level, resembles that of the lasso solution. Besides ANOVA models, the group-lasso penalty has been applied to generalized linear models in Dahinden et al. (2006), Meier et al. (2006) and Nardi and Rinaldo (2007) and to non-parametric problems in Bach (2007) and Ravikumar et al. (2007).

The general purpose of this paper is to prove for the group-lasso estimator described in Yuan and Lin (2006) the same type of optimality properties that have been established for the lasso estimator. In particular, we will derive conditions ensuring estimation and model selection consistency, prediction and estimation consistency, oracle properties and persistence. For the case of a fixed-dimensional parameter space, Bach (2007) derives some conditions for estimation and model selection consistency. For the double-asymptotic scenario in which the dimension of the parameter space grows with the sample size, a rigorous study of the performance of the group-lasso seems to be missing in the statistical literature. Our contributions include novel consistency and asymptotic normality results for the fixed-dimensional parameter space, model selection consistency when the number of predictors is larger than the sample size, oracle inequalities and persistence properties. Our methods of proofs are based on non-trivial extensions and generalizations of condition and results for the lasso procedure already in existence in the literature.

The paper is organized as follows. Section 2 introduces the group-lasso settings for least square problems. In section 3 we establish estimation and model selection consistency and asymptotic normality under the traditional scenario of increasing sample size and fixed parameter space. The conditions we impose are of different nature than the ones introduced in Bach (2007) and the results we obtain complement that analysis. In section 4 we investigate the properties of the group-lasso solution under the more complex, double-asymptotic scenario in which both the sample size and the model complexity grow simultaneously. In section 4.1 we provide a sufficient condition guaranteeing uniqueness of the group-lasso solution when the number of covariates is larger than the sample size. In section 4.2 we provide conditions for model selection consistency that holds even when the number of covariates grow at a larger rate than the sample size and in section 4.3 we derive finite sample bounds that can be used to establish consistency for estimation and prediction. Finally, in section 4.4 we derive two persistence properties. All the proofs are gathered in section 5 and in the Appendix.

2. The Group-Lasso Settings

Let $\mathcal{H}$ be an index set representing a class of linear subspaces of $\mathbb{R}^n$, each subspace being spanned by the columns of a $n \times d_h$ matrix X_h, where h ranges over $\mathcal{H}$. We will be assuming henceforth that the set $\mathcal{H}$ is known and has been assigned a total ordering, and we will always be using such an ordering. Let X be a $n \times d$ design matrix formed by concatenating the design matrices X_h, $h \in \mathcal{H}$, with $d = \sum_h d_h$. While we allow for non-zero correlations among groups, namely $\mathrm{X}_h^\top \mathrm{X}_{h'} \neq 0$ for $h \neq h'$, we will be making the simplifying assumption

$$\frac{1}{n}\mathrm{X}_h^\top \mathrm{X}_h = \mathrm{I}_{d_h}, \quad \forall h \in \mathcal{H},$$

where I_{d_h} denote the d_h-dimensional identity matrix, a condition that can be enforced via the Gram-Schmidt orthogonalization procedure.

For a subset $\mathcal{H}' \subset \mathcal{H}$, we will write $(\mathcal{H}')^c = \mathcal{H} \setminus \mathcal{H}'$ and, if $x \in \mathbb{R}^d$, we will use the notation $x_{\mathcal{H}'} = \mathrm{vec}\{x_h, h \in \mathcal{H}'\}$ for the d'-dimensional subvector comprised by the blocks of x indexed by $\mathcal{H}'$, where $d' = \sum_{h \in \mathcal{H}'} d_h$. Similarly, if $\mathcal{H}_1$ and $\mathcal{H}_2$ are two subsets of $\mathcal{H}$, and M a $d \times d$ matrix, we will write

$$\mathrm{M}_{\mathcal{H}_1, \mathcal{H}_2}$$

for the $(\sum_{h \in \mathcal{H}_1} d_h) \times (\sum_{h \in \mathcal{H}_2} d_h)$ block matrix, with blocks indexed by the subsets $\mathcal{H}_1$ and $\mathcal{H}_2$. In particular, if $\mathcal{H}_1 = \mathcal{H}_2$, we will simply write $\mathrm{M}_{\mathcal{H}_1}$.

We assume that the n-dimensional observed vector Y satisfies the linear model

$$Y = \mathrm{X}\beta^0 + \epsilon, \tag{1}$$

where ϵ is a n-dimensional vector of *iid* errors, with distributional properties to be specified below, and β^0 is the *unknown* d-dimensional vector of true coefficients. Then, the vector β^0 can be represented as $\mathrm{vec}\{\beta_h^0, h \in \mathcal{H}\}$, the concatenation of $|\mathcal{H}|$ vectors, where $\beta_h^0 \in \mathbb{R}^{d_h}$, for each $h \in \mathcal{H}$. Our crucial modeling assumption is that some of the subvectors of β^0 are zero and we will denote by $\mathcal{H}_0 = \{h \colon \beta_h^0 \neq 0\}$ the *unknown* index set of non-zero subvectors of β^0. Then, the true model complexity is given by $d_0 = \sum_{h \in \mathcal{H}_0} d_h < d$.

We consider the problem of estimating both β^0 and $\mathcal{H}_0$ in the non-trivial situation in which the cardinality $|\mathcal{H}_0|$ of the number or subspaces spanning the true mean vector of the response variable Y is smaller than the total number $|\mathcal{H}|$ of candidate subspaces. In essence, the estimation of the true underlying model $\mathcal{H}_0$ requires identifying, based on Y, the zero subvectors of β^0 and removing the blocks indexed by $\mathcal{H}_0^c$. This may be naturally formulated as a penalized least square problem with a ℓ_0 penalty on the cardinality of the subspaces included. Effectively, this entails considering all possible subsets of $\mathcal{H}$, an NP-hard task that is computationally infeasible, when $|\mathcal{H}|$ (and therefore d) is large. Instead, Yuan and Lin (2006) propose to use the group-lasso penalty, which is a convex relaxation to the ℓ_0 penalty based on the combination of the ℓ_1 penalty over the number of subspaces with the ℓ_2 penalty on the estimated coefficients of each subspace. The resulting group-lasso estimator is obtained as the solution to the convex problem

$$\inf_{\beta \in \mathbb{R}^d} \frac{1}{2n} \|Y - \mathrm{X}\beta\|_2^2 + \lambda \sum_h \lambda_h \|\beta_h\|_2, \tag{2}$$

where λ and $\{\lambda_h,\ h \in \mathcal{H}\}$, are tuning parameters that depend on the sample size n. A reasonable choice for λ_h is $\sqrt{d_h}$, so that larger subspaces are penalized more heavily. The group-lasso regularization is an extension of the lasso, or ℓ_1 penalty function, and consists of applying first the ℓ_2 penalty to individual blocks, to promote non-sparsity, and then the ℓ_1 norm to the resulting block norms, to promote block sparsity. Notice also that the group-lasso problem (2) includes as a special case the lasso and adaptive lasso (see Zou, 2006) problem in which $|\mathcal{H}| = d$ and each h correspond to the 1-dimensional subspace of $\mathbb{R}^n$ spanned by the corresponding column of the design matrix X.

Equation (2) is the Lagrangian function (with Lagrangian multipliers $\{\lambda\lambda_h, h \in \mathcal{H}\}$) of the equivalent convex problem

$$\begin{array}{c} \inf_{\beta \in \mathbb{R}^d} \frac{1}{2n} \|Y - \mathrm{X}\beta\|_2^2 \\ \text{s.t.} \quad \|\beta_h\|_2 \leq t_h, \end{array} \tag{3}$$

where $\{t_h, h \in \mathcal{H}\}$ are non-negative constants. In fact, there exists a correspondence between the coefficients $\{t_h, h \in \mathcal{H}\}$ and $\{\lambda\lambda_h, h \in \mathcal{H}\}$ of (3) and (2), respectively. In this article, we will mostly focus on the more popular, unconstrained formulation (2), which has, in particular, the advantage of letting one choose in a more direct way the regularization parameters. The constrained settings (3) will be used in Section 4.4 to establish persistence properties of the group-lasso.

In our analysis, we will study the asymptotic properties of the group-lasso estimator $\widehat{\beta}$, defined as a minimizer of (2), and of the associated group-lasso model selector

$$\widehat{\mathcal{H}} = \{h \colon \widehat{\beta}_h \neq 0\}. \tag{4}$$

We will consider two asymptotic regimes. In the simpler, traditional scenario, we assume that the model complexity is fixed and that only the sample size n increases. In the second, more modern, scenario, the model complexity increases with the sample size and we will then study the group-lasso solutions to a sequence of linear models in which $|\mathcal{H}|$ and $\{d_h, h \in \mathcal{H}\}$ grow with n. In fact, we allow for d to grow at a faster rate than n. For ease of readability, we will not make the dependence of $\mathcal{H}$, $\{d_h, h \in \mathcal{H}\}$, λ, X, ϵ and $\{\lambda_h, h \in \mathcal{H}\}$ on n explicit, although it will be apparent that all those quantities may change with n.

We conclude this section with some computational remarks. The subgradient conditions for the problem (2) are

$$\begin{array}{rl} -\frac{1}{n}\mathrm{X}_h^\top \left(Y - \mathrm{X}\widehat{\beta}\right) + \lambda\lambda_h \frac{\widehat{\beta}_h}{\|\widehat{\beta}_h\|_2} = 0 & \text{if} \quad \widehat{\beta}_h \neq 0 \\ -\frac{1}{n}\mathrm{X}_h^\top \left(Y - \mathrm{X}\widehat{\beta}\right) + \lambda\lambda_h z_h = 0 & \text{if} \quad \widehat{\beta}_h = 0, \end{array} \tag{5}$$

where z_h are generic vectors such that $\|z_h\|_2 \leq 1$ for all h. Because the objective function in (2) is convex on $\mathbb{R}^d$, the first-order conditions obtained by solving the sub-gradient equations produce the solutions to the group-lasso problem. By inspecting the sub-gradient conditions (5), Yuan and Lin (2006) devise a modification of the LARS algorithm by Efron et al. (2004) to account for the block structure of the penalty function that can be used to solve (2) numerically. Dahinden et al. (2006) improve on this method and develop a different computational strategy based on a block-coordinate gradient descent method in the context of logistic regression, which can be adapted to the present settings. See Dahinden et al. (2006) and, in particular, Zhao et al. (2007) for further details and some discussion on the computational aspects of the group-lasso estimator and on the choice of the regularization parameters.

2.1. *Example: ANOVA Models*

Consider an ANOVA design, arising from the cross-classification of K categorical variables, each taking value on a finite set $\mathcal{I}_k = \{1, \ldots, I_k\}$, for $k = 1, \ldots, K$. Let $\mathcal{I} = \bigotimes_k \mathcal{I}_k$ be the set of cells and $I = \prod_k I_k$ the total number of cells. Also, for each $i \in \mathcal{I}$, let n_i be the total number of observations in cell i. Then $\mathcal{H} = 2^{\mathcal{K}}$, the power set of $\mathcal{K} = \{1, \ldots, K\}$. Each $h \in \mathcal{H}_0$ represents an effect. For example, $h = \emptyset$ corresponds to the grand mean, a subset h with $|h| = 1$ to a main effect and, more generally, a subset h to an interaction effect among the variables indexed by h. The true model can be represented as $\mathcal{H}_0 \subseteq 2^{\mathcal{K}}$.

As h ranges over $\mathcal{H}$, $\mathbb{R}^I$ can be decomposed into the direct sum of orthogonal subspaces indexed by h, each with dimension $\prod_{j \in h}(I_j - 1)$ (see, e.g. Rinaldo, 2006). Let U_h be a matrix of full column rank spanning the subspace index by $h \subset \mathcal{K}$. Then, the columns of $\bigoplus_{h \in 2^{\mathcal{K}}} \mathrm{U}_h$ form a basis for $\mathbb{R}^{\mathcal{I}}$. Next, let T be a $n \times I$ matrix of the form

$$\begin{pmatrix} 1_1 & 0 & \ldots & 0 \\ 0 & 1_2 & \ldots & 0 \\ 0 & 0 & \ldots & 1_I \end{pmatrix},$$

where each 1_i is a n_i-dimensional vector of ones, and $\mathrm{X}_h = \mathrm{T}\mathrm{U}_h$. Notice that $\dim(\mathrm{range}(\mathrm{U}_h)) = \dim(\mathrm{range}(\mathrm{X}_h))$, provided each cell is positive. Then, the $n \times I$ matrix

$$\mathrm{X} = \bigoplus_{h \in 2^{\mathcal{K}}} \mathrm{X}_h$$

has full column rank and its columns span a I-dimensional subspace in $\mathbb{R}^n$. Notice that, for any $h \neq h' \in 2^{\mathcal{K}}$, while $\mathrm{U}_h^\top \mathrm{U}_{h'} = 0$, it is no longer the case that $\mathrm{X}_h^\top \mathrm{X}_{h'} = 0$ when the cells n_i differ, i.e. when the model is unbalanced or when some cells are empty. It is clear that the group-lasso settings described above include as special cases unbalanced and empty-cells ANOVA models, for which the usual decomposition of sums of square does not hold.

3. Fixed-d Asymptotics

In this section we derive conditions for model selection and estimation consistency when the sample size n increases, while the parameter space remains fixed. Our consistency results rely on different assumptions and slightly different settings than the analogous results in Bach (2007), and our analysis provides novel insights into this problem. Furthermore, we obtain rates of consistency and asymptotic normality.

We will use the classical assumptions for consistency of the ordinary least squares solutions:

(F1) $\lim_{n\to\infty} \frac{1}{n}\mathrm{X}^\top \mathrm{X} \to \mathrm{M}$, with M positive definite;
(F2) the errors ϵ_i are *iid* with mean zero and finite second moment σ^2.

To motivate our analysis, we first consider the necessary conditions for the group-lasso procedure to be model selection consistent, namely

$$\mathbb{P}\left\{\widehat{\mathcal{H}} = \mathcal{H}_0\right\} \to 1, \qquad \text{as } n \to \infty, \tag{6}$$

where $\widehat{\mathcal{H}}$ is defined in (4). An adaptation of Theorem 3 in Bach (2007) yields that, under assumption (F1) and (F2), (6) holds only if the weakly irreducibility condition

$$\max_{h' \in \mathcal{H}_0^c} \frac{1}{\lambda_{h'}} \left\| \mathrm{M}_{h',\mathcal{H}_0} \mathrm{M}_{\mathcal{H}_0}^{-1} \mathrm{B}_{\mathcal{H}_0} \beta_{\mathcal{H}_0} \right\|_2 \le 1, \tag{7}$$

is verified, where $\mathrm{B}_{\mathcal{H}_0}$ denotes the d_0-dimensional block-diagonal matrix with blocks

$$\left\{ \mathrm{I}_{d_h} \frac{\lambda_h}{\|\beta_h\|_2}, h \in \mathcal{H}_0 \right\}.$$

We remark that (7) generalizes an analogous necessary condition for model selection consistency of the lasso (see Zou, 2006; Zhao and Yu, 2006; Yuan and Lin, 2006). Below, we derive a different necessary condition for model selection consistency, which provides a rationale for the results we derive in the remainder of this section.

Proposition 3.1. *Under assumption (F1) and (F2), the model selection consistency property (6) holds only if*

$$\sqrt{n}\lambda\lambda_h \to \infty, \quad \forall h \notin \mathcal{H}_0.$$

Using the previous condition, it seems natural to consider sequences of penalty parameters such that $\lambda\lambda_h = O\left(\frac{1}{\sqrt{n}}\right)$ if $h \in \mathcal{H}_0$, and $\sqrt{n}\lambda\lambda_h \to \infty$ otherwise, which will also satisfy the weakly irreducibility condition (7). Implicitly, this idea is behind both Theorem 3.2 and 3.3.

The weak irreducibility condition and the other necessary condition of Proposition 3.1 both have the undesirable feature of depending on the unknown index set $\mathcal{H}_0$ of non-zero blocks. To remedy this problem, in the following result we describe an oracle procedure which automatically yields model selection consistency without knowing $\mathcal{H}_0$. This estimator is obtained as a direct generalization to the group-lasso framework of the adaptive lasso penalty put forward by Zou (2006). We let $\widehat{\beta}^{\mathrm{OLS}} = \left(\mathrm{X}^\top \mathrm{X}\right)^{-1} \mathrm{X}^\top Y$ for the least squares estimate of β. In the proof we essentially follow Zou (2006) and Knight and Fu (2000) and generalize their results to our settings.

Theorem 3.2. *Assume (F1) and (F2) and let* $\lambda_h = \frac{1}{\|\widehat{\beta}_h^{\mathrm{OLS}}\|_2^\gamma}$, *for some* $\gamma > 0$ *such that* $n^{(\gamma+1)/2}\lambda \to \infty$. *If* $\sqrt{n}\lambda \to 0$, *the model selection consistency property (6) is satisfied and, furthermore,*

$$\sqrt{n}(\widehat{\beta} - \beta^0) \xrightarrow{d} Z, \tag{8}$$

where $Z_{\mathcal{H}_0} \sim N_{d_0}(0, \sigma^2 \mathrm{M}_{\mathcal{H}_0}^{-1})$ *and* $Z_{\mathcal{H}_0^c} = 0$.

Remark. The only property of the ordinary least squares estimate $\widehat{\beta}^{\mathrm{OLS}}$ that was used in the proof is its $\sqrt{n}$-consistency. This is enough to guarantee that the penalty parameters $\{\lambda_h, h \in \mathcal{H}_0^c\}$ corresponding to the index set of the zero subvectors of β^0 are very large, with high probability for all n big enough. More generally, the Theorem remains true also when $\widehat{\beta}^{\mathrm{OLS}}$ is replaced by any a_n-consistent estimator, where $a_n \to \infty$, provided $a_n^\gamma \sqrt{n}\lambda \to \infty$.

We conclude this section with one final consistency result for the group-lasso estimator, which demands the knowledge of $\mathcal{H}_0$. Unlike the consistency results derived in (Bach, 2007, Section 2), the weakly irreducibility condition (7) is replaced by conditions on the asymptotic behavior of $\sqrt{n}\lambda\lambda_h$, $h \in \mathcal{H}$. Despite its reduced practical value, this result has the merit of showing explicitly that the penalty terms for the zero and non-zero blocks need to have a different asymptotic behavior.

Theorem 3.3. *Assume (F1), (F2) and further assume that the (possibly random) sequence* $\{a_n\}$, *with* $a_n = |\mathcal{H}_0| \max_{h \in \mathcal{H}_0} \lambda_h$, *satisfy* $\lambda a_n \xrightarrow{p} 0$ *as* $n \to \infty$. *Then* $\|\widehat{\beta} - \beta^0\|_2 = O_P(\frac{1}{\sqrt{n}} + \lambda a_n)$. *Furthermore, if* $\lambda a_n = O_P\left(\frac{1}{\sqrt{n}}\right)$ *and* $\sqrt{n}\lambda\lambda_h \xrightarrow{p} \infty$ *for each* $h \notin \mathcal{H}_0$, *then the conclusions of Theorem 3.2 still hold.*

Remark. The previous Theorem covers cases in which estimation consistency may hold (at a suboptimal rate) but not model selection consistency.

Theorem 3.2 and Theorem 3.3 both establish that the group-lasso estimator is asymptotically optimal, namely unbiased and efficient, and, therefore, offers the same asymptotic guarantees as the ordinary, unpenalized, least squares estimator. However, unlike the ordinary least squares, the group-lasso solutions comes equipped with a built-in penalty for sparsity, so that some of its blocks components will be zero. In fact, and this is key, as n increases, these zero components will be the same zero components of the true vector of coefficients β^0, with probability tending to 1. In contrast, the solutions to the ordinary least squares are all non-zero, thus making it much less effective at recovering $\mathcal{H}_0$.

4. Double Asymptotics

We turn now to the study of double-asymptotic scenario in which $|\mathcal{H}| \to \infty$ and the block-dimensions $\{d_h, h \in \mathcal{H}\}$ are allowed to change with n. In particular, this includes situations in which $d >> n$, i.e. d grows faster than n.

To simplify our derivations, we will enforce a normality assumption on the vector ϵ of errors:

(N) $\epsilon_i \sim N(0, \sigma^2)$, $1 \leq i \leq n$.

Specific cases in which this assumption can be relaxed are discussed as we proceed with our analysis.

4.1. Uniqueness of the Group-Lasso Solution

When $d > n$, there is a $(d-n)$ dimensional affine space of vectors satisfying the model equation (1). As a result, the solution to (2) needs not be unique and, therefore, it may no longer make sense to refer to "the" group-lasso estimator or model selector. To overcome this problem, we may want to impose the following condition, which is enough to guarantee uniqueness of the model representation (1) and, therefore, of the group-lasso solution:

(U(c)) $\max_{h' \neq h} \|\mathrm{X}_{h'}^\top \mathrm{X}_h\|_2 \leq \frac{\lambda_{h'}\lambda_h}{(1+2c)\delta|\mathcal{H}_0|}$, for some constants $c > 0$ and $\delta > \lambda^2_{\max}$,

where, for a $m \times p$ matrix A, $\|\mathrm{A}\|_2$ denotes the operator norm with respect to the Euclidian metric. In stating the assumption, we make explicit only the dependence on the more relevant constant c.

Proposition 4.1. *Under assumption* $(U(c))$*, if* β^1 *and* β^2 *satisfy (1) with* $|\{h \colon \beta^i_h \neq 0\}| \leq |\mathcal{H}_0|$*, for* $i = 1, 2$*, then* $\beta^1 = \beta^2$*.*

Remarks.

1. Assumption (U(c)) is the group-lasso equivalent of Assumption 2 in Lounici (2008) on the maximal mutual coherence between different columns of the design matrix X, which is
$$\max_{i \neq j} \frac{1}{n}|\mathrm{M}_{i,j}| \leq \frac{1}{\rho(1+2c)d_0},$$
where $\rho > 1$ and $c > 0$. We point out that uniqueness of the representation (1) follows also from this mutual coherence condition. However, assumption (U(c)) is more naturally tailored to the problem at hand and, furthermore, implies the important (RE($|\mathcal{H}_0|, c$)) condition (see Proposition 4.4 below), which is essential to establish the bounds derived in Section 4.3.
2. Alternatively, one may consider investigating conditions guaranteeing uniqueness of the group-lasso solution (2) directly, rather of the model representation, following the arguments used in Osborne et al. (2000) for the lasso

problem. Although it is apparent from their analysis that $|\widehat{\mathcal{H}}| \leq n$, i.e. the number of non-zero blocks is no larger than the sample size, extending the polyhedral arguments of Osborne et al. (2000, Section 3.1) to the group-lasso penalty appears problematic.

4.2. Sparsistency

In this section, we provide conditions for the model selection consistency (6) , or sparsistency, of the group-lasso model selector under the double asymptotic settings.

To this end, let $\mathcal{O}$ be the event that there exists a solution $\widehat{\beta}$ to (2) such that $\|\widehat{\beta}_h\|_2 > 0$ for all $h \in \mathcal{H}_0$, and $\widehat{\beta}_h = 0$ for all $h \in \mathcal{H}_0^c$. Then, the sparsistency property is

$$\mathbb{P}(\mathcal{O}) \to 1, \quad n \to \infty. \tag{9}$$

We will make the following assumptions:

(S1) the smallest eigenvalue of $\frac{1}{n}\left(\mathrm{X}_{\mathcal{H}_0}^\top \mathrm{X}_{\mathcal{H}_0}\right)$ is bounded below by a constant $C_{\min} > 0$;

(S2) letting $\alpha = \min_{h \in \mathcal{H}_0} \|\beta_h^0\|_\infty$,

$$\frac{1}{\alpha}\left[\sqrt{\frac{\log d_0}{n}} + \sqrt{d_0}\lambda \max_{h \in \mathcal{H}_0} \lambda_h\right] \to 0;$$

(S3) for some $0 < \epsilon < 1$ and every $h \in \mathcal{H}_0^c$,

$$\left\|\mathrm{X}_h^\top \mathrm{X}_{\mathcal{H}_0}\left(\mathrm{X}_{\mathcal{H}_0}^\top \mathrm{X}_{\mathcal{H}_0}\right)^{-1}\right\|_2 \leq \frac{1-\epsilon}{\sqrt{\sum_{h \in \mathcal{H}_0} \lambda_h^2}};$$

(S4)

$$\frac{1}{\lambda}\sqrt{\frac{\log(d-d_0)}{n}} \max_{h \in \mathcal{H}_0^c} \frac{\sqrt{d_h}}{\lambda_h} \to 0.$$

Theorem 4.2. *Under the assumptions (N) and (S1)-(S4), the sparsistency property (9) holds.*

Remarks. The conditions of Theorem 4.2 deserve a few comments.

1. From the proof, it can be seen that we can combine (S1) and (S2) into one assumption

$$\frac{1}{\alpha}\left[\sqrt{\frac{\log d_0}{nC_{\min}}} + \sqrt{d_0}\lambda \max_{h \in \mathcal{H}_0} \lambda_h\right] \to 0,$$

thus allowing the minimal eigenvalue of $\frac{1}{n}\left(\mathrm{X}_{\mathcal{H}_0}^\top \mathrm{X}_{\mathcal{H}_0}\right)$ to vanish at a rate slower than $\frac{1}{\alpha^2}\frac{\log d_0}{n}$.

2. The normality assumption (N) is by no means crucial. In fact, it is enough to require the errors to be independent, sub-gaussian random variables, with second moments bounded uniformly in n. Then, by applying, for example, Lemma 2.3 in Massart (2007), the conclusions of the Theorem would hold unchanged.
3. If $\lambda_h = \sqrt{d_h}$ then conditions (S4) simplifies to
$$\frac{\log(d-d_0)}{n\lambda^2} \rightarrow 0.$$
which is the same rate appearing in Equation 15 b) in Wainwright (2006) for the simpler lasso penalty.
4. It is apparent from condition (S4) that not only can d be much bigger then n, but it can in fact grow at at faster rate than n. In particular, condition (S4) formalizes quite explicitly the notion that the true model should be sparse in order for the group-lasso model selector to be successful.
5. Because the group-lasso solution may not be unique, Theorem 4.2 only implies the existence of a sequence of solutions guaranteeing sparsistency. In order to obtain a more satisfactory result, one may want to enforce also the uniqueness condition (U(c)), for some $c > 0$.

4.3. Inequalities for Prediction and Estimation

We now derive oracle inequalities for the prediction and estimation loss of the group-lasso estimator.

As a main technical step in our derivations (which generalizes standard arguments found, for example, in Bunea et al., 2007a,b; Bickel et al., 2007), the prediction and estimation bounds we establish hold on the event
$$\mathcal{A} = \bigcap_h \left\{ \frac{2}{\sqrt{n}} \left\| \mathrm{X}_h^\top \epsilon \right\|_2 < \sqrt{n}\lambda\lambda_h \right\}.$$

Therefore, we must impose conditions implying that $\mathcal{A}$ occurs with probability tending to 1, as both n and the model complexity increase. To that end, we formulate the asymptotic condition

(A) $\min_h \{ \frac{n}{\sigma^2}\lambda^2\lambda_h^2 - d_h \} - \log|\mathcal{H}| \rightarrow \infty$,

which will guarantee that the inequalities given below are meaningful for n large enough and also offers some characterizations of the rates of growth of the regularization parameters.

Lemma 4.3. *Assume (N) and (A). Then,* $\mathbb{P}(\mathcal{A}) \rightarrow 1$*, as* $n \rightarrow \infty$*.*

Remarks.

1. Assumption (A) provides general guidelines for choosing the tuning parameters λ and $\{\lambda_h, h \in \mathcal{H}\}$. In particular, if $\lambda_h = \sqrt{d_h}$ for each h, the condition reduces to
$$d_{\min}\frac{n}{\sigma^2}\lambda^2 - \log|\mathcal{H}| \rightarrow \infty,$$

where $d_{\min} = \min_h d_h$. For such a choice of λ_h, for example, we can use $\lambda = \sigma\sqrt{\frac{C_n}{n}}$, where C_n is such that

$$\frac{C_n d_{\min}}{\log|\mathcal{H}|} \to \infty.$$

Since $d_{\min} \geq 1$, we can set

$$\lambda = A\sigma\sqrt{\frac{\log|\mathcal{H}|}{n}},$$

for some $A > 1$.

2. Alternatively, and in less generality, if again $\lambda_h = \sqrt{d_h}$ for each h, we could consider the event

   $$\mathcal{A}' = \bigcap_{i=1}^{d} \left\{ \frac{2}{n} \left| \mathrm{X}_i^\top \epsilon \right| < \lambda \right\}.$$

   where X_i denotes the i-th column of the matrix X. Then, for

   $$\lambda = A\sigma\sqrt{\frac{\log d}{n}}$$

   with $A \geq 2$, a standard Gaussian tail bound (see, e.g., van de Geer, 2007, Lemma 3.8) yields

   $$1 - \mathbb{P}(\mathcal{A}') \leq 2\exp\left\{ -\frac{A^2}{8}\log d \right\},$$

   which vanishes provided $d \to \infty$. Notice that this case is covered by assumption (A). Then, using the event $\mathcal{A}'$ and Cauchy-Schwarz's inequality in equation (39) in the proof of Theorem 4.6, it is easy to see that the results of this section would hold with $\mathcal{A}$ replaced by $\mathcal{A}'$.
3. It appears that the Gaussianity assumption (N) is quite important in this context, as it is used in a fundamental way to establish condition (A). If, instead of the event $\mathcal{A}$, one considers the event $\mathcal{A}'$ (with the additional constraints $\lambda_h = \sqrt{d_h}$ for each h), then Gaussianity is not necessary and, with $\lambda = \sigma\sqrt{\frac{(\log d)^{(1+\eta)}}{n}}$ for some $\eta > 0$, one can still guarantee a vanishing probability for $\mathcal{A}'$ under the slightly stronger requirement $(\log d)^{(1+\delta)} = o(n)$ and some additional mild constraints. See Lounici (2008, Theorem 3) for a formal argument.

Another key assumption to our results is given below, where s is an integer and c a positive number:

(RE(s, c))

$$\min_{\mathcal{H}' \subseteq \mathcal{H}\colon |\mathcal{H}'| \leq s} \quad \min_{\beta \in \mathbb{R}^d\colon \gamma = \Lambda\beta, \sum_{h \in (\mathcal{H}')^c} \|\gamma_h\|_2 \leq c \sum_{h \in \mathcal{H}'} \|\gamma_h\|_2} \frac{\|\mathrm{X}\beta\|_2}{\sqrt{n}\|\beta_{\mathcal{H}'}\|_2} \equiv \kappa(h, c) > 0.$$

Here Λ is the $d \times d$ matrix with diagonal $\mathrm{vec}\{1_{d_h}\lambda_h, h \in \mathcal{H}\}$ and 1_{d_h} denotes the d_h-dimensional vector with entries all equal to 1. This assumption specializes the restricted eigenvalue assumption introduced by Bickel et al. (2007) to analyze the L_2 consistency property of the lasso procedure.

In particular, in the special case in which $s = |\mathcal{H}_0|$, the ($\mathrm{RE}(s, c)$) assumption is implied by the uniqueness assumption ($\mathrm{U}(c)$), as demonstrated in the next proposition.

Proposition 4.4. *Let* $\mathrm{U}(c, \delta)$ *be satisfied. Then, assumption* $\mathrm{RE}(|\mathcal{H}_0|, c)$ *holds.*

Our first result provides finite sample bounds for the prediction and estimation loss and for the number of non-zero blocks of the group-lasso estimator under the linear model (1), with unknown block-support set $\mathcal{H}_0$.

Theorem 4.5. *Assume (N) and* ($RE(|\mathcal{H}_0|, 3)$)*. On the event* $\mathcal{A}$*,*

$$\|\widehat{\beta} - \beta^0\|_2 \leq 16 \frac{\lambda}{\kappa_0^2 \lambda_{\min}} |\mathcal{H}_0|, \tag{10}$$

where $\lambda_{\min} = \min_h \lambda_h$*,*

$$\frac{1}{n}\|\mathrm{X}(\widehat{\beta} - \beta^0)\|_2^2 \leq \frac{16\lambda^2|\mathcal{H}_0|}{\kappa_0^2}, \tag{11}$$

and

$$|\widehat{\mathcal{H}}| \leq 64 \frac{C_{\max}|\mathcal{H}_0|}{\kappa_0^2 \lambda_{\min}^2}, \tag{12}$$

where $C_{\max}$ *is the largest eigenvalue of* $\frac{1}{n}\left(\mathrm{X}_{\mathcal{H}_0}^\top \mathrm{X}_{\mathcal{H}_0}\right)$ *and* $\kappa_0 = \kappa(|\mathcal{H}_0|, 3)$*.*

Next, we establish a more general oracle inequality for the prediction loss of the group-lasso estimator which covers the case of a mispecified model. Specifically, rather than assuming that the true model is linear, we consider the more general model

$$Y = f^0(\mathrm{X}) + \epsilon,$$

for some unknown, possibly non-linear, function f^0 of the covariates.

Theorem 4.6. *Under the assumptions (N) and RE(*$s, 3 + 4/\epsilon$*), on the event* $\mathcal{A}$*,*

$$\frac{1}{n}\|\mathrm{X}\widehat{\beta} - f^0(\mathrm{X})\|_2^2 \leq \inf_{\beta\colon\, |\mathcal{H}(\beta)| \leq s} \left\{ (1+\epsilon)\frac{1}{n}\|\mathrm{X}\beta - f^0(\mathrm{X})\|_2^2 + \frac{4}{\kappa^2}\left(\frac{4}{\epsilon} + \epsilon + 4\right)\lambda^2|\mathcal{H}(\beta)| \right\}, \tag{13}$$

where $\mathcal{H}(\beta) = \{h\colon \beta_h \neq 0\}$ *and* $\kappa = \kappa(s, 3 + 4/\epsilon)$*.*

Remarks.

1. Recall that, under our assumption (A), the event $\mathcal{A}^c$ has vanishing probability, so the bounds we obtain holds with large probability, for n big enough.

2. In both Theorems (4.6) and (13), we do not enforce the uniqueness condition (U(c)), and, therefore, the conclusions hold for *any* solution to (2). In fact, because of Proposition 4.4, we can replace the RE(s, c) conditions in both Theorems (4.6) and (13) by the appropriate U(c) conditions, which would guarantee the same results and also uniqueness of the group-lasso estimator.
3. The inequalities derived above directly generalize the corresponding bounds established by Bunea et al. (2007a) and Bickel et al. (2007) for the lasso problem.
4. From both Theorems, it is possible to get rates of prediction and estimation consistency of the group-lasso. These rates depend crucially on the choice of the tuning parameters compatible with assumption (A), in particular of λ. See Remark 1. after Lemma 4.3 for some comments on the possible values for λ. In particular, for $\lambda_h = \sqrt{d_h}$ and $\lambda = A\sigma\sqrt{\frac{\log|\mathcal{H}|}{n}}$, for some $A > 1$, we obtain rates that are comparable to lasso rates, with the number of parameters replaced by the number of blocks. This is due to the nature of our assumption (RE(s, c)).

4.4. Persistence

In this final section, we change our settings and adopt the double-asymptotic framework of Greenshtein and Ritov (2006) and Greenshtein (2006). Our goal is to study the risk consistency of the group-lasso solutions under a triangular array framework for the random vector $Q = (Y, X)$, where Y is the response variable and $X = (X_1, \dots, X_d)$ the vector of covariates. We are concerned with the predictive risk $R(\beta) = \mathbb{E}(Y - X\beta)^2$, where the expectation is with respect to the joint distribution $P_{(X,Y)}$ of Y and X.

Specifically, let $\widehat{\beta}_n$ be an estimator based on an *iid* sample $(Q^1, \dots, Q^n)$ of size n from $P_{(X,Y)}$ and let $R(\widehat{\beta}) = \mathbb{E}(Y - \widehat{\beta}X|Q^1, \dots, Q^n)$, for a new *iid* observation $(Y, X) \sim P_{(Y,X)}$. Just like above, we allow d to grow unbounded with n. Let $\{S_n\}$ be a sequence of sets of increasing dimensions. A sequence of estimators $\{\widehat{\beta}_n\}$ is said to be persistent with respect to $\{S_n\}$ if

$$R(\widehat{\beta}_n) - \inf_{\beta \in S_n} R(\beta) \xrightarrow{p} 0, \quad n \to \infty.$$

Notice that, in order for persistence to hold, it is not necessary for the best predictor of Y based on X to be linear.

We assume that the random covariates X have a grouping structure, which we represent using the same notation and conventions of Section 2. Accordingly, we consider the following two sequences of sets, each of them providing a different form of group penalty:

$$B_n = \left\{ \beta \colon \sum_h \sqrt{d_h} \|\beta_h\|_2 \leq b_n \right\}$$

and

$$C_n = \left\{ \beta \colon |\{h, \beta_h \neq 0\}| \leq c_n \right\},$$

for some sequences of numbers $\{b_n\}$ and $\{c_n\}$ to be determined.

Letting $\gamma = (-1, \beta_1, \ldots, \beta_p)$, we can write $R(\beta) = \gamma^\top \Sigma \gamma$, where $\Sigma = \mathbb{E} QQ^\top$. The empirical equivalent of this quantity is $\widehat{R}(\beta) = \gamma^\top \widehat{\Sigma} \gamma$, where

$$\widehat{\Sigma} = \frac{1}{n} \sum_i Q^i (Q^i)^\top.$$

In these new settings, the group-lasso estimator $\widehat{\beta}$ with respect to the sequence $\{S_n\}$ of sets of potential coefficients, which can be $\{B_n\}$ or $\{C_n\}$, is computed as

$$\widehat{\beta} = \min_{\beta \in S_n} \widehat{R}(\beta). \tag{14}$$

Following Zhou et al. (2007), we impose the conditions

(P1) $\mathbb{E}|Z_{jk}|^q \leq q! A^{q-2} B/2$, for each $j, k = 1, \ldots, d+1$;
(P2) $d \leq e^{n^\alpha}$,

where $Z_{jk} = Q_j Q_k - \mathbb{E} Q_j Q_k$, and A, B and α are some positive constants with $0 < \alpha \leq 1$.

Theorem 4.7. *Under the assumptions (P1) and (P2), the group lasso estimator defined in (14) is persistent with respect to $\{B_n\}$ if*

$$b_n = o\left(\left(\frac{n}{\log n} \right)^{1/4} \right). \tag{15}$$

It is persistent with respect to $\{C_n\}$ if

$$c_n = o\left(\left(\frac{n}{d_{\max}^2 \log n} \right)^{1/2} \right) \tag{16}$$

and the minimal eigenvalue of the covariance matrix of the predictors is positive.

Remarks.

1. Notice that (15) is implied by the stronger condition

$$b_n = o\left(\left(\frac{n}{d_{\max}^2 \log n} \right)^{1/4} \right),$$

which is of the same form as (16).

2. The definition of the set sequence $\{B_n\}_n$ can be generalized to

$$B_n = \left\{ \beta \colon \sum_h \lambda_h \|\beta_h\|_2 \leq b_n \right\},$$

and the results of Theorem 4.7 would remain true provided $\max_h \frac{\sqrt{d_h}}{\lambda_h} = O(1)$.

3. The persistence results (15) and (16) are easy generalizations of their lasso equivalents derived in Greenshtein and Ritov (2006) and Greenshtein (2006), the only difference being the additional term $d_{\max}$. For the choice $\lambda_h = \sqrt{d_h}$, for each h, this is precisely the extra term appearing also in Theorem 4.6.
4. Assumptions (P1) and (P2) are not the only options. See Greenshtein and Ritov (2006) and Greenshtein (2006) for alternative assumptions and derivations.

5. Proofs

Proof of Proposition 3.1. For every $h \in \mathcal{H}$, let $\lim \sqrt{n}\lambda\lambda_h = c_h \geq 0$. Next, we observe that, for any $u \in \mathbb{R}^d$, as $n \to \infty$,

$$\sqrt{n}\left(\left\|\beta^0_h + \frac{1}{\sqrt{n}}u_h\right\|_2 - \left\|\beta^0_h\right\|_2\right) \to \begin{cases} \frac{u_h^\top \beta^0_h}{\|\beta^0_h\|_2} & \text{if } \beta^0_h \neq 0 \\ \|u_h\|_2 & \text{if } \beta^0_h = 0. \end{cases}$$

Then, by the same arguments used in the proof of Theorem 3.2 below and by equation (18),

$$\sqrt{n}\left(\widehat{\beta} - \beta^0\right) \xrightarrow{d} u^* \equiv \operatorname{argmin}_{u \in \mathbb{R}^d} V(u),$$

where

$$V(u) = \frac{1}{2}u^\top \mathrm{M} u - 2u^\top W + \sum_h c_h \left(\frac{u_h^\top \beta^0_h}{\|\beta^0_h\|_2} 1_{\{\beta^0_h \neq 0\}} + \|u_h\|_2 1_{\{\beta^0_h = 0\}}\right),$$

with $W \sim N_d(0, \sigma^2\mathrm{M})$ (see also Knight and Fu, 2000; Zou, 2006).

We will prove the claim by showing that if $c_{h'} \neq \infty$ for some $h' \notin \mathcal{H}_0$, then

$$\mathbb{P}\left\{u^*_{h'} = 0\right\} < 1, \tag{17}$$

which will contradict the assumed model selection consistency (6). The optimal solution u^* must satisfy the first order optimality conditions

$$-2W_{\mathcal{H}_0} + \mathrm{M}_{\mathcal{H}_0} u^*_{\mathcal{H}_0} + \eta_{\mathcal{H}_0} = 0,$$

where $\eta_{\mathcal{H}_0} = \operatorname{vec}\left\{c_h \frac{\beta^0_h}{\|\beta^0_h\|_2}, h \in \mathcal{H}_0\right\}$, and

$$\| - 2W_h + \mathrm{M}_{h,\mathcal{H}_0} u^*_{\mathcal{H}_0}\|_2 \leq c_h, \quad \forall h \notin \mathcal{H}_0,$$

which together imply

$$\| - 2W_h + \mathrm{M}_{h,\mathcal{H}_0}\mathrm{M}^{-1}_{\mathcal{H}_0}\left(2W_{\mathcal{H}_0} - \eta_{\mathcal{H}_0}\right)\|_2 \leq c_h, \quad \forall h \notin \mathcal{H}_0.$$

Then, since $c_{h'} < \infty$,

$$\mathbb{P}\left\{u^*_{h'} = 0\right) \leq \mathbb{P}\left(\| - 2W_{h'} + \mathrm{M}_{h',\mathcal{H}_0}\mathrm{M}^{-1}_{\mathcal{H}_0}\left(2W_{\mathcal{H}_0} - \eta_{\mathcal{H}_0}\right)\|_2 \leq c_{h'}\right\} < 1,$$

thus proving (17). □

Proof of Theorem 3.2. We first show (8). Letting $\beta_n = \beta^0 + \frac{u}{\sqrt{n}}$, where $u \in \mathbb{R}^d$, the objective function (2) (multiplied by n) can be written as a function of u as

$$Q_n(u) = \frac{1}{2}\left\| \frac{1}{\sqrt{n}} \mathrm{X} u + \epsilon \right\|_2^2 + \sum_h n\lambda\lambda_h \left\| \beta_h^0 + \frac{1}{\sqrt{n}} u_h \right\|_2,$$

where $u = \sqrt{n}(\beta_n - \beta^0)$. Let $u^* = \mathrm{argmin}_{u \in \mathbb{R}^d} Q_n(u)$ and observe that $u^* = \sqrt{n}(\widehat{\beta} - \beta^0)$, where $\widehat{\beta}$ is the minimizer of (2). Next, write

$$\begin{array}{rcl} D_n(u) & = & Q_n(u) - Q_n(0) \\ & = & \left(\frac{1}{2} u^\top (\frac{1}{n}\mathrm{X}^\top \mathrm{X}) u - \frac{1}{\sqrt{n}} u^\top \mathrm{X}\epsilon \right) + \sqrt{n}\lambda \sum_h \lambda_h \sqrt{n} \left(\left\| \beta_h^0 + \frac{1}{\sqrt{n}} u_h \right\|_2 - \|\beta_h^0\|_2 \right) \\ & \equiv & I_{1,n} + \sum_h I_{2,n,h}. \end{array} \tag{18}$$

Note that D_n is strictly convex. If $\beta_h^0 \neq 0$, then $\lambda_h \xrightarrow{p} \frac{1}{\|\beta_h^0\|_2^\gamma}$ and

$$\sqrt{n}\left(\left\| \beta_h^0 + \frac{1}{\sqrt{n}} u_h \right\|_2 - \|\beta_h^0\|_2 \right) \to \frac{u_h^\top \beta_h^0}{\|\beta_h^0\|_2},$$

and, therefore, $I_{2,n,h}$ converges in probability to 0 by Slutsky theorem and the assumption $\sqrt{n}\lambda = o(1)$. If $\beta_h^0 = 0$, then, since $n^{\gamma/2}\|\widehat{\beta}_h^{\mathrm{OLS}}\|_2^\gamma = O_P(1)$ and $\sqrt{n}\left(\left\| \beta_h^0 + \frac{1}{\sqrt{n}} u_h \right\|_2 - \|\beta_h^0\|_2 \right) = \|u_h\|_2$, we obtain

$$I_{2,n,h} = \sqrt{n}\lambda\lambda_h \|u_h\|_2 = \|u_h\|_2 \lambda \frac{n^{(\gamma+1)/2}}{(\sqrt{n}\|\widehat{\beta}_h^{\mathrm{OLS}}\|_2)^\gamma} \xrightarrow{p} \infty, \tag{19}$$

where the second assumption in the statement was used. Because $I_{1,n} \xrightarrow{d} \frac{1}{2} u^\top \mathrm{M} u + W$, where $W \sim N_d(0, \sigma^2 \mathrm{M})$, and $|\mathcal{H}|$ is finite, it follows that $D_n(u) \xrightarrow{d} D(u)$, with

$$D(u) = \begin{cases} \frac{1}{2} u_{\mathcal{H}_0}^\top \mathrm{M} u_{\mathcal{H}_0} - 2 u_{\mathcal{H}_0}^\top W & \text{if } u_{\mathcal{H}_0} \neq 0 \\ \infty & \text{otherwise.} \end{cases}$$

The unique minimizer of $D(u)$ is $(\mathrm{M}_{\mathcal{H}_0}^{-1} W, 0)^\top$. By the argmax theorem in van der Vaart and Wellner (1998, Corollary 3.2.3) (or alternatively, the results in Geyer, 1994),

$$u_{\mathcal{H}_0}^{(n)} \xrightarrow{d} \mathrm{M}_{\mathcal{H}_0}^{-1} W \sim N_{d_0}(0, \sigma^2 \mathrm{M}_0^{-1})$$

and

$$u_{\mathcal{H}_0^c}^{(n)} \xrightarrow{p} 0,$$

and (8) is verified.

Next, we prove model selection consistency (6). Since $\widehat{\beta}$ is $\sqrt{n}$−consistent, for each $h \in \mathcal{H}_0$, $\widehat{\beta}_h \neq 0$ with arbitrarily high probability for sufficiently large n. Thus, we only need to show that, for each $h \notin \mathcal{H}_0$, $\widehat{\beta}_h = 0$ with arbitrarily high probability for sufficiently large n. Model selection consistency will then follow from the finiteness of $|\mathcal{H}_0|$. Suppose that, for some $h \notin \mathcal{H}_0$, $\widehat{\beta}_h \neq 0$. Then, from the subgradient conditions (5),

$$\mathrm{X}_h^\top \left(Y - \mathrm{X}\widehat{\beta} \right) = n\lambda\lambda_h \frac{\widehat{\beta}_h}{\|\widehat{\beta}_h\|_2}. \tag{20}$$

Because of $\sqrt{n}(\beta^0 - \widehat{\beta})$ is asymptotically normally distributed, and using our assumption on the design matrix,

$$\left\| \frac{1}{\sqrt{n}} \mathrm{X}_h^\top \left(Y - \mathrm{X}\widehat{\beta} \right) \right\|_2 = \left\| \frac{1}{n} \mathrm{X}_h^\top \mathrm{X} \sqrt{n}(\beta^0 - \widehat{\beta}) + \frac{1}{\sqrt{n}} \mathrm{X}_h^\top \epsilon \right\|_2 = O_P(1).$$

Furthermore, by the same arguments leading to (19), $\sqrt{n}\lambda\lambda_h \overset{p}{\rightarrow} \infty$. Then, the norm of the terms on two sides of equation (20) have different order of magnitude, as $n \rightarrow \infty$, which implies that $\widehat{\beta}_h$ does not satisfy that first order condition for being non-zero with increasing probability, and therefore $\widehat{\beta}_h = 0$ with probability tending to 1. □

Proof of Theorem 3.3. In the first part of the proof, we follow Fan and Li (2001). Let $\alpha_n = \frac{1}{\sqrt{n}} + \lambda a_n$ and $p_h \colon \mathbb{R}^{d_h} \rightarrow \mathbb{R}$ be a random function given by

$$p_h(x) = \lambda_h \|x\|_2$$

and write

$$Q(\beta) = \frac{1}{2}\|Y - \mathrm{X}\beta\|_2^2 + n\lambda \sum_h p_h(\beta_h),$$

and

$$D(\alpha_n u) = Q(\beta^0 + \alpha_n u) - Q(\beta^0),$$

for $u \in \mathbb{R}^d$. We will show that, for each ϵ, there exists a constant C such that, for large enough n,

$$\mathbb{P}\left\{ \inf_{u \colon \|u\|_2 = C} D(\alpha_n u) > 0 \right\} > 1 - \epsilon,$$

which implies the existence of a local minimizer inside the ball $\{\beta^0 + \alpha u \colon \|u\|_2 \leq C\}$ and therefore a solution $\widehat{\beta}$ such that $\|\widehat{\beta} - \beta^0\|_2 = O_P(\alpha_n)$. Since $p_h(0) = 0$ and $Y = \mathrm{X}\beta^0 + \epsilon$, we have

$$D(\alpha_n, u) \geq \frac{1}{2}\|Y - \mathrm{X}(\beta^0 - \alpha_n u)\|_2^2 - \frac{1}{2}\|Y - \mathrm{X}\beta^0\|_2^2 + n\lambda \sum_{h \in \mathcal{H}_0} \left(p_h(\beta_h^0 + \alpha_n u_h) - p_h(\beta_h^0) \right) \tag{21}$$

The first two terms in (21) can be written as

$$\frac{1}{2}\|\epsilon - \alpha_n \mathrm{X}u\|_2^2 - \frac{1}{2}\|\epsilon\|_2^2 = \frac{1}{2}\alpha_n^2 n u^\top \mathrm{M} u(1 + o(1)) - \alpha_n \langle \epsilon, \mathrm{X}u \rangle,$$

from which it follows easily that they are of order

$$O\left(\alpha_n^2 n\|u\|_2^2\right) - O_P\left(\alpha_n\sqrt{n}\|u\|_2\right). \tag{22}$$

The last term on the right hand side of (21) can be bounded as follows:

$$\begin{array}{rcl} n\lambda\sum_{h\in\mathcal{H}_0}\left(p_h(\beta_h^0+\alpha_n u_h)-p_h(\beta_h^0)\right) & \geq & -\lambda n\sum_{h\in\mathcal{H}_0}\lambda_h\alpha_n\|u_h\|_2 \\ & \geq & -\|u\|_2\alpha_n\lambda n a_n|\mathcal{H}_0| \\ & = & -\|u\|_2 O_P(n\alpha_n^2). \end{array}$$

Combining the previous display with (21) and (22), one can conclude that, for sufficiently large $C=\|u\|_2$, the positive term $O\left(\alpha_n^2 n\|u\|_2^2\right)$ dominates all the others.
If $\lambda a_n = O\left(\frac{1}{\sqrt{n}}\right)$, then $\sqrt{n}\left(\widehat{\beta}-\beta^0\right) = O_P(1)$. Then, since $\sqrt{n}\lambda\lambda_h \to \infty$ for each $h \notin \mathcal{H}_0$, the model selection consistency (6) follows from the same arguments used at the end of the proof of Theorem 3.2. Since the event $\{\widehat{\mathcal{H}} \neq \mathcal{H}\}$ has vanishing probability, asymptotic normality (8) is easily proved by restricting to the complementary event $\{\widehat{\mathcal{H}} = \mathcal{H}\}$ and applying the central limit theorem and Slutsky's theorem to equation (23) below, taking into account fact that $\lambda a_n \to 0$. □

Proof of Proposition 4.1. Let $\beta=\beta^1-\beta^2$. Then $X\beta=0$. Assume that $\beta\neq 0$. Using the same notation as in Proposition 4.4 with $s=2|\mathcal{H}_0|$, we get, by equation (30),

$$\frac{\|\mathrm{X}\beta\|_2^2}{n\|\gamma\|_2^2} \geq \frac{1}{\lambda_{\max}^2} - \frac{1}{(1+2c)\delta} > 0,$$

which gives a contradiction, since $\delta > \lambda_{\max}^2$. □

Proof of Theorem 4.2. The proof is an adaptation to the present settings of arguments use in Wainwright (2006). Let $\widehat{\mathcal{H}} = \{h\colon \widehat{\beta}_h \neq 0\}$ and set

$$\widehat{\eta}_{\mathcal{H}} = \mathrm{vec}\left\{\lambda_h\frac{\widehat{\beta}_h}{\|\widehat{\beta}_h\|_2}, h\in\widehat{\mathcal{H}}\right\}$$

and

$$\widehat{\eta}_{\mathcal{H}^c} = \mathrm{vec}\left\{\lambda_h z_h, h\in\widehat{\mathcal{H}}^c\right\},$$

where $\|z_h\|_2 \leq 1$. Using the subgradient conditions, the event $\mathcal{O}$ holds if and only if

$$\widehat{\beta}_{\mathcal{H}_0} = \beta_{\mathcal{H}_0} + \left(\frac{1}{n}\mathrm{X}_{\mathcal{H}_0}^\top\mathrm{X}_{\mathcal{H}_0}\right)^{-1}\left(\frac{1}{n}\mathrm{X}_{\mathcal{H}_0}^\top\epsilon - \lambda\widehat{\eta}_{\mathcal{H}_0}\right) \tag{23}$$

and

$$\lambda\widehat{\eta}_{\mathcal{H}_0^c} = \frac{1}{n}\mathrm{X}_{\mathcal{H}_0^c}^\top\epsilon + \frac{1}{n}\mathrm{X}_{\mathcal{H}_0^c}^\top\mathrm{X}_{\mathcal{H}_0}\left(\frac{1}{n}\mathrm{X}_{\mathcal{H}_0}^\top\mathrm{X}_{\mathcal{H}_0}\right)^{-1}\left(\lambda\widehat{\eta}_{\mathcal{H}_0} - \frac{1}{n}\mathrm{X}_{\mathcal{H}_0}^\top\epsilon\right) \tag{24}$$

We will use equations (23) and (24) to show

$$\mathbb{P}\left\{\|\widehat{\beta}_{\mathcal{H}_0} - \beta^0_{\mathcal{H}_0}\|_\infty < \alpha\right\} \to 1, \quad n \to \infty \tag{25}$$

and

$$\mathbb{P}\left\{\frac{1}{\lambda_h}\|\widehat{\eta}_h\|_2 < 1, \forall h \in \mathcal{H}_0^c\right\} \to 1, \quad n \to \infty, \tag{26}$$

respectively, where we recall that $\alpha = \min_{h \in \mathcal{H}_0} \|\beta^0_h\|_\infty$. In turn, (25) and (26) imply

$$\mathbb{P}(\mathcal{O}) \to 1, \quad n \to \infty,$$

as claimed. We begin with (25). Write, for simplicity, $\Sigma_0 = \frac{1}{n}\mathrm{X}^\top_{\mathcal{H}_0}\mathrm{X}_{\mathcal{H}_0}$ and consider the d_0-dimensional vector $Z = \Sigma_0^{-1}\frac{1}{n}\mathrm{X}^\top_{\mathcal{H}_0^c}\epsilon$. Then, $\mathbb{E}Z = 0$ and $\mathbb{V}Z = \frac{\sigma^2}{n}\Sigma_0^{-1}$, so that $\mathbb{V}Z_i \leq \frac{\sigma^2}{nC_{\min}}$ for each coordinate i of Z. Using standard results on the maximum of a Gaussian vector (see, e.g., Ledoux and Talagrand, 1991),

$$\mathbb{E}\|Z\|_\infty \leq 3\sigma\sqrt{\frac{\log d_0}{nC_{\min}}} \tag{27}$$

As for the second term on the right hand side of (23), we obtain

$$\lambda\|\Sigma_0^{-1}\widehat{\eta}_{\mathcal{H}_0}\|_\infty \leq \lambda\|\Sigma_0^{-1}\|_\infty\|\widehat{\eta}_{\mathcal{H}_0}\|_\infty \leq \lambda\frac{\sqrt{d_0}}{C_{\min}}\max_{h \in \mathcal{H}_0}\lambda_h, \tag{28}$$

where in the last inequality we use the bounds

$$\|\widehat{\eta}_{\mathcal{H}_0}\|_\infty \leq \max_{h \in \mathcal{H}_0}\|\widehat{\eta}_h\|_2 \leq \max_{h \in \mathcal{H}_0}\lambda_h,$$

and

$$\|\Sigma_0^{-1}\|_\infty \leq \sqrt{d_0}\|\Sigma_0^{-1}\|_2 \leq \frac{\sqrt{d_0}}{C_{\min}}.$$

By Markov inequality, and using (27) and (28),

$$\begin{aligned}\mathbb{P}\left(\|\widehat{\beta}_{\mathcal{H}_0} - \beta^0_{\mathcal{H}_0}\|_\infty > \alpha\right) &\leq \tfrac{\mathbb{E}\|\widehat{\beta}_{\mathcal{H}_0} - \beta^0_{\mathcal{H}_0}\|_\infty}{\alpha}\\ &\leq \tfrac{1}{\alpha}\left[\mathbb{E}\|Z\|_\infty + \lambda\|\Sigma_0^{-1}\widehat{\eta}_{\mathcal{H}_0}\|_\infty\right]\\ &\leq \tfrac{1}{\alpha}\left[3\sigma\sqrt{\tfrac{\log d_0}{nC_{\min}}} + \lambda\tfrac{\sqrt{d_0}}{C_{\min}}\max_{h \in \mathcal{H}_0}\lambda_h\right],\end{aligned}$$

which goes to zero under (S2), thus establishing (25).

Next, we show (26). Rewrite (24) as

$$\widehat{\eta}_{\mathcal{H}_0^c} = K + \frac{1}{\lambda}W,$$

where

$$K = \frac{1}{n}\mathrm{X}^\top_{\mathcal{H}_0^c}\mathrm{X}_{\mathcal{H}_0}\left(\frac{1}{n}\mathrm{X}^\top_{\mathcal{H}_0}\mathrm{X}_{\mathcal{H}_0}\right)^{-1}\widehat{\eta}_{\mathcal{H}_0}$$

and

$$W = \mathrm{X}_{\mathcal{H}_0^c}^\top \left[\mathrm{I} - \mathrm{X}_{\mathcal{H}_0} \left(\frac{1}{n} \mathrm{X}_{\mathcal{H}_0}^\top \mathrm{X}_{\mathcal{H}_0} \right)^{-1} \mathrm{X}_{\mathcal{H}_0}^\top \right] \frac{1}{n} \epsilon.$$

Then, for any $h \in \mathcal{H}_0^c$,

$$\frac{1}{\lambda_h} \|\widehat{\eta}_h\|_2 \leq \frac{1}{\lambda_h} \|K_h\|_2 + \frac{1}{\lambda} \frac{\sqrt{d_h}}{\lambda_h} \|W_h\|_\infty. \tag{29}$$

We bound the first term in the previous equation as follows,

$$\frac{1}{\lambda_h} \|K_h\|_2 \leq \left\| X_h^\top \mathrm{X}_{\mathcal{H}_0} \left(\mathrm{X}_{\mathcal{H}_0}^\top \mathrm{X}_{\mathcal{H}_0} \right)^{-1} \right\|_2 \sqrt{\sum_{h \in \mathcal{H}_0} \lambda_h^2} < 1 - \epsilon,$$

with the last inequality stemming from assumption (S3). As for the second term in (29), notice that $\mathbb{E} W_h = 0$ and

$$\mathbb{V} W_h = \frac{\sigma^2}{n} \mathrm{X}_h^\top \left[\mathrm{I} - \mathrm{X}_{\mathcal{H}_0} \left(\frac{1}{n} \mathrm{X}_{\mathcal{H}_0}^\top \mathrm{X}_{\mathcal{H}_0} \right)^{-1} \mathrm{X}_{\mathcal{H}_0}^\top \right] \mathrm{X}_h \leq \frac{\sigma^2}{n} \|\mathrm{X}_h\|_2^2 = \frac{\sigma^2}{n}.$$

By the same arguments used above,

$$\mathbb{E} \|W\|_\infty \leq 3\sigma \sqrt{\frac{\log(d - d_0)}{n}},$$

hence, in virtue of Markov's inequality,

$$\mathbb{P}\left(\frac{1}{\lambda} \max_{h \in \mathcal{H}_0^c} \frac{\sqrt{d_h}}{\lambda_h} \|W\|_\infty > \frac{\epsilon}{2} \right) \leq 6 \max_{h \in \mathcal{H}_0^c} \frac{\sqrt{d_h}}{\lambda_h} \frac{\sigma}{\lambda \epsilon} \sqrt{\frac{\log(d - d_0)}{n}}.$$

Therefore, using assumption (S4),

$$\mathbb{P}\left(\|\widehat{\eta}_h\|_2 > 1 - \epsilon/2, \text{ for some } h \in \mathcal{H}_0^c \right) \to 0,$$

which gives (26). The proof is now complete. □

Proof of Theorem 4.6. The proof follows closely (Bickel et al., 2007, Theorem 5.1) and is essentially based on Lemma 6.1 in the Appendix. Let $\beta \in \mathbb{R}^d$ be arbitrary, with $\mathcal{H}(\beta) \leq s$. On the event $\mathcal{A}$, if

$$4\lambda \sum_{h \in \mathcal{H}'} \lambda_h \|\widehat{\beta}_h - \beta_h\|_2 \leq \epsilon \frac{1}{n} \|\mathrm{X}(\widehat{\beta} - \beta^0)\|_2^2,$$

the claim holds trivially from the first inequality in (37). Consider instead the complementary case

$$\mathcal{A}_1 = \left\{ \epsilon \frac{1}{n} \|\mathrm{X}(\widehat{\beta} - \beta^0)\|_2^2 < 4\lambda \sum_{h \in \mathcal{H}'} \lambda_h \|\widehat{\beta}_h - \beta_h\|_2 \right\}.$$

On the event $\mathcal{A} \cap \mathcal{A}_1$, from the first inequality in (37), we get

$$\sum_{h \in (\mathcal{H}')^c} \lambda_h \|\widehat{\beta}_h - \beta_h\|_2 \leq \left(3 + \frac{4}{\epsilon}\right) \sum_{h \in \mathcal{H}'} \lambda_h \|\widehat{\beta}_h - \beta_h\|_2.$$

Using the assumption $\mathrm{RE}(s, 3 + 4/\epsilon)$,we obtain, still on $\mathcal{A} \cap \mathcal{A}_1$,

$$\sum_{h \in \mathcal{H}'} \lambda_h^2 \|\widehat{\beta}_h - \beta_h\|_2^2 \leq \frac{1}{\kappa^2} \frac{1}{n} \|\mathrm{X}(\widehat{\beta} - \beta)\|_2^2.$$

Thus, by the second inequality in (37), on $\mathcal{A} \cap \mathcal{A}_1$,

$$\frac{1}{n}\|\mathrm{X}(\widehat{\beta}-\beta^0)\|_2^2 \leq \frac{1}{n}\|\mathrm{X}(\beta-\beta_0)\|_2^2 + 4\lambda \frac{\sqrt{|\mathcal{H}(\beta)|}}{\kappa} \left(\frac{1}{\sqrt{n}} \|\mathrm{X}(\widehat{\beta} - \beta^0)\|_2 + \frac{1}{\sqrt{n}} \|\mathrm{X}\beta^0 - \mathrm{X}\beta\|_2 \right).$$

This expression is of the same form as inequality (A.3) in Bunea et al. (2007a). Following their arguments, we get that, for any $a > 1$,

$$\frac{1}{n}\left(\|\mathrm{X}(\widehat{\beta} - \beta^0)\|_2^2\right) \leq \frac{a+1}{a-1}\frac{1}{n}\|\mathrm{X}(\beta - \beta^0)\|_2^2 + \frac{8a^2}{\kappa^2(a-1)}\lambda^2 |\mathcal{H}(\beta)|,$$

and (13) is established by setting $\epsilon = \frac{2}{a-1}$. □

Proof of Proposition 4.4. We adapt the arguments used in Lounici (2008, Lemma 2). Let $\beta \in \mathbb{R}^d$ such that $\mathcal{H}' \equiv \mathcal{H}(\beta)$, $|\mathcal{H}'| \leq s$ and $\sum_{(\mathcal{H}')^c} \|\gamma_h\|_2 \leq c \sum_{\mathcal{H}'} \|\gamma_h\|_2$, where $\gamma = \Lambda\beta$. Then,

$$\begin{aligned} \frac{\|\mathrm{X}\beta_{\mathcal{H}'}\|_2^2}{n\|\gamma_{\mathcal{H}'}\|_2^2} &\geq \frac{1}{\lambda^2_{\max}} - \frac{1}{\|\gamma_{\mathcal{H}'}\|_2^2} \sum_{h,h' \in \mathcal{H}'} \beta_h^\top \left(\tfrac{1}{n}\mathrm{X}_h^\top \mathrm{X}_{h'}\right) \beta_{h'} \\ &\geq \frac{1}{\lambda^2_{\max}} - \frac{1}{(2c+1)\delta s} \frac{\left(\sum_{h \in \mathcal{H}'} \lambda_h \|\beta_h\|_2\right)^2}{\|\gamma_{\mathcal{H}'}\|_2^2}, \end{aligned} \tag{30}$$

where assumption (U) is used in the second inequality. Denoting with $\mathrm{X}_{\mathcal{H}'}$ the submatrix of X comprised by $\{\mathrm{X}_h, h \in \mathcal{H}'\}$, the last inequality yields

$$\begin{aligned} \frac{\|\mathrm{X}\beta\|_2^2}{n\|\gamma_{\mathcal{H}'}\|_2^2} &\geq \frac{\|\mathrm{X}\beta_{\mathcal{H}'}\|_2^2}{n\|\gamma_{\mathcal{H}'}\|_2^2} + 2\frac{\beta_{\mathcal{H}'}^\top \mathrm{X}_{\mathcal{H}'}^\top \mathrm{X}_{(\mathcal{H}')^c}\beta_{(\mathcal{H}')^c}}{n\|\gamma_{\mathcal{H}'}\|_2^2} \\ &\geq \frac{1}{\lambda^2_{\max}} - \frac{1}{(2c+1)\delta s}\frac{\left(\sum_{h\in\mathcal{H}'} \lambda_h \|\beta_h\|_2\right)^2}{\|\gamma_{\mathcal{H}'}\|_2^2} - \frac{2}{\|\gamma_{\mathcal{H}'}\|_2^2} \sum_{h\in\mathcal{H}', h' \in (\mathcal{H}')^c} \beta_h^\top \left(\tfrac{1}{n}\mathrm{X}_h^\top \mathrm{X}_{h'}\right)\beta_{h'} \\ &\geq \frac{1}{\lambda^2_{\max}} - \frac{1}{(2c+1)\delta s}\frac{\left(\sum_{h\in\mathcal{H}'} \lambda_h \|\beta_h\|_2\right)^2}{\|\gamma_{\mathcal{H}'}\|_2^2} - \frac{2}{(2c+1)\delta s \|\gamma_{\mathcal{H}'}\|_2^2} \left(\sum_{h\in\mathcal{H}'} \lambda_h \|\beta_h\|_2\right)\left(\sum_{h\in(\mathcal{H}')^c} \lambda_h\|\beta_h\|_2\right) \\ &= \frac{1}{\lambda^2_{\max}} - \frac{1}{(2c+1)\delta s}\frac{\left(\sum_{h\in\mathcal{H}'} \|\gamma_h\|_2\right)^2}{\|\gamma_{\mathcal{H}'}\|_2^2} - \frac{2}{(2c+1)\delta s}\frac{\left(\sum_{h\in\mathcal{H}'} \|\gamma_h\|_2\right)\left(\sum_{h\in(\mathcal{H}')^c} \|\gamma_h\|_2\right)}{\|\gamma_{\mathcal{H}'}\|_2^2} \\ &\geq \frac{1}{\lambda^2_{\max}} - \frac{1}{(2c+1)\delta s}\frac{\left(\sum_{h\in\mathcal{H}'} \|\gamma_h\|_2\right)^2}{\|\gamma_{\mathcal{H}'}\|_2^2} - \frac{2c}{(2c+1)\delta s}\frac{\left(\sum_{h\in\mathcal{H}'} \|\gamma_h\|_2\right)^2}{\|\gamma_{\mathcal{H}'}\|_2^2} \\ &\geq \frac{1}{\lambda^2_{\max}} - \frac{s}{(2c+1)\delta s} - \frac{2cs}{(2c+1)\delta s} \\ &\geq \frac{1}{\lambda^2_{\max}} - \frac{1}{\delta}, \end{aligned}$$

where we have used Cauchy-Schwarz's inequality in the third and fourth line and assumption (U) in the third line. Since $\delta > \lambda^2_{\max}$ by assumption, we obtain $\kappa(s, c) > 0$. □

Proof of Theorem 4.5. In Lemma 6.1 we can now set $f^0(\mathrm{X}) = \mathrm{X}\beta^0$ and $\beta = \beta^0$. Throughout the proof, all the inequalities are valid on the set $\mathcal{A}$. The first inequality in (37) implies that

$$\sum_h \lambda\lambda_h \|\widehat{\beta}_h - \beta_h^0\|_2 \leq 4 \sum_{h \in \mathcal{H}_0} \lambda\lambda_h \|\widehat{\beta}_h - \beta_h^0\|_2,$$

from which it follows

$$\sum_{h \in \mathcal{H}_0^c} \lambda_h \|\widehat{\beta}_h - \beta_h^0\|_2 \leq 3 \sum_{h \in \mathcal{H}_0} \lambda_h \|\widehat{\beta}_h - \beta_h^0\|_2. \tag{31}$$

Similarly, using the second inequality in (37),

$$\frac{1}{n}\|\mathrm{X}(\widehat{\beta} - \beta^0)\|_2^2 \leq 4\lambda\sqrt{|\mathcal{H}_0|}\sqrt{\sum_{h \in \mathcal{H}_0} \lambda_h^2 \|\widehat{\beta}_h - \beta_h^0\|_2^2}. \tag{32}$$

Combining (31) and (32), and using assumption (RE($|\mathcal{H}_0|, 3$)), we obtain

$$\sqrt{\sum_{h \in \mathcal{H}_0} \lambda_h^2 \|\widehat{\beta} - \beta_0\|_2^2} \leq \frac{4\lambda\sqrt{|\mathcal{H}_0|}}{\kappa_0^2}, \tag{33}$$

which yields (11).

Next, in virtue of (33), and using Cauchy-Schwarz's inequality

$$\left\|\Lambda(\widehat{\beta} - \beta^0)\right\|_2 \leq \sum_h \lambda_h \|\widehat{\beta} - \beta^0\|_2 \leq 4 \sum_{h \in \mathcal{H}_0} \lambda_h \|\widehat{\beta} - \beta^0\|_2 \leq 4\sqrt{|\mathcal{H}_0|}\sqrt{\sum_{h \in \mathcal{H}_0} \lambda_h^2 \|\widehat{\beta} - \beta_0\|_2^2},$$

which is bounded by $16\frac{\lambda}{\kappa_0^2}|\mathcal{H}_0|$. This implies

$$\|\widehat{\beta} - \beta^0\|_2 \leq 16\frac{\lambda}{\kappa_0^2 \lambda_{\min}}|\mathcal{H}_0|,$$

which is (10).

In order to show (12), we first show that

$$|\widehat{\mathcal{H}}| \leq \frac{16}{9}\frac{C_{\max}}{\lambda^2\lambda_{\min}^2}\frac{1}{n}\|\mathrm{X}(\beta_0 - \widehat{\beta})\|_2^2. \tag{34}$$

From the subgradient conditions, we get, for each h,

$$\frac{1}{n}\mathrm{X}_h^\top\left(\mathrm{X}(\beta^0 - \widehat{\beta})\right) + \frac{1}{n}\mathrm{X}_h^\top\epsilon = \lambda\lambda_h z_h,$$

where $z_h = \frac{\widehat{\beta}_h}{\|\widehat{\beta}_h\|_2}$ if $\widehat{\beta}_h \neq 0$ and z_h is any vector with ℓ_2 norm bounded by 1 if $\widehat{\beta}_h = 0$. Then, by the triangle inequality,

$$\frac{1}{n}\left\|\mathrm{X}_h^\top\left(\mathrm{X}(\widehat{\beta} - \beta^0)\right)\right\|_2 \geq \lambda\lambda_h - \frac{1}{n}\left\|\mathrm{X}_h^\top\epsilon\right\|_2 \geq \frac{1}{2}\lambda\lambda_h,$$

for each h. It then follows that

$$\frac{1}{n^2}\sum_{h\in\widehat{\mathcal{H}}}\left\|\mathrm{X}_h^\top\left(\mathrm{X}(\widehat{\beta}-\beta^0)\right)\right\|_2^2 \geq |\widehat{\mathcal{H}}|\lambda^2\lambda^2_{\min}\frac{1}{4}. \tag{35}$$

On the other hand, since

$$\mathrm{X}\mathrm{X}^\top=\sum_h \mathrm{X}_h\mathrm{X}_h^\top,$$

we also have

$$\frac{1}{n^2}\sum_{h\in\widehat{\mathcal{H}}}\left\|\mathrm{X}_h^\top\left(\mathrm{X}(\widehat{\beta}-\beta^0)\right)\right\|_2^2 \leq \frac{1}{n^2}\left(\mathrm{X}(\widehat{\beta}-\beta^0)\right)^\top \mathrm{X}\mathrm{X}^\top\left(\mathrm{X}(\widehat{\beta}-\beta^0)\right) \leq \frac{C_{\max}}{n}\|\mathrm{X}(\widehat{\beta}-\beta^0)\|_2^2, \tag{36}$$

where the last inequality follows from the fact that $\frac{1}{n}\mathrm{X}^\top\mathrm{X}$ and $\frac{1}{n}\mathrm{X}\mathrm{X}^\top$ have the same maximal eigenvalue. Combining (35) and (36),

$$|\widehat{\mathcal{H}}|\leq 4\frac{C_{\max}}{\lambda^2\lambda^2_{\min}}\frac{1}{n}\|\mathrm{X}(\beta^0-\widehat{\beta})\|_2^2,$$

which is (34). Inserting equation (11) in (34), we obtain (12). □

Proof of Theorem 4.7. Following the results of section A, part IV of Zhou et al. (2007), assumptions (P1) and (P2) coupled with Berstein's inequality yield

$$\max_{j,k}\left|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\right| = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

Then,

$$\begin{array}{rcl}
\sup_{\beta\in B_n}|R(\beta)-\widehat{R}(\beta)| & = & \sup_{\beta\in B_n}\left|\gamma^\top(\Sigma-\widehat{\Sigma})\gamma\right| \\
& \leq & \sup_{\beta\in B_n}\max_{j,k}\left|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\right|\|\gamma\|_1^2 \\
& \leq & \sup_{\beta\in B_n}\left|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\right|\left(1+\sum_h\sqrt{d_h}\|\beta_h\|_2\right)^2 \\
& \leq & \max_{j,k}\left|\Sigma_{j,k}-\widehat{\Sigma}_{j,k}\right|(1+b_n)^2 \\
& = & o_P(1).
\end{array}$$

where, in the second inequality, we used the bound $\|\gamma\|_1\leq 1+\sum_h\sqrt{d_h}\|\beta_h\|_2$ and the last step follows from (15). Therefore,

$$\sup_{\beta\in B_n}|R(\beta)-\widehat{R}(\beta)|\xrightarrow{p}0,$$

which implies persistence with respect to $\{B_n\}$, since

$$|R(\widehat{\beta}_n)-\inf_{\beta\in B_n}R(\beta)|\leq 2\sup_{\beta\in B_n}|R(\beta)-\widehat{R}(\beta)|.$$

The second part of the statement follows for the simple chain of inequalities

$$\begin{array}{rcl} \sum_h \sqrt{d_h}\|\beta_h\|_2 & = & \sum_h \sqrt{d_h}\|\beta_h\|_2 I_{\{\beta_h \neq 0\}} \\ & \leq & \|\beta\|_2 \sqrt{\sum_h d_h I_{\{\beta_h \neq 0\}}} \\ & \leq & C\sqrt{d_{\max}}\sqrt{|\{h, \beta_h \neq 0\}|}, \end{array}$$

where $\|\beta\|_2 \leq C$ holds uniformly over n for some constant C in virtue of (P1) and the assumed positivity of the minimal eigenvalue of the covariance matrix of the predictors. Under (16), this implies $C_n \subset B_n$ for each n and thus persistency with respect to $\{C_n\}_n$. □

6. Appendix

Proof of Lemma 4.3. Let $V_h = \frac{1}{\sqrt{n}\sigma} \mathrm{X}_h^\top \epsilon$, so that $V_h \sim N_{d_h}(0, \mathrm{I})$ and $\|V_h\|_2^2 \sim \chi^2_{d_h}$. By the union bound,

$$\begin{array}{rcl} \mathbb{P}(\mathcal{A}^c) \leq \sum_h \mathbb{P}\left(\|V_h\|_2^2 \geq \frac{1}{4}\frac{n}{\sigma^2}\lambda^2\lambda_h^2\right) & = & \sum_h \mathbb{P}\left(\|V_h\|_2^2 - d_h \geq \frac{1}{4}\frac{n}{\sigma^2}\lambda^2\lambda_h^2 - d_h\right) \\ & = & \sum_h \mathbb{P}\left(\|V_h\|_2^2 - d_h \geq \sqrt{2 d_h} x_h\right), \end{array}$$

where $x_h = \frac{1}{\sqrt{2}}\left(\frac{1}{4}\frac{\frac{n}{\sigma^2}\lambda^2\lambda_h^2}{\sqrt{d_h}} - \sqrt{d_h}\right)$. For large enough n, we can apply the tail bound inequality for a variable distributed like $\chi^2_{d_h}$ (see, e.g. Cavalier et al., 2002), yielding

$$\mathbb{P}(\mathcal{A}^c) \leq \sum_h \exp\left\{-\frac{x_h^2}{2\left(1 + x_h\sqrt{\frac{2}{d_h}}\right)}\right\}.$$

Because of (A), for large enough n,

$$\exp\left\{-\frac{x_h^2}{2\left(1 + x_h\sqrt{\frac{2}{d_h}}\right)}\right\} \leq \exp\left\{-\frac{x_h^2}{3\sqrt{2}\frac{x_h}{\sqrt{d_h}}}\right\} = \exp\left\{-\frac{1}{3\sqrt{2}}\left(\frac{n}{\sigma^2}\lambda^2\lambda_h^2 - d_h\right)\right\},$$

from which it follows, once again using (A), that

$$\mathbb{P}(\mathcal{A}^c) \leq \exp\left\{\log|\mathcal{H}| - \frac{1}{3\sqrt{2}}\min_h\left(\frac{n}{\sigma^2}\lambda^2\lambda_h^2 - d_h\right)\right\} \to 0.$$

This concludes the proof. □

Lemma 6.1. *Let* $\mathbb{E}Y = f^0(\mathrm{X})$*, for some function* f^0 *and assume (N). On the event* $\mathcal{A}$*, for any* $\beta \in \mathbb{R}^d$ *with block support set* $\mathcal{H}' = \{h \colon \beta_h \neq 0\}$*,*

$$\begin{array}{rcl} \frac{1}{n}\|\mathrm{X}\widehat{\beta} - f^0(\mathrm{X})\|_2^2 + \sum_h \lambda\lambda_h\|\widehat{\beta}_h - \beta_h\|_2 & \leq & \frac{1}{n}\|\mathrm{X}\beta - f^0(\mathrm{X})\|_2^2 + 4\lambda\sum_{h\in\mathcal{H}'}\lambda_h\|\widehat{\beta}_h - \beta_h\|_2 \\ & \leq & \frac{1}{n}\|\mathrm{X}\beta - f^0(\mathrm{X})\|_2^2 + 4\lambda\sqrt{|\mathcal{H}'|}\sqrt{\sum_{h\in\mathcal{H}'}\lambda_h^2\|\widehat{\beta}_h - \beta_h\|_2^2}, \end{array} \tag{37}$$

Proof of Lemma 6.1. Following the derivation in Bunea et al. (2007a), for an arbitrary $\beta \in \mathbb{R}^d$ with block support set $\mathcal{H}'$, it holds that

$$\frac{1}{2n}\|\mathrm{X}\widehat{\beta}-f^0(\mathrm{X})\|_2^2 \le \frac{1}{2n}\|\mathrm{X}\beta-f^0(\mathrm{X})\|_2^2+\sum_h \lambda\lambda_h\|\beta_h\|_2-\sum_h \lambda\lambda_h\|\widehat{\beta}_h\|_2+\sum_h W_h^\top(\widehat{\beta}_h-\beta_h), \tag{38}$$

where $W_h = \frac{1}{n}\mathrm{X}_h^\top \epsilon$. By Cauchy-Schwarz's inequality, on the event $\mathcal{A}$,

$$\sum_h |W_h^\top(\widehat{\beta}_h-\beta_h)| \le \frac{1}{2}\sum_h \lambda\lambda_h\|\widehat{\beta}_h-\beta_h\|_2. \tag{39}$$

Using the last display, and adding and subtracting $\frac{1}{2}\sum_h \lambda\lambda_h\|\widehat{\beta}_h-\beta_h\|_2$ to both sides of (38), the term

$$\frac{1}{2n}\|\mathrm{X}\widehat{\beta}-f^0(\mathrm{X}))\|_2^2+\frac{1}{2}\sum_h \lambda\lambda_h\|\widehat{\beta}_h-\beta_h\|_2$$

is bounded by

$$\frac{1}{2n}\|\mathrm{X}\beta-f^0(\mathrm{X})\|_2^2+\sum_h \lambda\lambda_h\|\widehat{\beta}_h-\beta_h\|_2+\sum_h \lambda\lambda_h\|\beta_h\|_2-\sum_h \lambda\lambda_h\|\widehat{\beta}_h\|_2,$$

which, in turn, is no larger than

$$\frac{1}{2n}\|\mathrm{X}\beta-f^0(\mathrm{X})\|_2^2+\sum_{h\in\mathcal{H}'} \lambda\lambda_h\|\widehat{\beta}_h-\beta_h\|_2+\sum_{h\in\mathcal{H}'} \lambda\lambda_h\left(\|\beta_h\|_2-\|\widehat{\beta}_h\|_2\right),$$

all the above inequalities being valid on $\mathcal{A}$. Then, from (38), and applying the triangle inequality to the last display, we obtain, still on $\mathcal{A}$,

$$\begin{aligned}\tfrac{1}{2n}\|\mathrm{X}\widehat{\beta}-f^0(\mathrm{X})\|_2^2+\tfrac{1}{2}\textstyle\sum_h \lambda\lambda_h\|\widehat{\beta}_h-\beta_h\|_2 &\le \tfrac{1}{2n}\|\mathrm{X}\beta-f^0(\mathrm{X})\|_2^2+2\lambda\textstyle\sum_{h\in\mathcal{H}'}\lambda_h\|\widehat{\beta}_h-\beta_h\|_2\\ &\le \tfrac{1}{2n}\|\mathrm{X}\beta-f^0(\mathrm{X})\|_2^2+2\lambda\sqrt{|\mathcal{H}'|}\sqrt{\textstyle\sum_{h\in\mathcal{H}'}\lambda_h^2\|\widehat{\beta}_h-\beta_h\|_2^2},\end{aligned}$$

where the second inequality stems from Cauchy-Schwarz's inequality. The last expression, multiplied by 2, is (37). □

7. Acknowledgments

We thank the anonymous referees and, in particular, the associate editor for detailed and constructivce comments that led to a much improved presentation.

References

Bach, F. (2007). Consistency of the group Lasso and multiple kernel learning, to appear in *Journal of Machine Learning.*

Bickel, P.J., Ritov, Y. and Tsybakov, A. B. (2007). A Simultaneous analysis of Lasso and Dantzig selector. Submitted to *Annals of Statistics.*

Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Aggregation for Gaussian regression, *The Annals of Statistics,* 35(4), 1674–1697.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2007b). Sparsity oracle inequalities for the lasso, *Electronic Journal of Statistics,* 1, 169194.

Cavalier, L. , Golubev, G. K. , Picard, D. and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems, *Annals of Statistics,* 30 843–874.

Dahinden, C., Parmiggiani, G., Emerick, M.C. and Bühlmann, P. (2006). Sparse Contingency Tables and High-Dimensional Log-Linealr Models for Alternative Splicing in Full-Length cDNA Libraries, Research Report 132, Swiss Federal Institute of Technology.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression, *Annals of Statistics,* 32, 407499.

Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle properties, Journal of the American Statistical Association, 96(456), 1348–1360.

Fan J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *Annals of Statistics,* 32(3), 928–961.

Geyer, C. (1994). On the asymptotics of constrained-M estimation, *Annals of Statistics,*, 22, 1993 – 2010.

Gilbert, A. C. and Strauss, M. J. (2006). Algorithms for Simultaneous Sparse Approximation Part II: Convex Relaxation, *Signal Processing,* 86, 572588

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional predictor selection and the virtue of overparametrization, *Bernoulli,* 10, 971–988.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1 constraint, *Annals of Statistics,* 34(5), 2367–2386.

Kim, Y., Kim, J. and Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica,* 16(2).

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators, *Annals of Statistics,* 28(5), 1356–1378.

Koltchinskii, V. (2005). Sparsity in Penalized Empirical Risk Minimization, manuscript.

Ledoux, M. and Talagrand, M. (1991). Probability in Banach spaces: isoperimetry and processes. Springer-Verlag.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators, *Electronic Journal of Statistics,* 2, 90–102.

Massart, P. (2007). Concentration Inequalities and Model Selection, Lecture Notes in Mathematics, Vol. 1896, Springer.

Meier, L., van der Geer, S. and Bühlmann, P. (2006). The Group Lasso for Logistic Regression, *Journal of the Royal Statistical Society,* Series B, 70(1), 53-7.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso, *Annals of Statistics,* 34(3), 1436–1462.

Meinshausen, N. and Yu, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data, to appear in the *Annals of Statistics.*

Nardi, N. and Rinaldo, A. (2007). The Log-linear Group-Lasso Estimator and Its Asymptotic Properties, manuscript.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000). On the LASSO and its Dual, *Journal of Computational and Graphical Statistics,* 9(2), 319–337.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, *Annals of Statistics,* 16(1), 356–366.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2007). Sparse Additive Models, manuscript.

Rinaldo (2006). Computing Maximum Likelihood Estimates in Log-Linear Models, Technical report, Department of Statistics, Carnegie Mellon University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society,* Series B, 58(1), 267–288.

van de Geer, S.A. (2007). Oracle inequalities and Regularization, in Lectures on Empirical Processes, European Mathematical Society.

van der Vaart, A.W. and Wellner, J.A. (1998). *Weak Convergence and Empirical Processes,* Springer.

Wainwright, M. J. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity, Technical Report 708, Department of Statistics, UC Berkeley.

Wainwright, M. J. (2007). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting, Technical Report, UC Berkeley, Department of Statistics .

Yuan, M. and Lin Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society,* B, 68(1), 49–67.

Yuan, M. and Lin Y. (2006). On the non-negative garrotte estimator, *Journal of the Royal Statistical Society,* B, 69(2), 143–161.

Zhang, T. (2007). Some Sharp Performance Bounds for Least Squares Regression with L_1 Regularization, manuscript.

Zhang, H. and Huang, J. (2007). The sparsity and bias of the Lasso selection in high-dimensional linear regression, to appear in the *Annals of Statistics.*

Zhao, P., Rocha, G. and Yu, B. (2008). Grouped and hierarchical model selection through composite absolute penalties, manuscript.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso, *Journal of Machine Learning Research,* 7, 2541–2563.

Zhou, S., Lafferty, J. and Wasserman, L. (2007). Compressed Regression, manuscript.

Zhou, N. and Zhu, J. (2007). Group Variable Selection via Hierarchical Lasso and Its Oracle Property, manuscript.

Zou, H. (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association,* 101(476), 1418–1429.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society,* Series B, 67(2):301–320, 2005.