Taking Advantage of Sparsity in Multi-Task Learning

Karim Lounici
LPMA and CREST
3, Av. Pierre Larousse,
92240 Malakoff, France
karim.lounici@ensae.fr

Massimiliano Pontil
Department of Computer Science
University College London
Malet Pl., WC1E, London, UK
m.pontil@cs.ucl.ac.uk

Alexandre B. Tsybakov
LPMA and CREST
3, Av. Pierre Larousse,
92240 Malakoff, France
alexandre.tsybakov@ensae.fr

Sara A. van de Geer
Seminar für Statistik, ETH Zürich
LEO D11, 8092 Zürich
Switzerland
geer@stat.math.ethz.ch

Abstract

We study the problem of estimating multiple linear regression equations for the purpose of both prediction and variable selection. Following recent work on multi-task learning [1], we assume that the sparsity patterns of the regression vectors are included in the same set of small cardinality. This assumption leads us to consider the Group Lasso as a candidate estimation method. We show that this estimator enjoys nice sparsity oracle inequalities and variable selection properties. The results hold under a certain restricted eigenvalue condition and a coherence condition on the design matrix, which naturally extend recent work in [3, 19]. In particular, in the multi-task learning scenario, in which the number of tasks can grow, we are able to remove completely the effect of the number of predictor variables in the bounds. Finally, we show how our results can be extended to more general noise distributions, of which we only require the variance to be finite.[1]

1 Introduction

We study the problem of estimating multiple regression equations under sparsity assumptions on the underlying regression coefficients. More precisely, we consider multiple Gaussian regression models,

$$\begin{aligned} y_1 &= X_1\beta_1^* + W_1 \\ y_2 &= X_2\beta_2^* + W_2 \\ &\vdots \\ y_T &= X_T\beta_T^* + W_T \end{aligned} \tag{1.1}$$

where, for each $t = 1, \ldots, T$, we let X_t be a prescribed $n \times M$ design matrix, β_t^* the unknown vector of regression coefficients and y_t an n-dimensional vector of observations. We assume that $W_1, \ldots, W_T$ are *i.i.d.* zero mean random vectors.

We are interested in estimation methods which work well even when the number of parameters in each equation is much larger than the number of observations, that is, $M \gg n$. This situation may arise in many practical applications in which the predictor variables are inherently high dimensional, or it may be "costly" to observe response variables, due to difficult experimental procedures, see, for example [1] for a discussion.

Examples in which this estimation problem is relevant range from multi-task learning [1, 8, 20, 24] and conjoint analysis (see, for example, [13, 18] and references therein) to longitudinal data analysis [10] as well as the analysis of panel data [14, 28], among others. In particular, multi-task learning provides a main motivation for our study. In that setting each regression equation corresponds to a different learning task (the classification case can be treated similarly); in addition to the requirement that $M \gg n$, we also allow for the number of tasks T to be much larger than n. Following [1] we assume that there are only few common important variables which are shared by the tasks. A general goal of this paper is to study the implications of this assumption from a statistical learning view point, in particular, to quantify the advantage provided by the large number of tasks to learn the underlying vectors $\beta_1^*, \ldots, \beta_T^*$ as well as to select common variables shared by the tasks.

Our study pertains and draws substantial ideas from the recently developed area of compressed sensing and sparse estimation (or sparse recovery), see [3, 7, 11] and references therein. A central problem studied therein is that of estimating the parameters of a (single) Gaussian regression model. Here, the term "sparse" means that most of the components of the underlying M-dimensional regression vector are equal to zero. A main motivation for sparse estimation comes from the observation that in many practical applications M is much larger than the number n of observations but the underlying model is sparse, see [7, 11] and references therein. Un-

[1] The first author should be considered for Mark Fulk award.

der this circumstance ordinary least squares will not work. A more appropriate method for sparse estimation is the ℓ_1-norm penalized least squares method, which is commonly referred to as the Lasso method. In fact, it has been recently shown by different authors, under different conditions on the design matrix, that the Lasso satisfies sparsity oracle inequalities, see [3, 5, 6, 26] and references therein. Closest to our study in this paper is [3], which relies upon a Restricted Eigenvalue (RE) assumption. The results of these works make it possible to estimate the parameter β even in the so-called "*p much larger than n*" regime (in our notation, the number of predictor variables p corresponds to MT).

In this paper, we assume that the vectors $\beta_1^*, \ldots, \beta_T^*$ are not only sparse but also have their sparsity patterns included in the same set of small cardinality s. In other words, the response variable associated with each equation in (1.1) depends only on some members of a small subset of the corresponding predictor variables, which is preserved across the different equations. This assumption, that we further refer to as *structured sparsity assumption*, is motivated by some recent work on multi-task learning [1]. It naturally leads to an extension of the Lasso method, the so-called Group Lasso [29], in which the error term is the average residual error across the different equations and the penalty term is a mixed $(2,1)$-norm. The structured sparsity assumption induces a relation between the responses and, as we shall see, can be used to improve estimation.

The paper is organized as follows. In Section 2 we define the estimation method and comment on previous related work. In Section 3 we study the oracle properties of this estimator when the errors W_t are Gaussian. Our main results concern upper bounds on the prediction error and the distance between the estimator and the true regression vector β^*. Specifically, Theorem 3.3 establishes that under the above structured sparsity assumption on β^*, the prediction error is essentially of the order of s/n. In particular, in the multi-task learning scenario, in which T can grow, we are able to remove completely the effect of the number of predictor variables in the bounds. Next, in Section 4, under a stronger condition on the design matrices, we describe a simple modification of our method and show that it selects the correct sparsity pattern with an overwhelming probability (Theorem 4.3). We also find the rates of convergence of the estimators for mixed $(2,1)$-norms with $1 \le p \le \infty$ (Corollary 4.4). The techniques of proofs build upon and extend those of [3] and [19]. Finally, in Section 5 we discuss how our results can be extended to more general noise distributions, of which we only require the variance to be finite.

2 Method and related work

In this section we first introduce some notation and then describe the estimation method which we analyze in the paper. As stated above, our goal is to estimate T linear regression functions identified by the parameters $\beta_1^*, \ldots, \beta_T^* \in \mathbb{R}^M$. We may write the model (1.1) in compact notation as

$$y = X\beta^* + W \tag{2.1}$$

where y and W are the nT-dimensional random vectors formed by stacking the vectors $y_1, \ldots, y_T$ and the vectors $W_1, \ldots, W_T$, respectively. Likewise β^* denotes the vector obtained by stacking the regression vectors $\beta_1^*, \ldots, \beta_T^*$. Unless otherwise specified, all vectors are meant to be column vectors. Thus, for every $t \in \mathbb{N}_T$, we write $y_t = (y_{ti} : i \in \mathbb{N}_n)^\top$ and $W_t = (W_{ti} : i \in \mathbb{N}_n)^\top$, where, hereafter, for every positive integer k, we let $\mathbb{N}_k$ be the set of integers from 1 and up to k. The $nT \times MT$ block diagonal design matrix X has its t-th block formed by the $n \times M$ matrix X_t. We let $x_{t1}^\top, \ldots, x_{tn}^\top$ be the row vectors forming X_t and $(x_{ti})_j$ the j-th component of the vector x_{ti}. Throughout the paper we assume that x_{ti} are deterministic.

For every $\beta \in \mathbb{R}^{MT}$ we introduce $(\beta)^j \equiv \beta^j = (\beta_{tj} : t \in \mathbb{N}_T)^\top$, that is, the vector formed by the coefficients corresponding to the j-th variable. For every $1 \le p < \infty$ we define the mixed $(2,p)$-norm of β as

$$\|\beta\|_{2,p} = \left(\sum_{j=1}^M \left(\sum_{t=1}^T \beta_{tj}^2 \right)^{\frac{p}{2}} \right)^{\frac{1}{p}} = \left(\sum_{j=1}^M \|\beta^j\|^p \right)^{\frac{1}{p}}$$

and the $(2,\infty)$-norm of β as

$$\|\beta\|_{2,\infty} = \max_{1 \le j \le M} \|\beta^j\|,$$

where $\|\cdot\|$ is the standard Euclidean norm.

If $J \subseteq \mathbb{N}_M$ we let $\beta_J \in \mathbb{R}^{MT}$ be the vector formed by stacking the vectors $(\beta^j I\{j \in J\} : j \in \mathbb{N}_M)$, where $I\{\cdot\}$ denotes the indicator function. Finally we set $J(\beta) = \{j : \beta^j \ne 0,\ j \in \mathbb{N}_M\}$ and $M(\beta) = |J(\beta)|$ where $|J|$ denotes the cardinality of set $J \subset \{1, \ldots, M\}$. The set $J(\beta)$ contains the indices of the relevant variables shared by the vectors $\beta_1, \ldots, \beta_T$ and the number $M(\beta)$ quantifies the level of structured sparsity across those vectors.

We have now accumulated the sufficient information to introduce the estimation method. We define the empirical residual error

$$\hat{S}(\beta) = \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{ti}^\top \beta_t - y_{ti})^2 = \frac{1}{nT} \|X\beta - y\|^2$$

and, for every $\lambda > 0$, we let our estimator $\hat{\beta}$ be a solution of the optimization problem [1]

$$\min_\beta \hat{S}(\beta) + 2\lambda \|\beta\|_{2,1}. \tag{2.2}$$

In order to study the statistical properties of this estimator, it is useful to derive the optimality condition for a solution of the problem (2.2). Since the objective function in (2.2) is convex, $\hat{\beta}$ is a solution of (2.2) if and only if 0 (the MT-dimensional zero vector) belongs to the subdifferential of the objective function. In turn, this condition is equivalent to the requirement that

$$-\nabla \hat{S}(\hat{\beta}) \in 2\lambda \partial \left(\sum_{j=1}^M \|\hat{\beta}^j\| \right),$$

where ∂ denotes the subdifferential (see, for example, [4] for more information on convex analysis). Note that

$$\partial \left(\sum_{j=1}^M \|\beta^j\| \right) = \Bigg\{ \theta \in \mathbb{R}^{MT} : \theta^j = \frac{\beta^j}{\|\beta^j\|} \text{ if } \beta^j \ne 0, \quad \|\theta^j\| \le 1, \text{ if } \beta^j = 0,\ j \in \mathbb{N}_M \Bigg\}.$$

Thus, $\hat\beta$ is a solution of (2.2) if and only if

$$\frac{1}{nT}(X^\top(y - X\hat\beta))^j = \lambda\frac{\hat\beta^j}{\|\hat\beta^j\|}, \quad \text{if } \hat\beta^j \neq 0 \qquad (2.3)$$

$$\frac{1}{nT}\|(X^\top(y - X\hat\beta))^j\| \leq \lambda, \qquad \text{if } \hat\beta^j = 0. \qquad (2.4)$$

We now comment on previous related work. Our estimator is a special case of the Group Lasso estimator [29]. Several papers analyzing statistical properties of the Group Lasso appeared quite recently [2, 9, 15, 17, 21, 22, 23, 25]. Most of them are focused on the Group Lasso for additive models [15, 17, 22, 25] or generalized linear models [21]. Special choice of groups is studied in [9]. Discussion of the Group Lasso in a relatively general setting is given by Bach [2] and Nardi and Rinaldo [23]. Bach [2] assumes that the predictors x_{ti} are random with a positive definite covariance matrix and proves results on consistent selection of sparsity pattern $J(\beta^*)$ when the dimension of the model ($p = MT$ in our case) is fixed and $n \to \infty$. Nardi and Rinaldo [23] consider a setting that covers ours and address the issue of sparsity oracle inequalities in the spirit of [3]. However, their bounds are too coarse (see comments in Section 3 below). Obozinski et al. [25] consider the case where all the matrices X_i are the same and all their rows are independent Gaussian random vectors with the same covariance matrix. They show that the resulting estimator achieves consistent selection of the sparsity pattern and that there may be some improvement with respect to the usual Lasso. Except for this very particular example, theoretical advantages of the group Lasso as compared to the usual Lasso were not featured in the literature. Note also that Obozinski et al. [25] focused on the consistent selection, whereas it remained unclear whether there is some improvement in the prediction properties as compared to the usual Lasso.

One of the aims of this paper is to show that such an improvement does exist. In particular, our Theorem 3.3 implies that the prediction bound for the Group Lasso estimator that we use here is better than for the standard Lasso under the same assumptions. Furthermore, we demonstrate that as the number of tasks T increases the dependence of the bound on M disappears, provided that M grows at the rate slower than $\exp(\sqrt{T})$.

Finally, we note that we recently came across the work [16], whose results are similar to those in Section 3 below.

3 Sparsity oracle inequality

Let $1 \leq s \leq M$ be an integer that gives an upper bound on the structured sparsity $M(\beta^*)$ of the true regression vector β^*. We make the following assumption.

Assumption 3.1 *There exists a positive number $\kappa = \kappa(s)$ such that*

$$\min\left\{\frac{\|X\Delta\|}{\sqrt{n}\|\Delta_J\|} \ : \ |J| \leq s, \Delta \in \mathbb{R}^{MT} \setminus \{0\}, \|\Delta_{J^c}\|_{2,1} \leq 3\|\Delta_J\|_{2,1}\right\} \geq \kappa,$$

where J^c denotes the complement of the set of indices J.

To emphasize the dependency of Assumption 3.1 on s, we will sometimes refer to it as Assumption RE(s). This is a natural extension to our setting of the Restricted Eigenvalue assumption for the usual Lasso and Dantzig selector from [3]. The ℓ_1 norms are now replaced by the mixed (2,1)-norms. Note that, however, the analogy is not complete. In fact, the sample size n in the usual Lasso setting corresponds to nT in our case, whereas in Assumption 3.1 we consider $\sqrt{\Delta^\top X^\top X\Delta/n}$ and not $\sqrt{\Delta^\top X^\top X\Delta/(nT)}$. This is done in order to have a correct normalization of κ without compulsory dependence on T (if we use the term $\sqrt{\Delta^\top X^\top X\Delta/(nT)}$ in Assumption 3.1, then $\kappa \sim T^{-1/2}$ even in the case of the identity matrix $X^\top X/n$).

Several simple sufficient conditions for Assumption 3.1 with $T = 1$ are given in [3]. Similar sufficient conditions can be stated in our more general setting. For example, it is enough to suppose that each of the matrices $X_t^\top X_t/n$ is positive definite or satisfies a Restricted Isometry condition as in [7] or the coherence condition (cf. Lemma 4.2 below).

Lemma 3.2 *Consider the model* (1.1) *for $M \geq 2$ and $T, n \geq 1$. Assume that the random vectors $W_1, \dots, W_T$ are i.i.d. Gaussian with zero mean and covariance matrix $\sigma^2 I_{n\times n}$, all diagonal elements of the matrix $X^\top X/n$ are equal to* 1 *and $M(\beta^*) \leq s$. Let*

$$\lambda = \frac{2\sigma}{\sqrt{nT}}\left(1 + \frac{A\log M}{\sqrt{T}}\right)^{1/2},$$

where $A > 8$ and let $q = \min(8\log M, A\sqrt{T}/8)$. Then with probability at least $1 - M^{1-q}$, for any solution $\hat\beta$ of problem (2.2) *and all $\beta \in \mathbb{R}^{MT}$ we have*

$$\frac{1}{nT}\|X(\hat\beta - \beta^*)\|^2 + \lambda\|\hat\beta - \beta\|_{2,1} \leq \qquad (3.1)$$
$$\leq \frac{1}{nT}\|X(\beta - \beta^*)\|^2 + 4\lambda\sum_{j\in J(\beta)}\|\hat\beta^j - \beta^j\|,$$

$$\frac{1}{nT}\max_{1\leq j\leq M}\|(X^\top X(\beta^* - \hat\beta))^j\| \leq \frac{3}{2}\lambda, \qquad (3.2)$$

$$M(\hat\beta) \leq \frac{4\phi_{\max}}{\lambda^2 nT^2}\|X(\hat\beta - \beta^*)\|^2, \qquad (3.3)$$

where $\phi_{\max}$ is the maximum eigenvalue of the matrix $X^\top X/n$.

Proof: For all $\beta \in \mathbb{R}^{MT}$, we have

$$\frac{1}{nT}\|X\hat\beta - y\|^2 + 2\lambda\sum_{j=1}^M\|\hat\beta^j\| \leq \frac{1}{nT}\|X\beta - y\|^2 + 2\lambda\sum_{j=1}^M\|\beta^j\|$$

which, using $y = X\beta^* + W$, is equivalent to

$$\frac{1}{nT}\|X(\hat\beta - \beta^*)\|^2 \leq \frac{1}{nT}\|X(\beta - \beta^*)\|^2$$
$$+ \frac{2}{nT}W^\top X(\hat\beta - \beta) + 2\lambda\sum_{j=1}^M\left(\|\beta^j\| - \|\hat\beta^j\|\right). \qquad (3.4)$$

By Hölder's inequality, we have that

$$W^\top X(\hat\beta - \beta) \leq \|X^\top W\|_{2,\infty}\|\hat\beta - \beta\|_{2,1}$$

where

$$\|X^\top W\|_{2,\infty} = \max_{1\le j\le M} \sqrt{\sum_{t=1}^{T}\left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2}.$$

Consider the random event

$$\mathcal{A} = \left\{\frac{1}{nT}\|X^\top W\|_{2,\infty} \le \frac{\lambda}{2}\right\}.$$

Since we assume all diagonal elements of the matrix $X^\top X/n$ to be equal to 1, the random variables

$$V_{tj} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(x_{ti})_j W_{ti},$$

$t = 1,\dots,T$, are *i.i.d.* standard Gaussian. Using this fact we can write, for any $j = 1,\dots,M$,

$$\begin{aligned}
&\Pr\left(\sum_{t=1}^{T}\left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2 \ge \frac{\lambda^2(nT)^2}{4}\right)\\
&= \Pr\left(\chi_T^2 \ge \frac{\lambda^2 nT^2}{4\sigma^2}\right)\\
&= \Pr\left(\chi_T^2 \ge T + A\sqrt{T}\log M\right),
\end{aligned}$$

where χ_T^2 is a chi-square random variable with T degrees of freedom. We now apply Lemma A.1, the union bound and the fact that $A > 8$ to get

$$\Pr(\mathcal{A}^c) \le M\exp\left(-\frac{A\log M}{8}\min\left(\sqrt{T}, A\log M\right)\right) \le M^{1-q}.$$

It follows from (3.4) that, on the event $\mathcal{A}$.

$$\begin{aligned}
&\frac{1}{nT}\|X(\hat\beta-\beta^*)\|^2 + \lambda\sum_{j=1}^{M}\|\hat\beta^j-\beta^j\| \le\\
&\frac{1}{nT}\|X(\beta-\beta^*)\|^2 + 2\lambda\sum_{j=1}^{M}\left(\|\hat\beta^j-\beta^j\| + \|\beta^j\| - \|\hat\beta^j\|\right)\\
&\le \frac{1}{nT}\|X(\beta-\beta^*)\|^2 + 4\lambda\sum_{j\in J(\beta)}\|\hat\beta^j-\beta^j\|,
\end{aligned}$$

which coincides with (3.1). To prove (3.2), we use the inequality

$$\frac{1}{nT}\max_{1\le j\le M}\|(X^\top(y-X\hat\beta))^j\| \le \lambda, \tag{3.5}$$

which follows from (2.3) and (2.4). Then,

$$\begin{aligned}
&\frac{1}{nT}\|(X^\top(X(\hat\beta-\beta^*)))^j\| \le\\
&\frac{1}{nT}\|(X^\top(X\hat\beta-y))^j\| + \frac{1}{nT}\|(X^\top W)^j\|,
\end{aligned}$$

where we have used $y = X\beta^* + W$ and the triangle inequality. The result then follows by combining the last inequality with inequality (3.5) and using the definition of the event $\mathcal{A}$.

Finally, we prove (3.3). First, observe that, on the event $\mathcal{A}$,

$$\frac{1}{nT}\|(X^\top X(\hat\beta-\beta^*))^j\| \ge \frac{\lambda}{2}, \quad \text{if } \hat\beta^j \ne 0.$$

This fact follows from (2.3), (2.1) and the definition of the event $\mathcal{A}$. The following chain yields the result:

$$\begin{aligned}
M(\hat\beta) &\le \frac{4}{\lambda^2(nT)^2}\sum_{j\in J(\hat\beta)}\|(X^\top X(\hat\beta-\beta^*))^j\|^2\\
&\le \frac{4}{\lambda^2(nT)^2}\sum_{j=1}^{M}\|(X^\top X(\hat\beta-\beta^*))^j\|^2\\
&= \frac{4}{\lambda^2(nT)^2}\|X^\top X(\hat\beta-\beta^*)\|^2\\
&\le \frac{4\phi_{\max}}{\lambda^2 nT^2}\|X(\hat\beta-\beta^*)\|^2,
\end{aligned}$$

where, in the last line we have used the fact that the eigenvalues of $X^\top X/n$ are bounded from above by $\Phi_{\max}$. ■

We are now ready to state the main result of this section.

Theorem 3.3 *Consider the model* (1.1) *for* $M \ge 2$ *and* $T, n \ge 1$. *Assume that the random vectors* $W_1,\dots,W_T$ *are i.i.d. Gaussian with zero mean and covariance matrix* $\sigma^2 I_{n\times n}$, *all diagonal elements of the matrix* $X^\top X/n$ *are equal to* 1 *and* $M(\beta^*) \le s$. *Furthermore let Assumption 3.1 hold with* $\kappa = \kappa(s)$ *and let* $\phi_{\max}$ *be the largest eigenvalue of the matrix* $X^\top X/n$. *Let*

$$\lambda = \frac{2\sigma}{\sqrt{nT}}\left(1 + \frac{A\log M}{\sqrt{T}}\right)^{1/2},$$

where $A > 8$ *and let* $q = \min(8\log M, A\sqrt{T}/8)$. *Then with probability at least* $1 - M^{1-q}$, *for any solution* $\hat\beta$ *of problem* (2.2) *we have*

$$\frac{1}{nT}\|X(\hat\beta-\beta^*)\|^2 \le \frac{64\sigma^2}{\kappa^2}\frac{s}{n}\left(1+\frac{A\log M}{\sqrt{T}}\right) \tag{3.6}$$

$$\frac{1}{\sqrt{T}}\|\hat\beta-\beta^*\|_{2,1} \le \frac{32\sigma}{\kappa^2}\frac{s}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}} \tag{3.7}$$

$$M(\hat\beta) \le \frac{64\phi_{\max}}{\kappa^2}s. \tag{3.8}$$

If, in addition, Assumption RE(2s) holds, then with the same probability for any solution $\hat\beta$ *of problem* (2.2) *we have*

$$\frac{1}{\sqrt{T}}\|\hat\beta-\beta^*\| \le \frac{8\sqrt{10}\sigma}{\kappa^2(2s)}\sqrt{\frac{s}{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}. \tag{3.9}$$

Proof: We act similarly to the proof of Theorem 6.2 in [3]. Let $J = J(\beta^*) = \{j : (\beta^*)^j \ne 0\}$. By inequality (3.1) with $\beta = \beta^*$ we have, on the event $\mathcal{A}$, that

$$\begin{aligned}
\frac{1}{nT}\|X(\hat\beta-\beta^*)\|^2 &\le 4\lambda\sum_{j\in J}\|\hat\beta^j-\beta^{*j}\|\\
&\le 4\lambda\sqrt{s}\|(\hat\beta-\beta^*)_J\|.
\end{aligned} \tag{3.10}$$

Moreover by the same inequality, on the event $\mathcal{A}$, we have $\sum_{j=1}^{M}\|\hat\beta^j - \beta^{*j}\| \le 4\sum_{j\in J}\|\hat\beta^j - \beta^{*j}\|$, which implies that $\sum_{j\in J^c}\|\hat\beta^j-\beta^{*j}\|\le 3\sum_{j\in J}\|\hat\beta^j-\beta^{*j}\|$. Thus, by Assumption 3.1

$$\|(\hat\beta-\beta^*)_J\|\le\frac{\|X(\hat\beta-\beta^*)\|}{\kappa\sqrt{n}}. \tag{3.11}$$

Now, (3.6) follows from (3.10) and (3.11). Inequality (3.7) follows again by noting that

$$\sum_{j=1}^M\|\hat\beta^j-\beta^{*j}\|\le 4\sum_{j\in J}\|\hat\beta^j-\beta^{*j}\|\le 4\sqrt{s}\|(\hat\beta-\beta^*)_J\|$$

and then using (3.6). Inequality (3.8) follows from (3.3) and (3.6).

Finally, we prove (3.9). Let $\Delta=\hat\beta-\beta^*$ and let J' be the set of indices in J^c corresponding to s maximal in absolute value norms $\|\Delta^j\|$. Consider the set $J_{2s}=J\cup J'$. Note that $|J_{2s}|\leqslant 2s$. Let $\|\Delta^{(k)}_{J^c}\|$ denote the k-th largest norm in the set $\{\|\Delta^j\|:\,j\in J^c\}$. Then, clearly,

$$\|\Delta^{(k)}_{J^c}\|\le\sum_{j\in J^c}\|\Delta^j\|/k=\|\Delta_{J^c}\|_{2,1}/k.$$

This and the fact that $\|\Delta_{J^c}\|_{2,1}\le 3\|\Delta_J\|_{2,1}$ on the event $\mathcal{A}$ implies

$$\begin{aligned}\sum_{j\in J^c_{2s}}\|\Delta^j\|^2 &\le \sum_{k=s+1}^{\infty}\frac{\|\Delta_{J^c}\|^2_{2,1}}{k^2}\\ &\le \frac{\|\Delta_{J^c}\|^2_{2,1}}{s}\le\frac{9\|\Delta_J\|^2_{2,1}}{s}\\ &\le 9\sum_{j\in J}\|\Delta^j\|^2\le 9\sum_{j\in J_{2s}}\|\Delta^j\|^2.\end{aligned}$$

Therefore, on $\mathcal{A}$ we have

$$\|\Delta\|^2\le 10\sum_{j\in J_{2s}}\|\Delta^j\|^2\equiv 10\|\Delta_{J_{2s}}\|^2 \tag{3.12}$$

and also from (3.10):

$$\frac{1}{nT}\|X\Delta\|^2\le 4\lambda\sqrt{s}\|\Delta_{J_{2s}}\|. \tag{3.13}$$

In addition, $\|\Delta_{J^c}\|_{2,1}\le 3\|\Delta_J\|_{2,1}$ easily implies that

$$\|\Delta_{J^c_{2s}}\|_{2,1}\le 3\|\Delta_{J_{2s}}\|_{2,1}.$$

Combining (3.13) with Assumption RE($2s$) we find that on the event $\mathcal{A}$ it holds that

$$\|\Delta_{J_{2s}}\|\le\frac{4\lambda\sqrt{s}\,T}{\kappa^2(2s)}.$$

This inequality and (3.12) yield (3.9). ■

Theorem 3.3 is valid for any fixed n,M,T; the approach is non-asymptotic. Some relations between these parameters are relevant in the particular applications and various asymptotics can be derived as corollaries. For example, in multi-task learning it is natural to assume that $T\ge n$, and the motivation for our approach is the strongest if also $M\gg n$. The bounds of Theorem 3.3 are meaningful if the sparsity index s is small as compared to the sample size n and the logarithm of the dimension $\log M$ is not too large as compared to $\sqrt{T}$.

Note also that the values T and $\sqrt{T}$ in the denominators of the right-hand sides of (3.6), (3.7), and (3.9) appear quite naturally. For instance, the norm $\|\hat\beta-\beta^*\|_{2,1}$ in (3.7) is a sum of M terms each of which is a Euclidean norm of a vector in $\mathbb{R}^T$, and thus it is of the order $\sqrt{T}$ if all the components are equal. Therefore, (3.7) can be interpreted as a correctly normalized "error per coefficient" bound.

We now state several important conclusions. They are all valid for the general Group Lasso, and not only in the multi-task learning setup. They key point for their validity is the structured sparsity assumption.

1. *Theorem 3.3 applies to the general Group Lasso setting.* Indeed, the proofs in this section do not use the fact that the matrix $X^\top X$ is block-diagonal. The only restriction on $X^\top X$ is given in Assumption 3.1. For example, Assumption 3.1 is obviously satisfied if $X^\top X/(nT)$ (the correctly normalized Gram matrix of the regression model (2.1)) has a positive minimal eigenvalue.

2. *The dependence on the dimension M is negligible for large T.* Indeed, the bounds of Theorem 3.3 become independent of M if we choose the number of tasks T larger than $\log^2 M$. A striking fact is that no relation between the sample size n and the dimension M is required. This is quite in contrast to the previous results on sparse recovery where the assumption $\log M=o(n)$ was considered as *sine qua non* constraint. For example, Theorem 3.3 gives meaningful bounds if $M=\exp(n^\gamma)$ for arbitrarily large $\gamma>0$, provided that $T>n^{2\gamma}$. This is due to the structured sparsity assumption, and is not conditioned by the block-diagonal (multi-task) structure of the regression matrices.

3. *Our estimator admits better risk bounds than the usual Lasso.* Let us explain this point considering the example of the prediction error bound (3.6). Indeed, for the same multi-task setup, we can apply a usual Lasso estimator $\hat\beta^L$, that is a solution of the following optimization problem

$$\min_\beta S(\beta)+2\lambda'\sum_{t=1}^T\sum_{j=1}^M|\beta_{tj}|$$

where $\lambda'>0$ is a tuning parameter. We will use the bounds of [3] for the prediction error of $\hat\beta^L$. For a fair comparison with Theorem 3.3, we assume that we are in the most favorable situation where $M<n$, each of the matrices $\frac{1}{n}X_t^TX_t$ is positive definite and has minimal eigenvalue greater than κ^2. This implies both Assumption 3.1 and the Restricted Eigenvalue assumption as stated in [3]. Next, we assume, as in Theorem 3.3, that the diagonal elements of the matrix $X^\top X/n$ are equal to 1.

To use the results of [3], we note that the parameters n, M, s therein correspond to $n'=nT$, $M'=MT$, $s'=sT$ in our setup, and the minimal eigenvalue of the matrix $\frac{1}{n'}X^TX=\frac{1}{nT}X^TX$ is greater than $(\kappa')^2\equiv\kappa^2/T$. Another particularity is that, due to our normalization, the diagonal elements of the matrix $\frac{1}{nT}X^TX$ are equal to $1/T$, and not to 1, as in [3]. This results in the fact that the correct λ' is by a $\sqrt{T}$ factor smaller than that given in [3]:

$$\lambda' = A' \frac{\sigma}{\sqrt{T}} \sqrt{\frac{\log(MT)}{nT}},$$

where $A' > 2\sqrt{2}$. We can then act as in the proof of inequality (7.8) from [3] (cf. (B.31) in [3]) to obtain that, with probability at least $1 - (MT)^{1-\frac{(A')^2}{8}}$, it holds

$$\begin{aligned} \frac{1}{nT}||X(\hat{\beta}^L - \beta^*)||^2 &\leq \frac{16 s'(\lambda')^2}{(\kappa')^2} \\ &= \frac{16(A')^2}{\kappa^2}\sigma^2 s \frac{\log(MT)}{n}. \end{aligned}$$

Comparing with (3.6) we conclude that if $\log M$ is not too large as compared to $\sqrt{T}$ the rate of prediction bound (3.6) for the Group Lasso is by a factor of $\log(MT)$ better than for the usual Lasso under the same assumptions. Let us emphasize that the improvement is only due to the property that β^* is structured sparse.

Finally, we note that [23] follow the scheme of the proof of [3] to derive similar in spirit to ours but coarse oracle inequalities. Their results do not explain the advantages discussed in the points 1–3 above. Indeed, the tuning parameter λ chosen in [23], pp. 614–615, is larger than our λ by at least a factor of $\sqrt{T}$. As a consequence, the corresponding bounds in the oracle inequalities of [23] are larger than ours by positive powers of T.

4 Coordinate-wise estimation and selection of sparsity pattern

In this section, we show how from any solution of the problem (2.2) we can reliably estimate the correct sparsity pattern with high probability.

We first introduce some more notation. We define the Gram matrix of the design $\Psi = \frac{1}{n} X^\top X$. Note that Ψ is a $MT \times MT$ block-diagonal matrix with T blocks of dimension $M \times M$ each. We denote these blocks by $\Psi_t = \frac{1}{n} X_t^\top X_t \equiv (\Psi_{tj,tk})_{j,k=1,\dots,M}$.

In this section we assume that the following condition holds true.

Assumption 4.1 *The elements $\Psi_{tj,tk}$ of the Gram matrix Ψ satisfy*

$$\Psi_{tj,tj} = 1, \ \ \forall 1 \leqslant j \leqslant M, \, 1 \leqslant t \leqslant T,$$

and

$$\max_{1 \leqslant t \leqslant T, j \neq k} |\Psi_{tj,tk}| \leqslant \frac{1}{7\alpha s},$$

for some integer $s \geqslant 1$ and some constant $\alpha > 1$.

Note that the above assumption on Ψ implies Assumption 3.1 as we prove in the following lemma.

Lemma 4.2 *Let Assumption 4.1 be satisfied. Then Assumption 3.1 is satisfied with $\kappa = \sqrt{1 - \frac{1}{\alpha}}$.*

Proof: For any subset J of $\{1, \dots, M\}$ such that $|J| \leqslant s$ and any $\Delta \in \mathbb{R}^{MT}$ such that $\|\Delta_{J^c}\|_{2,1} \leqslant 3\|\Delta_J\|_{2,1}$, we have

$$\begin{aligned} \frac{\Delta_J^\top \Psi \Delta_J}{\|\Delta_J\|^2} &= 1 + \frac{\Delta_J^\top (\Psi - I_{MT \times MT})\Delta_J}{\|\Delta_J\|^2} \\ &\geqslant 1 - \frac{1}{7\alpha s} \frac{\left(\sum_{j \in J} \sum_{t=1}^T |\Delta_{tj}|\right)^2}{\|\Delta_J\|^2} \\ &\geqslant 1 - \frac{1}{7\alpha} \end{aligned}$$

where we have used Assumption 4.1 and the Cauchy-Schwarz inequality. Next, using consecutively Assumption 4.1, the Cauchy-Schwarz inequality and the inequality $\|\Delta_{J^c}\|_{2,1} \leqslant 3\|\Delta_J\|_{2,1}$ we obtain

$$\begin{aligned} \frac{|\Delta_{J^c}^\top \Psi \Delta_J|}{\|\Delta_J\|^2} &\leqslant \frac{1}{7\alpha s} \frac{\sum_{t=1}^T \sum_{j \in J} \sum_{k \in J^c} |\Delta_{tj}||\Delta_{tk}|}{\|\Delta_J\|^2} \\ &\leqslant \frac{1}{7\alpha s} \frac{\sum_{j \in J, k \in J^c} \|\Delta^j\| \|\Delta^k\|}{\|\Delta_J\|^2} \\ &\leqslant \frac{3}{7\alpha s} \frac{\|\Delta_J\|_{2,1}^2}{\|\Delta_J\|^2} \\ &\leqslant \frac{3}{7\alpha}. \end{aligned}$$

Combining these inequalities we find

$$\frac{\Delta^\top \Psi \Delta}{\|\Delta_J\|^2} \geqslant \frac{\Delta_J^\top \Psi \Delta_J}{\|\Delta_J\|^2} + \frac{2\Delta_{J^c}^\top \Psi \Delta_J}{\|\Delta_J\|^2} \geqslant 1 - \frac{1}{\alpha} > 0.$$

■

Note also that, by an argument as in [19], it is not hard to show that under Assumption 4.1 the vector β^* satisfying (2.1) is unique.

Theorem 3.3 provides bounds for compound measures of risk, that is, depending simultaneously on all the vectors β^j. An important question is to evaluate the performance of estimators for each of the components β^j separately. The next theorem provides a bound of this type and, as a consequence, a result on the selection of sparsity pattern.

Theorem 4.3 *Consider the model (1.1) for $M \geqslant 2$ and $T, n \geqslant 1$. Let the assumptions of Lemma 3.2 be satisfied and let Assumption 4.1 hold with the same s. Set*

$$c = \left(3 + \frac{32}{7(\alpha - 1)}\right)\sigma.$$

Let λ, A and $W_1, \dots, W_T$ be as in Lemma 3.2. Then with probability at least $1 - M^{1-q}$, where $q = \min(8 \log M, ByA\sqrt{T}/8)$, for any solution $\hat{\beta}$ of problem (2.2) we have

$$\frac{1}{\sqrt{T}} \|\hat{\beta} - \beta^*\|_{2,\infty} \leqslant \frac{c}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}. \tag{4.1}$$

If, in addition,

$$\min_{j \in J(\beta^*)} \frac{1}{\sqrt{T}} \|(\beta^*)^j\| > \frac{2c}{\sqrt{n}} \sqrt{1 + \frac{A \log M}{\sqrt{T}}}, \tag{4.2}$$

then with the same probability for any solution $\hat\beta$ of problem (2.2) the set of indices

$$\hat J = \left\{ j : \frac{1}{\sqrt{T}}\|\hat\beta^j\| > \frac{c}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}\right\} \quad (4.3)$$

estimates correctly the sparsity pattern $J(\beta^*)$, that is,

$$\hat J = J(\beta^*).$$

Proof: Set $\Delta = \hat\beta - \beta^*$. We have

$$\|\Delta\|_{2,\infty} \leqslant \|\Psi\Delta\|_{2,\infty} + \|(\Psi - I_{MT\times MT})\Delta\|_{2,\infty}. \quad (4.4)$$

Using Assumption 4.1 we obtain

$$\|(\Psi - I_{MT\times MT})\Delta\|_{2,\infty} =$$

$$\max_{1\leqslant j\leqslant M}\left[\sum_{t=1}^T\left(\sum_{k=1:k\neq j}^M |\Psi_{tj,tk}||\Delta_{tk}|\right)^2\right]^{1/2} \leq$$

$$\max_{1\leqslant j\leqslant M}\left[\max_{1\leqslant t\leqslant T, j\neq k}|\Psi_{tj,tk}|^2\sum_{t=1}^T\left(\sum_{k=1:k\neq j}^M |\Delta_{tk}|\right)^2\right]^{1/2} \leq$$

$$\frac{1}{7\alpha s}\left[\sum_{t=1}^T\left(\sum_{k=1}^M |\Delta_{tk}|\right)^2\right]^{1/2}. \quad (4.5)$$

By the Minkowski inequality for the Euclidean norm in $\mathbb{R}^T$,

$$\left[\sum_{t=1}^T\left(\sum_{k=1}^M |\Delta_{tk}|\right)^2\right]^{1/2} \leqslant \|\Delta\|_{2,1}. \quad (4.6)$$

Combining the three above displays we get

$$\|\Delta\|_{2,\infty} \leqslant \|\Psi\Delta\|_{2,\infty} + \frac{1}{7\alpha s}\|\Delta\|_{2,1}.$$

Thus, by Lemma 3.2 and Theorem 3.1, with probability at least $1 - M^{1-q}$,

$$\|\Delta\|_{2,\infty} \leqslant \left(\frac{3}{2} + \frac{16}{7\alpha\kappa^2}\right)\lambda T.$$

By Lemma 4.2, $\alpha\kappa^2 = \alpha - 1$, which yields the first result of the theorem. The second result follows from the first one in an obvious way. ■

Assumption of type (4.2) is inevitable in the context of selection of sparsity pattern. It says that the vectors $(\beta^*)^j$ cannot be arbitrarily close to 0 for j in the pattern. Their norms should be at least somewhat larger than the noise level.

The second result of Theorem 4.3 (selection of sparsity pattern) can be compared with [2, 23] who considered the Group Lasso. There are several differences. First, our estimator $\hat J$ is based on thresholding of the norms $\|\hat\beta^j\|$, while [2, 23] take instead the set where these norms do not vanish. In practice, the latter is known to be a poor selector; it typically overestimates the true sparsity pattern. Second, [2, 23] consider specific asymptotic settings, while our result holds for any fixed n, M, T. Different kinds of asymptotics can be therefore obtained as simple corollaries. Finally, note that the estimator $\hat\beta$ is not necessarily unique. Though [23] does not discuss this fact, the proof there only shows that *there exists a subsequence of solutions $\hat\beta$ of (2.2)* such that the set $\{j : \|\hat\beta^j\| \neq 0\}$ coincides with the sparsity pattern $J(\beta^*)$ in some specified asymptotics (we note that the "if and only if" claim before formula (23) in [23] is not proved). In contrast, the argument in Theorem 4.3 does not require any analysis of the uniqueness issues, though it is not excluded that the solution is indeed unique. It guarantees that *simultaneously for all solutions $\hat\beta$ of (2.2)* and any fixed n, M, T the correct selection is done with high probability.

Theorems 3.3 and 4.3 imply the following corollary.

Corollary 4.4 *Consider the model (1.1) for $M \geqslant 2$ and $T, n \geqslant 1$. Let the assumptions of Lemma 3.2 be satisfied and let Assumption 4.1 holds with the same s. Let λ, A and $W_1, \dots, W_T$ be as in Lemma 3.2. Then with probability at least $1 - M^{1-q}$, where $q = \min(8\log M, A\sqrt{T}/8)$, for any solution $\hat\beta$ of problem (2.2) and any $1 \leq p < \infty$ we have*

$$\frac{1}{\sqrt{T}}\|\hat\beta - \beta^*\|_{2,p} \leqslant c_1\sigma\frac{s^{1/p}}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}, \quad (4.7)$$

where

$$c_1 = \left(\frac{32\alpha}{\alpha-1}\right)^{1/p}\left(3 + \frac{32}{7(\alpha-1)}\right)^{1-\frac{1}{p}}.$$

If, in addition, (4.2) holds, then with the same probability for any solution $\hat\beta$ of problem (2.2) and any $1 \leq p < \infty$ we have

$$\frac{1}{\sqrt{T}}\|\hat\beta - \beta^*\|_{2,p} \leqslant c_1\sigma\frac{|\hat J|^{1/p}}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}, \quad (4.8)$$

where $\hat J$ is defined in (4.3).

Proof: Set $\Delta = \hat\beta - \beta$. For any $p \geqslant 1$ we have

$$\frac{1}{\sqrt{T}}\|\Delta\|_{2,p} \leqslant \left(\frac{1}{\sqrt{T}}\|\Delta\|_{2,1}\right)^{\frac{1}{p}}\left(\frac{1}{\sqrt{T}}\|\Delta\|_{2,\infty}\right)^{1-\frac{1}{p}}.$$

Combining (3.7), (4.1) with $\kappa = \sqrt{1-\frac{1}{\alpha}}$ and the above display yields the first result. ■

Inequalities (4.1) and (4.8) provide confidence intervals for the unknown parameter β^* in mixed (2,p)-norms.

For averages of the coefficients β_{tj} we can establish a sign consistency result which is somewhat stronger than the result in Theorem 4.3. For any $\beta \in \mathbb{R}^M$, define $\vec{\text{sign}}(\beta) = (\text{sign}(\beta^1), \dots, \text{sign}(\beta^M))^\top$ where

$$\text{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

Introduce the averages

$$a_j^* = \frac{1}{T}\sum_{t=1}^T \beta_{tj}^*, \quad \hat a_j = \frac{1}{T}\sum_{t=1}^T \hat\beta_{tj}.$$

Consider the threshold $\tau = \frac{c}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}$ and define a thresholded estimator

$$\tilde{a}_j = \hat{a}_j I\{|\hat{a}_j| > \tau\}.$$

Let $\tilde{a}$ and a^* be the vectors with components $\tilde{a}_j$ and a_j^*, $j = 1,\ldots,M$, respectively. We need the following additional assumption.

Assumption 4.5 *It holds that*

$$\min_{j\in J(a^*)} |a_j^*| \geq \frac{2c}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}.$$

This assumption says that we cannot recover arbitrarily small components. Similar assumptions are standard in the literature on sign consistency (see, for example, [19] for more details and references).

Theorem 4.6 *Consider the model (1.1) for $M \geqslant 2$ and $T, n \geqslant 1$. Let the assumptions of Lemma 3.2 be satisfied and let Assumption 4.1 hold with the same s. Let λ and A be defined as in Lemma 3.2 and c as in Theorem 4.3. Then with probability at least $1 - M^{1-q}$, where $q = \min(8\log M, A\sqrt{T}/8)$, for any solution $\hat{\beta}$ of problem (2.2) we have*

$$\max_{1\leqslant j\leqslant M} |\hat{a}_j - a_j^*| \leqslant \frac{c}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}.$$

If, in addition, Assumption 4.5 holds, then with the same probability, for any solution $\hat{\beta}$ of problem (2.2), $\tilde{a}$ recovers the sign pattern of a^*:

$$\vec{\text{sign}}(\tilde{a}) = \vec{\text{sign}}(a^*).$$

Proof: Note that for every $j \in \mathbb{N}_M$

$$|\hat{a}_j - a_j^*| \leq \frac{1}{\sqrt{T}}\|\hat{\beta}-\beta^*\|_{2,\infty} \leq \frac{c}{\sqrt{n}}\sqrt{1+\frac{A\log M}{\sqrt{T}}}.$$

The proof is then similar to that of Theorem 4.3. ■

5 Non-Gaussian noise

In this section, we only assume that the random variables $W_{ti}, i \in \mathbb{N}_n, t \in \mathbb{N}_T$, are independent with zero mean and finite variance $\mathbb{E}[W_{ti}^2] \leqslant \sigma^2$. In this case the results remain similar to those of the previous sections, though the concentration effect is weaker. We need the following technical assumption

Assumption 5.1 *The matrix X is such that*

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\max_{1\leqslant j\leqslant M}|(x_{ti})_j|^2 \leqslant c',$$

for a constant $c' > 0$.

This assumption is quite mild. It is satisfied for example, if all $(x_{ti})_j$ are bounded in absolute value by a constant uniformly in i, t, j. We have the two following theorems.

Theorem 5.2 *Consider the model* (1.1) *for $M \geq 3$ and $T, n \geq 1$. Assume that the random vectors $W_1,\ldots,W_T$ are independent with zero mean and finite variance $\mathbb{E}[W_{ti}^2] \leqslant \sigma^2$, all diagonal elements of the matrix $X^\top X/n$ are equal to 1 and $M(\beta^*) \leq s$. Let also Assumption 5.1 be satisfied. Furthermore let κ be defined as in Assumption 3.1 and $\phi_{\max}$ be the largest eigenvalue of the matrix $X^\top X/n$. Let*

$$\lambda = \sigma\sqrt{\frac{(\log M)^{1+\delta}}{nT}}, \quad \delta > 0.$$

Then with probability at least $1 - \frac{(2e\log M - e)c'}{(\log M)^{1+\delta}}$, for any solution $\hat{\beta}$ of problem (2.2) *we have*

$$\frac{1}{nT}\|X(\hat{\beta}-\beta^*)\|^2 \leq \frac{16}{\kappa^2}\sigma^2 s\frac{(\log M)^{1+\delta}}{n}, \tag{5.1}$$

$$\frac{1}{\sqrt{T}}\|\hat{\beta}-\beta^*\|_{2,1} \leq \frac{16}{\kappa^2}\sigma s\sqrt{\frac{(\log M)^{1+\delta}}{n}}, \tag{5.2}$$

$$M(\hat{\beta}) \leq \frac{64\phi_{\max}}{\kappa^2}s. \tag{5.3}$$

If, in addition, Assumption RE(2s) holds, then with the same probability for any solution $\hat{\beta}$ of problem (2.2) *we have*

$$\frac{1}{T}\|\hat{\beta}-\beta^*\|^2 \leq \frac{160}{\kappa^4(2s)}\sigma^2 s\frac{(\log M)^{1+\delta}}{n}.$$

Theorem 5.3 *Consider the model (1.1) for $M \geqslant 3$ and $T, n \geqslant 1$. Let the assumptions of Theorem 5.2 be satisfied and let Assumption 4.1 hold with the same s. Set*

$$c = \left(\frac{3}{2} + \frac{1}{7(\alpha-1)}\right)\sigma.$$

Let λ be as in Theorem as in 5.2. Then with probability at least $1 - \frac{(2e\log M - e)c'}{(\log(MT))^{1+\delta}}$, for any solution $\hat{\beta}$ of problem (2.2) we have

$$\frac{1}{\sqrt{T}}\|\hat{\beta}-\beta^*\|_{2,\infty} \leqslant c\sqrt{\frac{(\log M)^{1+\delta}}{n}}.$$

If, in addition, it holds that

$$\min_{j\in J(\beta^*)} \frac{1}{\sqrt{T}}\|(\beta^*)^j\| > 2c\sqrt{\frac{(\log M)^{1+\delta}}{n}},$$

then with the same probability for any solution $\hat{\beta}$ of problem (2.2) the set of indices

$$\hat{J} = \left\{j : \frac{1}{\sqrt{T}}\|\hat{\beta}^j\| > c\sqrt{\frac{(\log M)^{1+\delta}}{n}}\right\}$$

estimates correctly the sparsity pattern $J(\beta^*)$:

$$\hat{J} = J(\beta^*).$$

The proofs of these theorems are similar to the ones of Theorems 3.3 and 4.3 up to a modification of the bound on $P(\mathcal{A}^c)$ in Lemma 3.2. We consider now the event

$$\mathcal{A} = \left\{\max_{j=1}^{M}\sqrt{\sum_{t=1}^{T}\left(\sum_{i=1}^{n}(x_{ti})_j W_{ti}\right)^2} \leq \lambda nT\right\}.$$

The Markov inequality yields that

$$\Pr(\mathcal{A}^c) \leqslant \frac{\sum_{t=1}^{T} \mathbb{E}[\max_{1\leqslant j\leqslant M} \left(\sum_{i=1}^{n} (x_{ti})_j W_{ti}\right)^2]}{(\lambda n T)^2}.$$

Then we use Lemma A.2 given below with the random vectors

$$Y_{ti} = ((x_{ti})_1 W_{ti}/n, \ldots, (x_{ti})_M W_{ti}/n) \in \mathbb{R}^M,$$

$\forall i \in \mathbb{N}_n, \forall t \in \mathbb{N}_T$. We get that

$$\Pr(\mathcal{A}^c) \quad \leqslant \quad \frac{2e\log M - e}{\lambda^2 n T}\sigma^2 \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n} \max_{1\leqslant j\leqslant M} |(x_{ti})_j|^2.$$

By the definition of λ in Theorem 5.2 and Assumption 5.1 we obtain

$$\Pr(\mathcal{A}^c) \quad \leqslant \quad \frac{(2e\log M - e)c'}{(\log M)^{1+\delta}}.$$

■

Thus, we see that under the finite variance assumption on the noise, the dependence on the dimension M cannot be made negligible for large T.

Acknowledgments

Part of this work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778 as well as by the EPSRC Grant EP/D071542/1.

A Auxiliary results

Here we collect two auxiliary results which are used in the above analysis. The first result is a useful bound on the tail of the chi-square distribution.

Lemma A.1 *Let χ_T^2 be a chi-square random variable with T degrees of freedom. Then*

$$\Pr(\chi_T^2 > T + x) \leq \exp\left(-\frac{1}{8}\min\left(x, \frac{x^2}{T}\right)\right)$$

for all $x > 0$.

Proof: By the Wallace inequality [27] we have

$$\Pr(\chi_T^2 > T + x) \leq \Pr(\mathcal{N} > z(x)),$$

where $\mathcal{N}$ is the standard normal random variable and $z(x) = \sqrt{x - T\log(1 + x/T)}$. The result now follows from inequalities $\Pr(\mathcal{N} > z(x)) \leq \exp(-z^2(x)/2)$ and

$$u - \log(1+u) \geq \frac{u^2}{2(1+u)} \geq \frac{1}{4}\min\left(u, u^2\right), \ \forall u > 0.$$

■

The next result is a version of Nemirovski's inequality (see [12], Corollary 2.4 page 5).

Lemma A.2 *Let $Y_1, \ldots, Y_n \in \mathbb{R}^M$ be independent random vectors with zero means and finite variance, and let $M \geqslant 3$. Then*

$$\mathbb{E}\left[|\sum_{i=1}^{n} Y_i|_\infty^2\right] \leqslant (2e\log M - e)\sum_{i=1}^{n}\mathbb{E}\left[|Y_i|_\infty^2\right],$$

where $|\cdot|_\infty$ is the ℓ_∞ norm.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[2] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[3] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics (to appear)*.

[4] J.M. Borwein and A.S. Lewis. *Convex Analysis And Nonlinear Optimization: Theory And Examples*. Springer, 2006.

[5] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.

[6] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[7] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2005.

[8] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2008.

[9] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variable Lasso. http://hal.archives-ouvertes.fr/hal-00145160/fr/. 2007.

[10] P. Diggle. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

[11] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

[12] L. Dümbgen, S.A. van de Geer, and J.A. Wellner. Nemirovski's inequalities revisited. Available on Arxiv, 2008.

[13] T. Evgeniou, M. Pontil, and O. Toubia. A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science*, (26):805–818, 2007.

[14] C. Hsiao. *Analysis of Panel Data*. Cambridge University Press, 2003.

[15] J. Huang, J.L. Horowitz, and F Wei. Variable selection in nonparametric additive models. Manuscript. 2008.

[16] J. Huang and T. Zhang. The benefit of group sparsity. Available on Arxiv, 2009.

[17] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Conference on Learning Theory, COLT-2008*, pages 229–238, 2008.

[18] P. J. Lenk, W. S. DeSarbo, P. E. Green, and M. R. Young. Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.

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

[20] A. Maurer. Bounds for linear multi-task learning. *The Journal of Machine Learning Research*, 7:117–139, 2006.

[21] L. Meier, S. van de Geer, and P. Buhlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*, 70(1):53–57, 2006.

[22] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. arXiv:0806.4115. 2008.

[23] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group Lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

[24] G. Obozinski, M.J. Wainwright, and M.I. Jordan. Union support recovery in high-dimensional multivariate regression. 2008.

[25] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 22, 2007.

[26] S.A. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36(2):614, 2008.

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

[28] J.M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

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