Journal of Machine Learning Research 12 (2011) 2415-2435 Submitted 8/10; Revised 2/11; Published 7/11

# Union Support Recovery in Multi-task Learning

**Mladen Kolar** MLADENK@CS.CMU.EDU
**John Lafferty** LAFFERTY@CS.CMU.EDU
**Larry Wasserman** LARRY@STAT.CMU.EDU
*School of Computer Science*
*Carnegie Mellon University*
*5000 Forbes Avenue*
*Pittsburgh, PA 15213, USA*

**Editor:** Hui Zou

## Abstract

We sharply characterize the performance of different penalization schemes for the problem of selecting the relevant variables in the multi-task setting. Previous work focuses on the regression problem where conditions on the design matrix complicate the analysis. A clearer and simpler picture emerges by studying the Normal means model. This model, often used in the field of statistics, is a simplified model that provides a laboratory for studying complex procedures.

**Keywords:** high-dimensional inference, multi-task learning, sparsity, normal means, minimax estimation

## 1. Introduction

We consider the problem of estimating a sparse signal in the presence of noise. It has been empirically observed, on various data sets ranging from cognitive neuroscience (Liu et al., 2009) to genome-wide association mapping studies (Kim et al., 2009), that considering related estimation tasks jointly can improve estimation performance. Because of this, joint estimation from related tasks or *multi-task learning* has received much attention in the machine learning and statistics community (see for example Turlach et al., 2005; Zou and Yuan, 2008; Zhang, 2006; Negahban and Wainwright, 2009; Obozinski et al., 2011; Lounici et al., 2009; Liu et al., 2009; Lounici et al., 2010; Argyriou et al., 2008; Kim et al., 2009, and references therein). However, the theory behind multi-task learning is not yet settled.

An example of multi-task learning is the problem of estimating the coefficients of several multiple regressions

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \qquad j \in [k] \tag{1}$$

where $\mathbf{X}_j \in \mathbb{R}^{n \times p}$ is the design matrix, $\mathbf{y}_j \in \mathbb{R}^n$ is the vector of observations, $\boldsymbol{\epsilon}_j \in \mathbb{R}^n$ is the noise vector and $\boldsymbol{\beta}_j \in \mathbb{R}^p$ is the unknown vector of regression coefficients for the $j$-th task, with $[k] = \{1, \ldots, k\}$.

When the number of variables $p$ is much larger than the sample size $n$, it is commonly assumed that the regression coefficients are jointly sparse, that is, there exists a small subset $S \subset [p]$ of the regression coefficients, with $s := |S| \ll n$, that are non-zero for all or most of the tasks.

The model in (1) under the joint sparsity assumption was analyzed in, for example, Obozinski et al. (2011), Lounici et al. (2009), Negahban and Wainwright (2009), Lounici et al. (2010) and

Kolar and Xing (2010). Obozinski et al. (2011) propose to minimize the penalized least squares objective with a mixed $(2,1)$-norm on the coefficients as the penalty term. The authors focus on consistent estimation of the support set $S$, albeit under the assumption that the number of tasks $k$ is fixed. Negahban and Wainwright (2009) use the mixed $(\infty,1)$-norm to penalize the coefficients and focus on the exact recovery of the non-zero pattern of the regression coefficients, rather than the support set $S$. For a rather limited case of $k = 2$, the authors show that when the regression do not share a common support, it may be harmful to consider the regression problems jointly using the mixed $(\infty,1)$-norm penalty. Kolar and Xing (2010) address the feature selection properties of simultaneous greedy forward selection. However, it is not clear what the benefits are compared to the ordinary forward selection done on each task separately. In Lounici et al. (2009) and Lounici et al. (2010), the focus is shifted from the consistent selection to benefits of the joint estimation for the prediction accuracy and consistent estimation. The number of tasks $k$ is allowed to increase with the sample size. However, it is assumed that all tasks share the same features; that is, a relevant coefficient is non-zero for all tasks.

Despite these previous investigations, the theory is far from settled. A simple clear picture of when sharing between tasks actually improves performance has not emerged. In particular, to the best of our knowledge, there has been no previous work that sharply characterizes the performance of different penalization schemes on the problem of selecting the relevant variables in the multi-task setting.

In this paper we study multi-task learning in the context of the *many Normal means model*. This is a simplified model that is often useful for studying the theoretical properties of statistical procedures. The use of the many Normal means model is fairly common in statistics but appears to be less common in machine learning. Our results provide a sharp characterization of the sparsity patterns under which the Lasso procedure performs better than the group Lasso. Similarly, our results characterize how the group Lasso (with the mixed $(2,1)$ norm) can perform better when each non-zero row is dense.

### 1.1 The Normal Means Model

The simplest Normal means model has the form

$$Y_i = \mu_i + \sigma\varepsilon_i, \quad i = 1,\ldots,p \tag{2}$$

where $\mu_1,\ldots,\mu_p$ are unknown parameters and $\varepsilon_1,\ldots,\varepsilon_p$ are independent, identically distributed Normal random variables with mean 0 and variance 1. There are a variety of results (Brown and Low, 1996; Nussbaum, 1996) showing that many learning problems can be converted into a Normal means problem. This implies that results obtained in the Normal means setting can be transferred to many other settings. As a simple example, consider the nonparametric regression model $Z_i = m(i/n) + \delta_i$ where $m$ is a smooth function on $[0,1]$ and $\delta_i \sim N(0,1)$. Let $\phi_1,\phi_2,\ldots,$ be an orthonormal basis on [0,1] and write $m(x) = \sum_{j=1}^{\infty} \mu_j \phi_j(x)$ where $\mu_j = \int_0^1 m(x)\phi_j(x)dx$. To estimate the regression function $m$ we need only estimate $\mu_1,\mu_2,\ldots,$. Let $Y_j = n^{-1}\sum_{i=1}^n Z_i \phi_j(i/n)$. Then $Y_j \approx N(\mu_j,\sigma^2)$ where $\sigma^2 = 1/n$. This has the form of (2) with $\sigma = 1/\sqrt{n}$. Hence this regression problem can be converted into a Normal means model.

However, the most important aspect of the Normal means model is that it allows a clean setting for studying complex problems. In this paper, we consider the following Normal means model. Let

$$Y_{ij} \sim \begin{cases} (1-\varepsilon)\mathcal{N}(0,\sigma^2)+\varepsilon\mathcal{N}(\mu_{ij},\sigma^2) & j\in[k], \quad i\in S \\ N(0,\sigma^2) & j\in[k], \quad i\in S^c \end{cases} \tag{3}$$

where $(\mu_{ij})_{i,j}$ are unknown real numbers, $\sigma=\sigma_0/\sqrt{n}$ is the variance with $\sigma_0>0$ known, $(Y_{ij})_{i,j}$ are random observations, $\varepsilon\in[0,1]$ is the parameter that controls the sparsity of features across tasks and $S\subset[p]$ is the set of relevant features. Let $s=|S|$ denote the number of relevant features. Denote the matrix $M\in\mathbb{R}^{p\times k}$ of means

| | Tasks | | | |
|---|---|---|---|---|
| | 1 | 2 | … | $k$ |
| 1 | $\mu_{11}$ | $\mu_{12}$ | $\cdots$ | $\mu_{1k}$ |
| 2 | $\mu_{21}$ | $\mu_{22}$ | $\cdots$ | $\mu_{2k}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $p$ | $\mu_{p1}$ | $\mu_{p2}$ | $\cdots$ | $\mu_{pk}$ |

and let $\boldsymbol{\theta}_i=(\mu_{ij})_{j\in[k]}$ denote the $i$-th row of the matrix $M$. The set $S^c=[p]\backslash S$ indexes the zero rows of the matrix $M$ and the associated observations are distributed according to the Normal distribution with zero mean and variance $\sigma^2$. The rows indexed by $S$ are non-zero and the corresponding observation are coming from a mixture of two Normal distributions. The parameter $\varepsilon$ determines the proportion of observations coming from a Normal distribution with non-zero mean. The reader should regard each column as one vector of parameters that we want to estimate. The question is whether sharing across columns improves the estimation performance.

It is known from the work on the Lasso that in regression problems, the design matrix needs to satisfy certain conditions in order for the Lasso to correctly identify the support $S$ (see van de Geer and Bühlmann, 2009, for an extensive discussion on the different conditions). These regularity conditions are essentially unavoidable. However, the Normal means model (3) allows us to analyze the estimation procedure in (4) and focus on the scaling of the important parameters $(n,k,p,s,\varepsilon,\mu_{\min})$ for the success of the support recovery. Using the model (3) and the estimation procedure in (4), we are able to identify regimes in which estimating the support is more efficient using the ordinary Lasso than with the multi-task Lasso and vice versa. Our results suggest that the multi-task Lasso does not outperform the ordinary Lasso when the features are not considerably shared across tasks; thus, practitioners should be careful when applying the multi-task Lasso without knowledge of the task structure.

An alternative representation of the model is

$$Y_{ij}=\begin{cases} \mathcal{N}(\xi_{ij}\mu_{ij},\sigma^2) & j\in[k], \quad i\in S \\ N(0,\sigma^2) & j\in[k], \quad i\in S^c \end{cases}$$

where $\xi_{ij}$ is a Bernoulli random variable with success probability $\varepsilon$. Throughout the paper, we will set $\varepsilon=k^{-\beta}$ for some parameter $\beta\in[0,1)$; $\beta<1/2$ corresponds to dense rows and $\beta>1/2$ corresponds to sparse rows. Let $\mu_{\min}$ denote the following quantity $\mu_{\min}=\min|\mu_{ij}|$.

Under the model (3), we analyze penalized least squares procedures of the form

$$\widehat{\boldsymbol{\mu}}=\underset{\boldsymbol{\mu}\in\mathbb{R}^{p\times k}}{\operatorname{argmin}}\ \frac{1}{2}||\mathbf{Y}-\boldsymbol{\mu}||_F^2+\operatorname{pen}(\boldsymbol{\mu}) \tag{4}$$

where $||A||_F = \sum_{jk} A_{jk}^2$ is the Frobenious norm, pen$(\cdot)$ is a penalty function and $\boldsymbol{\mu}$ is a $p \times k$ matrix of means. We consider the following penalties:

1. the $\ell_1$ penalty
$$\text{pen}(\boldsymbol{\mu}) = \lambda \sum_{i\in[p]} \sum_{j\in[k]} |\mu_{ij}|,$$
which corresponds to the Lasso procedure applied on each task independently, and denote the resulting estimate as $\widehat{\boldsymbol{\mu}}^{\ell_1}$

2. the mixed $(2,1)$-norm penalty
$$\text{pen}(\boldsymbol{\mu}) = \lambda \sum_{i\in[p]} ||\boldsymbol{\theta}_i||_2,$$
which corresponds to the multi-task Lasso formulation in Obozinski et al. (2011) and Lounici et al. (2009), and denote the resulting estimate as $\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_2}$

3. the mixed $(\infty,1)$-norm penalty
$$\text{pen}(\boldsymbol{\mu}) = \lambda \sum_{i\in[p]} ||\boldsymbol{\theta}_i||_\infty,$$
which correspond to the multi-task Lasso formulation in Negahban and Wainwright (2009), and denote the resulting estimate as $\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_\infty}$.

For any solution $\widehat{\boldsymbol{\mu}}$ of (4), let $S(\widehat{\boldsymbol{\mu}})$ denote the set of estimated non-zero rows

$$S(\widehat{\boldsymbol{\mu}}) = \{i \in [p] \ : \ ||\widehat{\boldsymbol{\theta}}_i||_2 \neq 0\}.$$

We establish sufficient conditions under which $\mathbb{P}[S(\widehat{\boldsymbol{\mu}}) \neq S] \leq \alpha$ for different methods. These results are complemented with necessary conditions for the recovery of the support set $S$.

In this paper, we focus our attention on the three penalties outlined above. There is a large literature on the penalized least squares estimation using concave penalties as introduced in Fan and Li (2001). These penalization methods have better theoretical properties in the presence of the design matrix, especially when the design matrix is far from satisfying the irrepresentable condition (Zhao and Yu, 2006). In the Normal means model, due to the lack of the design matrix, there is no advantage to concave penalties in terms of variable selection.

### 1.2 Overview of the Main Results

The main contributions of the paper can be summarized as follows.

1. We establish a lower bound on the parameter $\mu_{\min}$ as a function of the parameters $(n,k,p,s,\beta)$. Our result can be interpreted as follows: for any estimation procedure there exists a model given by (3) with non-zero elements equal to $\mu_{\min}$ such that the estimation procedure will make an error when identifying the set $S$ with probability bounded away from zero.

2. We establish the sufficient conditions on the signal strength $\mu_{\min}$ for the Lasso and both variants of the group Lasso under which these procedures can correctly identify the set of non-zero rows $S$.

By comparing the lower bounds with the sufficient conditions, we are able to identify regimes in which each procedure is optimal for the problem of identifying the set of non-zero rows $S$. Furthermore, we point out that the usage of the popular group Lasso with the mixed $(\infty,1)$ norm can be disastrous when features are not perfectly shared among tasks. This is further demonstrated through an empirical study.

### 1.3 Organization of the Paper

The paper is organized as follows. We start by analyzing the lower bound for any procedure for the problem of identifying the set of non-zero rows in Section 2. In Section 3 we provide sufficient conditions on the signal strength $\mu_{\min}$ for the Lasso and the group Lasso to be able to detect the set of non-zero rows $S$. In the following section, we propose an improved approach to the problem of estimating the set $S$. Results of a small empirical study are reported in Section 4. We close the paper by a discussion of our findings.

## 2. Lower Bound on the Support Recovery

In this section, we derive a lower bound for the problem of identifying the correct variables. In particular, we derive conditions on $(n,k,p,s,\varepsilon,\mu_{\min})$ under which any method is going to make an error when estimating the correct variables. Intuitively, if $\mu_{\min}$ is very small, a non-zero row may be hard to distinguish from a zero row. Similarly, if $\varepsilon$ is very small, many elements in a row will be zero and, again, as a result it may be difficult to identify a non-zero row. Before, we give the main result of the section, we introduce the class of models that are going to be considered.

Let

$$\mathcal{F}[\mu] := \{\boldsymbol{\theta} \in \mathbb{R}^k \,:\, \min_j |\theta_j| \geq \mu\}$$

denote the set of feasible non-zero rows. For each $j \in \{0,1,\ldots,k\}$, let $\mathcal{M}(j,k)$ be the class of all the subsets of $\{1,\ldots,k\}$ of cardinality $j$. Let

$$\mathbb{M}[\mu,s] = \bigcup_{\omega \in \mathcal{M}(s,p)} \left\{(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_p)' \in \mathbb{R}^{p\times k} \,:\, \boldsymbol{\theta}_i \in \mathcal{F}[\mu] \text{ if } i \in \omega,\ \boldsymbol{\theta}_i = \mathbf{0} \text{ if } i \notin \omega\right\} \tag{5}$$

be the class of all feasible matrix means. For a matrix $M \in \mathbb{M}[\mu,s]$, let $\mathbb{P}_M$ denote the joint law of $\{Y_{ij}\}_{i\in[p],j\in[k]}$. Since $\mathbb{P}_M$ is a product measure, we can write $\mathbb{P}_M = \otimes_{i\in[p]}\mathbb{P}_{\boldsymbol{\theta}_i}$. For a non-zero row $\boldsymbol{\theta}_i$, we set

$$\mathbb{P}_{\boldsymbol{\theta}_i}(A) = \int \mathcal{N}(A;\widehat{\boldsymbol{\theta}},\sigma^2\mathbf{I}_k)d\nu(\widehat{\boldsymbol{\theta}}), \qquad A \in \mathcal{B}(\mathbb{R}^k),$$

where $\nu$ is the distribution of the random variable $\sum_{j\in[k]}\mu_{ij}\xi_j e_j$ with $\xi_j \sim$ Bernoulli$(k^{-\beta})$ and $\{e_j\}_{j\in[k]}$ denoting the canonical basis of $\mathbb{R}^k$. For a zero row $\boldsymbol{\theta}_i = \mathbf{0}$, we set

$$\mathbb{P}_{\mathbf{0}}(A) = \mathcal{N}(A;\mathbf{0},\sigma^2\mathbf{I}_k), \qquad A \in \mathcal{B}(\mathbb{R}^k).$$

With this notation, we have the following result.

**Theorem 1** *Let*

$$\mu^2_{\min} = \mu^2_{\min}(n,k,p,s,\varepsilon,\beta) = \ln\left(1+u+\sqrt{2u+u^2}\right)\sigma^2$$

*where*

$$u = \frac{\ln\left(1+\frac{\alpha^2(p-s+1)}{2}\right)}{2k^{1-2\beta}}.$$

*If* $\alpha \in (0, \frac{1}{2})$ *and* $k^{-\beta}u < 1$*, then for all* $\mu \leq \mu_{\min}$*,*

$$\inf_{\widehat{\mu}} \sup_{M \in \mathbb{M}[\mu,s]} \mathbb{P}_M[S(\widehat{\mu}) \neq S(M)] \geq \frac{1}{2}(1-\alpha)$$

*where* $\mathbb{M}[\mu,s]$ *is given by* (5)*.*

The result can be interpreted in words in the following way: whatever the estimation procedure $\widehat{\mu}$, there exists some matrix $M \in \mathbb{M}[\mu_{\min}, s]$ such that the probability of incorrectly identifying the support $S(M)$ is bounded away from zero. In the next section, we will see that some estimation procedures achieve the lower bound given in Theorem 1.

## 3. Upper Bounds on the Support Recovery

In this section, we present sufficient conditions on $(n, p, k, \varepsilon, \mu_{\min})$ for different estimation procedures, so that

$$\mathbb{P}[S(\widehat{\boldsymbol{\mu}}) \neq S] \leq \alpha.$$

Let $\alpha', \delta' > 0$ be two parameters such that $\alpha' + \delta' = \alpha$. The parameter $\alpha'$ controls the probability of making a type one error

$$\mathbb{P}[\exists i \in [p] \ : \ i \in S(\widehat{\boldsymbol{\mu}}) \text{ and } i \notin S] \leq \alpha',$$

that is, the parameter $\alpha'$ upper bounds the probability that there is a zero row of the matrix $M$ that is estimated as a non-zero row. Likewise, the parameter $\delta'$ controls the probability of making a type two error

$$\mathbb{P}[\exists i \in [p] \ : \ i \notin S(\widehat{\boldsymbol{\mu}}) \text{ and } i \in S] \leq \delta',$$

that is, the parameter $\delta'$ upper bounds the probability that there is a non-zero row of the matrix $M$ that is estimated as a zero row.

The control of the type one and type two errors is established through the tuning parameter $\lambda$. It can be seen that if the parameter $\lambda$ is chosen such that, for all $i \in S$, it holds that $\mathbb{P}[i \notin S(\widehat{\boldsymbol{\mu}})] \leq \delta'/s$ and, for all $i \in S^c$, it hold that $\mathbb{P}[i \in S(\widehat{\boldsymbol{\mu}})] \leq \alpha'/(p-s)$, then using the union bound we have that $\mathbb{P}[S(\widehat{\boldsymbol{\mu}}) \neq S] \leq \alpha$. In the following subsections, we will use the outlined strategy to choose $\lambda$ for different estimation procedures.

### 3.1 Upper Bounds for the Lasso

Recall that the Lasso estimator is given as

$$\widehat{\boldsymbol{\mu}}^{\ell_1} = \underset{\boldsymbol{\mu} \in \mathbb{R}^{p \times k}}{\operatorname{argmin}} \ \frac{1}{2}||\mathbf{Y} - \boldsymbol{\mu}||_F^2 + \lambda||\boldsymbol{\mu}||_1.$$

It is easy to see that the solution of the above estimation problem is given as the following soft-thresholding operation

$$\widehat{\mu}_{ij}^{\ell_1} = \left(1 - \frac{\lambda}{|Y_{ij}|}\right)_+ Y_{ij}, \tag{6}$$

where $(x)_+ := \max(0,x)$. From (6), it is obvious that $i \in S(\widehat{\boldsymbol{\mu}}^{\ell_1})$ if and only if the maximum statistic, defined as

$$M_k(i) = \max_j |Y_{ij}|,$$

satisfies $M_k(i) \geq \lambda$. Therefore it is crucial to find the critical value of the parameter $\lambda$ such that

$$\begin{cases} \mathbb{P}[M_k(i) < \lambda] & < & \delta'/s & i \in S \\ \mathbb{P}[M_k(i) \geq \lambda] & < & \alpha'/(p-s) & i \in S^c. \end{cases}$$

We start by controlling the type one error. For $i \in S^c$ it holds that

$$\mathbb{P}[M_k(i) \geq \lambda] \leq k\mathbb{P}[|\mathcal{N}(0,\sigma^2)| \geq \lambda] \leq \frac{2k\sigma}{\sqrt{2\pi}\lambda} \exp\big(-\frac{\lambda^2}{2\sigma^2}\big) \tag{7}$$

using a standard tail bound for the Normal distribution. Setting the right hand side to $\alpha'/(p-s)$ in the above display, we obtain that $\lambda$ can be set as

$$\lambda = \sigma\sqrt{2\ln\frac{2k(p-s)}{\sqrt{2\pi}\alpha'}} \tag{8}$$

and (7) holds as soon as $2\ln\frac{2k(p-s)}{\sqrt{2\pi}\alpha'} \geq 1$. Next, we deal with the type two error. Let

$$\pi_k = \mathbb{P}[|(1-\varepsilon)\mathcal{N}(0,\sigma^2) + \varepsilon\mathcal{N}(\mu_{\min},\sigma^2)| > \lambda]. \tag{9}$$

Then for $i \in S$, $\mathbb{P}[M_k(i) < \lambda] \leq \mathbb{P}[\mathrm{Bin}(k,\pi_k) = 0]$, where $\mathrm{Bin}(k,\pi_k)$ denotes the binomial random variable with parameters $(k,\pi_k)$. Control of the type two error is going to be established through careful analysis of $\pi_k$ for various regimes of problem parameters.

**Theorem 2** *Let $\lambda$ be defined by* (8)*. Suppose $\mu_{\min}$ satisfies one of the following two cases:*

*(i)* $\mu_{\min} = \sigma\sqrt{2r\ln k}$ *where*

$$r > \left(\sqrt{1+C_{k,p,s}} - \sqrt{1-\beta}\right)^2$$

*with*

$$C_{k,p,s} = \frac{\ln\frac{2(p-s)}{\sqrt{2\pi}\alpha'}}{\ln k}$$

*and* $\lim_{n\to\infty} C_{k,p,s} \in [0,\infty)$*;*

*(ii)* $\mu_{\min} \geq \lambda$ *when*

$$\lim_{n\to\infty} \frac{\ln k}{\ln(p-s)} = 0$$

*and* $k^{1-\beta}/2 \geq \ln(s/\delta')$.

*Then*

$$\mathbb{P}[S(\widehat{\boldsymbol{\mu}}^{\ell_1}) \neq S] \leq \alpha.$$

The proof is given in Section 6.2. The two different cases describe two different regimes characterized by the ratio of $\ln k$ and $\ln(p-s)$.

Now we can compare the lower bound on $\mu_{\min}^2$ from Theorem 1 and the upper bound from Theorem 2. Without loss of generality we assume that $\sigma = 1$. We have that when $\beta < 1/2$ the lower bound is of the order $O\left(\ln\left(k^{\beta-1/2}\ln(p-s)\right)\right)$ and the upper bound is of the order $\ln(k(p-s))$. Ignoring the logarithmic terms in $p$ and $s$, we have that the lower bound is of the order $\widetilde{O}(k^{\beta-1/2})$ and the upper bound is of the order $\widetilde{O}(\ln k)$, which implies that the Lasso does not achieve the lower bound when the non-zero rows are dense. When the non-zero rows are sparse, $\beta > 1/2$, we have that both the lower and upper bound are of the order $\widetilde{O}(\ln k)$ (ignoring the terms depending on $p$ and $s$).

### 3.2 Upper Bounds for the Group Lasso

Recall that the group Lasso estimator is given as

$$\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_2} = \underset{\boldsymbol{\mu}\in\mathbb{R}^{p\times k}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{Y}-\boldsymbol{\mu}||_F^2 + \lambda\sum_{i\in[p]}||\boldsymbol{\theta}_i||_2,$$

where $\boldsymbol{\theta}_i = (\mu_{ij})_{j\in[k]}$. The group Lasso estimator can be obtained in a closed form as a result of the following thresholding operation (see, for example, Friedman et al., 2010)

$$\widehat{\boldsymbol{\theta}}_i^{\ell_1/\ell_2} = \left(1 - \frac{\lambda}{||Y_{i\cdot}||^2}\right)_+ Y_{i\cdot} \tag{10}$$

where $Y_{i\cdot}$ is the $i^{\text{th}}$ row of the data. From (10), it is obvious that $i \in S(\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_2})$ if and only if the statistic defined as

$$S_k(i) = \sum_j Y_{ij}^2,$$

satisfies $S_k(i) \geq \lambda$. The choice of $\lambda$ is crucial for the control of type one and type two errors. We use the following result, which directly follows from Theorem 2 in Baraud (2002).

**Lemma 3** *Let $\{Y_i = f_i + \sigma\xi_i\}_{i\in[n]}$ be a sequence of independent observations, where $f = \{f_i\}_{i\in[n]}$ is a sequence of numbers, $\xi_i \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\sigma$ is a known positive constant. Suppose that $t_{n,\alpha} \in \mathbb{R}$ satisfies $\mathbb{P}[\chi_n^2 > t_{n,\alpha}] \leq \alpha$. Let*

$$\phi_\alpha = I\{\sum_{i\in[n]} Y_i^2 \geq t_{n,\alpha}\sigma^2\}$$

*be a test for $f = 0$ versus $f \neq 0$. Then the test $\phi_\alpha$ satisfies*

$$\mathbb{P}[\phi_\alpha = 1] \leq \alpha$$

*when $f = 0$ and*

$$\mathbb{P}[\phi_\alpha = 0] \leq \delta$$

*for all $f$ such that*

$$||f||_2^2 \geq 2(\sqrt{5}+4)\sigma^2\ln\left(\frac{2e}{\alpha\delta}\right)\sqrt{n}.$$

**Proof** This follows immediately from Theorem 2 in Baraud (2002). ■

It follows directly from lemma 3 that setting

$$\lambda = t_{n,\alpha'/(p-s)}\sigma^2 \tag{11}$$

will control the probability of type one error at the desired level, that is,

$$\mathbb{P}[S_k(i) \geq \lambda] \leq \alpha'/(p-s), \qquad \forall i \in S^c.$$

The following theorem gives us the control of the type two error.

**Theorem 4** *Let* $\lambda = t_{n,\alpha'/(p-s)}\sigma^2$. *Then*

$$\mathbb{P}[S(\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_2}) \neq S] \leq \alpha$$

*if*

$$\mu_{\min} \geq \sigma\sqrt{2(\sqrt{5}+4)}\sqrt{\frac{k^{-1/2+\beta}}{1-c}}\sqrt{\ln\frac{2e(2s-\delta')(p-s)}{\alpha'\delta'}}$$

*where* $c = \sqrt{2\ln(2s/\delta')/k^{1-\beta}}$.

The proof is given in Section 6.3.

Using Theorem 1 and Theorem 4 we can compare the lower bound on $\mu_{\min}^2$ and the upper bound. Without loss of generality we assume that $\sigma = 1$. When each non-zero row is dense, that is, when $\beta < 1/2$, we have that both lower and upper bounds are of the order $\widetilde{O}(k^{\beta-1/2})$ (ignoring the logarithmic terms in $p$ and $s$). This suggest that the group Lasso performs better than the Lasso for the case where there is a lot of feature sharing between different tasks. Recall from previous section that the Lasso in this setting does not have the optimal dependence on $k$. However, when $\beta > 1/2$, that is, in the sparse non-zero row regime, we see that the lower bound is of the order $\widetilde{O}(\ln(k))$ whereas the upper bound is of the order $\widetilde{O}(k^{\beta-1/2})$. This implies that the group Lasso does not have optimal dependence on $k$ in the sparse non-zero row setting.

### 3.3 Upper Bounds for the Group Lasso with the Mixed $(\infty, 1)$ Norm

In this section, we analyze the group Lasso estimator with the mixed $(\infty, 1)$ norm, defined as

$$\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_\infty} = \underset{\boldsymbol{\mu}\in\mathbb{R}^{p\times k}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{Y}-\boldsymbol{\mu}||_F^2 + \lambda\sum_{i\in[p]}||\boldsymbol{\theta}_i||_\infty,$$

where $\boldsymbol{\theta}_i = (\mu_{ij})_{j\in[k]}$. The closed form solution for $\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_\infty}$ can be obtained (see Liu et al., 2009), however, we are only going to use the following lemma.

**Lemma 5** *(Liu et al., 2009)* $\widehat{\boldsymbol{\theta}}_i^{\ell_1/\ell_\infty} = \mathbf{0}$ *if and only if* $\sum_j |Y_{ij}| \leq \lambda$.

**Proof** See the proof of Proposition 5 in Liu et al. (2009). ■

Suppose that the penalty parameter $\lambda$ is set as

$$\lambda = k\sigma\sqrt{2\ln\frac{k(p-s)}{\alpha'}}. \tag{12}$$

It follows immediately using a tail bound for the Normal distribution that

$$\mathbb{P}[\sum_j |Y_{ij}| \geq \lambda] \leq k \max_j \mathbb{P}[|Y_{ij}| \geq \lambda/k] \leq \alpha'/(p-s), \qquad \forall i \in S^c,$$

which implies that the probability of the type one error is controlled at the desired level.

**Theorem 6** *Let the penalty parameter $\lambda$ be defined by* (12)*. Then*

$$\mathbb{P}[S(\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_\infty}) \neq S] \leq \alpha$$

*if*

$$\mu_{\min} \;\geq\; \frac{1+\tau}{1-c} k^{-1+\beta} \lambda$$

*where* $c = \sqrt{2\ln(2s/\delta')/k^{1-\beta}}$ *and* $\tau = \sigma\sqrt{2k\ln\frac{2s-\delta'}{\delta'}}/\lambda$.

The proof is given in Section 6.4.

Comparing upper bounds for the Lasso and the group Lasso with the mixed $(2,1)$ norm with the result of Theorem 6, we can see that both the Lasso and the group Lasso have better dependence on $k$ than the group Lasso with the mixed $(\infty,1)$ norm. The difference becomes more pronounced as $\beta$ increases. This suggest that we should be very cautious when using the group Lasso with the mixed $(\infty,1)$ norm, since as soon as the tasks do not share exactly the same features, the other two procedures have much better performance on identifying the set of non-zero rows.

## 4. Simulation Results

We conduct a small-scale empirical study of the performance of the Lasso and the group Lasso (both with the mixed $(2,1)$ norm and with the mixed $(\infty,1)$ norm). Our empirical study shows that the theoretical findings of Section 3 describe sharply the behavior of procedures even for small sample studies. In particular, we demonstrate that as the minimum signal level $\mu_{\min}$ varies in the model (3), our theory sharply determines points at which probability of identifying non-zero rows of matrix $M$ successfully transitions from 0 to 1 for different procedures.

The simulation procedure can be described as follows. Without loss of generality we let $S = [s]$ and draw the samples $\{Y_{ij}\}_{i\in[p],j\in[k]}$ according to the model in (3). The total number of rows $p$ is varied in $\{128, 256, 512, 1024\}$ and the number of columns is set to $k = \lfloor p\log_2(p)\rfloor$. The sparsity of each non-zero row is controlled by changing the parameter $\beta$ in $\{0, 0.25, 0.5, 0.75\}$ and setting $\varepsilon = k^{-\beta}$. The number of non-zero rows is set to $s = \lfloor \log_2(p)\rfloor$, the sample size is set to $n = 0.1p$ and $\sigma_0 = 1$. The parameters $\alpha'$ and $\delta'$ are both set to 0.01. For each setting of the parameters, we report our results averaged over 1000 simulation runs. Simulations with other choices of parameters $n, s$ and $k$ have been tried out, but the results were qualitatively similar and, hence, we do not report them here.

The regularization parameter $\lambda$ is chosen according to Equations (8), (11) and (12), which assume that the noise level $\sigma_0$ is known. In practice, estimating the standard deviation of the noise in high-dimensions is a hard problem and practitioners often use cross-validation as a data-driven way to choose the penalty parameter. For recent work on data-driven tuning of the penalty parameters, we refer the reader to Arlot and Bach (2009).

### 4.1 Lasso

We investigate the performance on the Lasso for the purpose of estimating the set of non-zero rows, $S$. Figure 1 plots the probability of success as a function of the signal strength. On the same figure we plot the probability of success for the group Lasso with both $(2,1)$ and $(\infty,1)$-mixed norms. Using theorem 2, we set

$$\mu_{\text{lasso}} = \sqrt{2(r+0.001)\ln k} \tag{13}$$

where $r$ is defined in theorem 2. Next, we generate data according to (3) with all elements $\{\mu_{ij}\}$ set to $\mu = \rho\mu_{\text{lasso}}$, where $\rho \in [0.05, 2]$. The penalty parameter $\lambda$ is chosen as in (8). Figure 1 plots probability of success as a function of the parameter $\rho$, which controls the signal strength. This probability transitions very sharply from 0 to 1. A rectangle on a horizontal line represents points at which the probability $\mathbb{P}[\widehat{S} = S]$ is between 0.05 and 0.95. From each subfigure in Figure 1, we can observe that the probability of success for the Lasso transitions from 0 to 1 for the same value of the parameter $\rho$ for different values of $p$, which indicates that, except for constants, our theory correctly characterizes the scaling of $\mu_{\min}$. In addition, we can see that the Lasso outperforms the group Lasso (with $(2,1)$-mixed norm) when each non-zero row is very sparse (the parameter $\beta$ is close to one).

### 4.2 Group Lasso

Next, we focus on the empirical performance of the group Lasso with the mixed $(2,1)$ norm. Figure 2 plots the probability of success as a function of the signal strength. Using theorem 4, we set

$$\mu_{\text{group}} = \sigma\sqrt{2(\sqrt{5}+4)}\sqrt{\frac{k^{-1/2+\beta}}{1-c}}\sqrt{\ln\frac{(2s-\delta')(p-s)}{\alpha'\delta'}} \tag{14}$$

where $c$ is defined in theorem 4. Next, we generate data according to (3) with all elements $\{\mu_{ij}\}$ set to $\mu = \rho\mu_{\text{group}}$, where $\rho \in [0.05, 2]$. The penalty parameter $\lambda$ is given by (11). Figure 2 plots probability of success as a function of the parameter $\rho$, which controls the signal strength. A rectangle on a horizontal line represents points at which the probability $\mathbb{P}[\widehat{S} = S]$ is between 0.05 and 0.95. From each subfigure in Figure 2, we can observe that the probability of success for the group Lasso transitions from 0 to 1 for the same value of the parameter $\rho$ for different values of $p$, which indicated that, except for constants, our theory correctly characterizes the scaling of $\mu_{\min}$. We observe also that the group Lasso outperforms the Lasso when each non-zero row is not too sparse, that is, when there is a considerable overlap of features between different tasks.

### 4.3 Group Lasso with the Mixed $(\infty,1)$ Norm

Next, we focus on the empirical performance of the group Lasso with the mixed $(\infty,1)$ norm. Figure 3 plots the probability of success as a function of the signal strength. Using theorem 6, we set

$$\mu_{\text{infty}} = \frac{1+\tau}{1-c}k^{-1+\beta}\lambda \tag{15}$$

where $\tau$ and $c$ are defined in theorem 6 and $\lambda$ is given by (12). Next, we generate data according to (3) with all elements $\{\mu_{ij}\}$ set to $\mu = \rho\mu_{\text{infty}}$, where $\rho \in [0.05, 2]$. Figure 3 plots probability of success as a function of the parameter $\rho$, which controls the signal strength. A rectangle on a horizontal line represents points at which the probability $\mathbb{P}[\widehat{S} = S]$ is between 0.05 and 0.95. From each subfigure

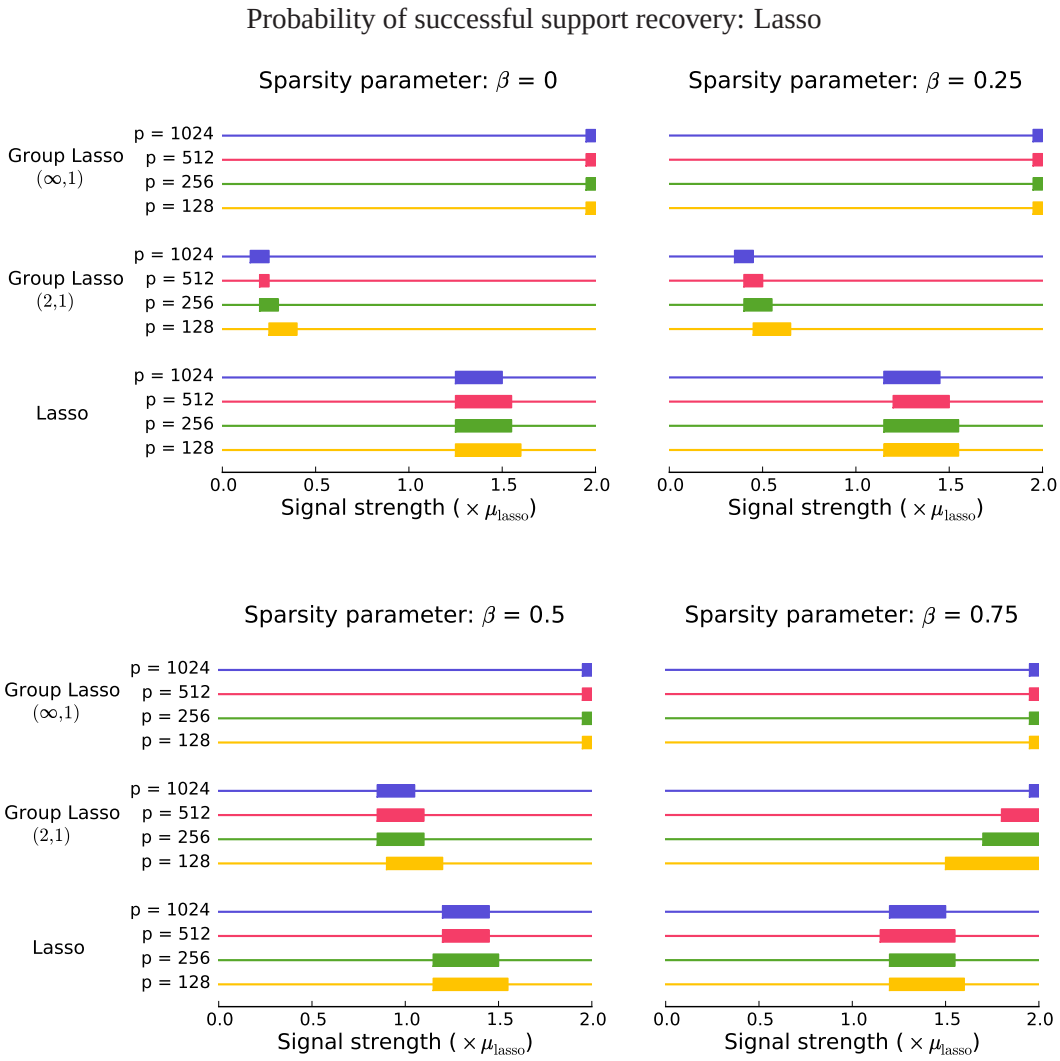


Figure 1: The probability of success for the Lasso for the problem of estimating $S$ plotted against the signal strength, which is varied as a multiple of $\mu_{\text{lasso}}$ defined in (13). A rectangle on each horizontal line represents points at which the probability $\mathbb{P}[\widehat{S} = S]$ is between 0.05 and 0.95. To the left of the rectangle the probability is smaller than 0.05, while to the right the probability is larger than 0.95. Different subplots represent the probability of success as the sparsity parameter β changes.

in Figure 3, we can observe that the probability of success for the group Lasso transitions from 0 to 1 for the same value of the parameter ρ for different values of $p$, which indicated that, except for constants, our theory correctly characterizes the scaling of $\mu_{\text{min}}$. We also observe that the group Lasso with the mixed $(\infty, 1)$ norm never outperforms the Lasso or the group Lasso with the mixed $(2, 1)$ norm.

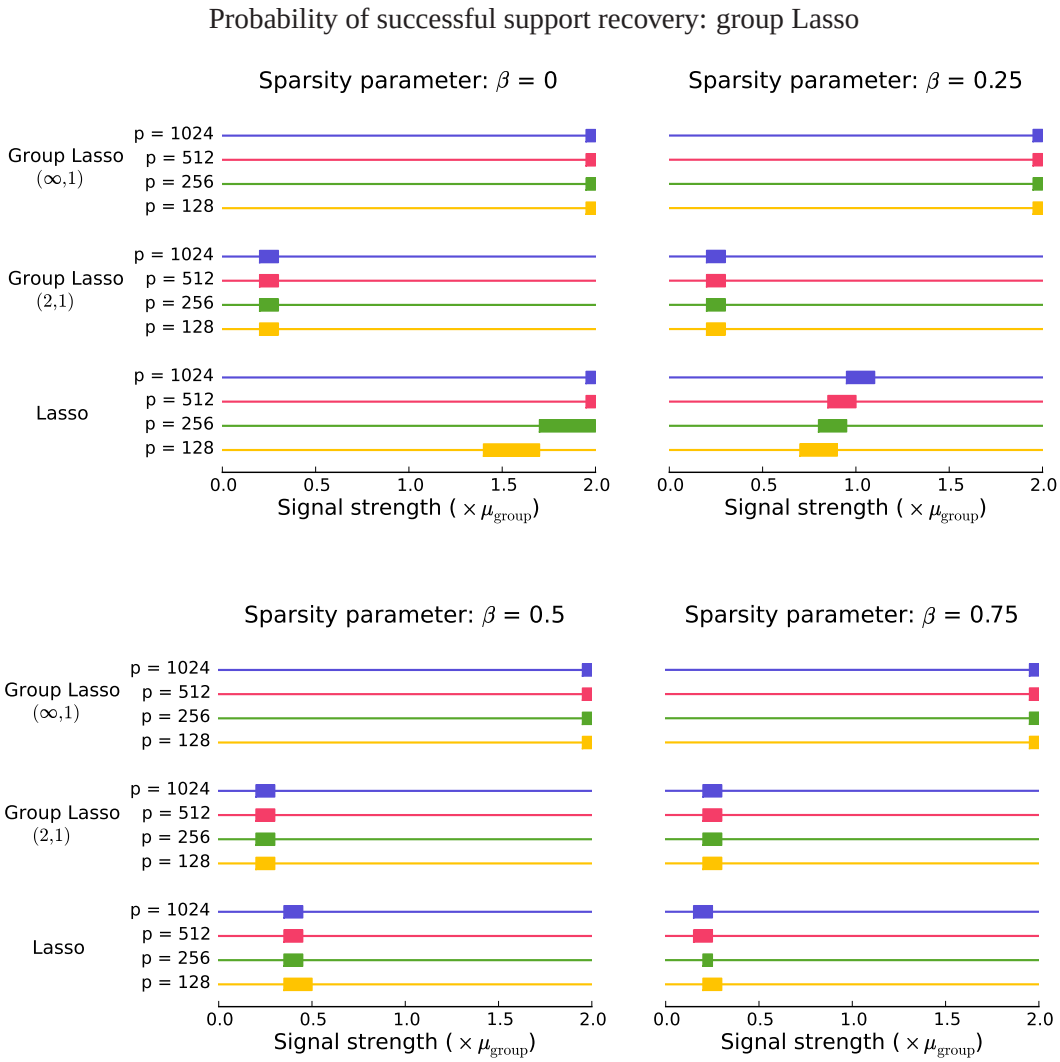


Figure 2: The probability of success for the group Lasso for the problem of estimating $S$ plotted against the signal strength, which is varied as a multiple of $\mu_{\text{group}}$ defined in (14). A rectangle on each horizontal line represents points at which the probability $\mathbb{P}[\widehat{S} = S]$ is between 0.05 and 0.95. To the left of the rectangle the probability is smaller than 0.05, while to the right the probability is larger than 0.95. Different subplots represent the probability of success as the sparsity parameter β changes.

## 5. Discussion

We have studied the benefits of task sharing in sparse problems. Under many scenarios, the group lasso outperforms the lasso. The $\ell_1/\ell_2$ penalty seems to be a much better choice for the group lasso than the $\ell_1/\ell_\infty$. However, as pointed out to us by Han Liu, for screening, where false discoveries are less important than accurate recovery, it is possible that the $\ell_1/\ell_\infty$ penalty could be useful. From the results in Section 3, we can further conclude that the Lasso procedure performs better than the group Lasso when each non-zero row is sparse, while the group Lasso (with the mixed $(2,1)$ norm)

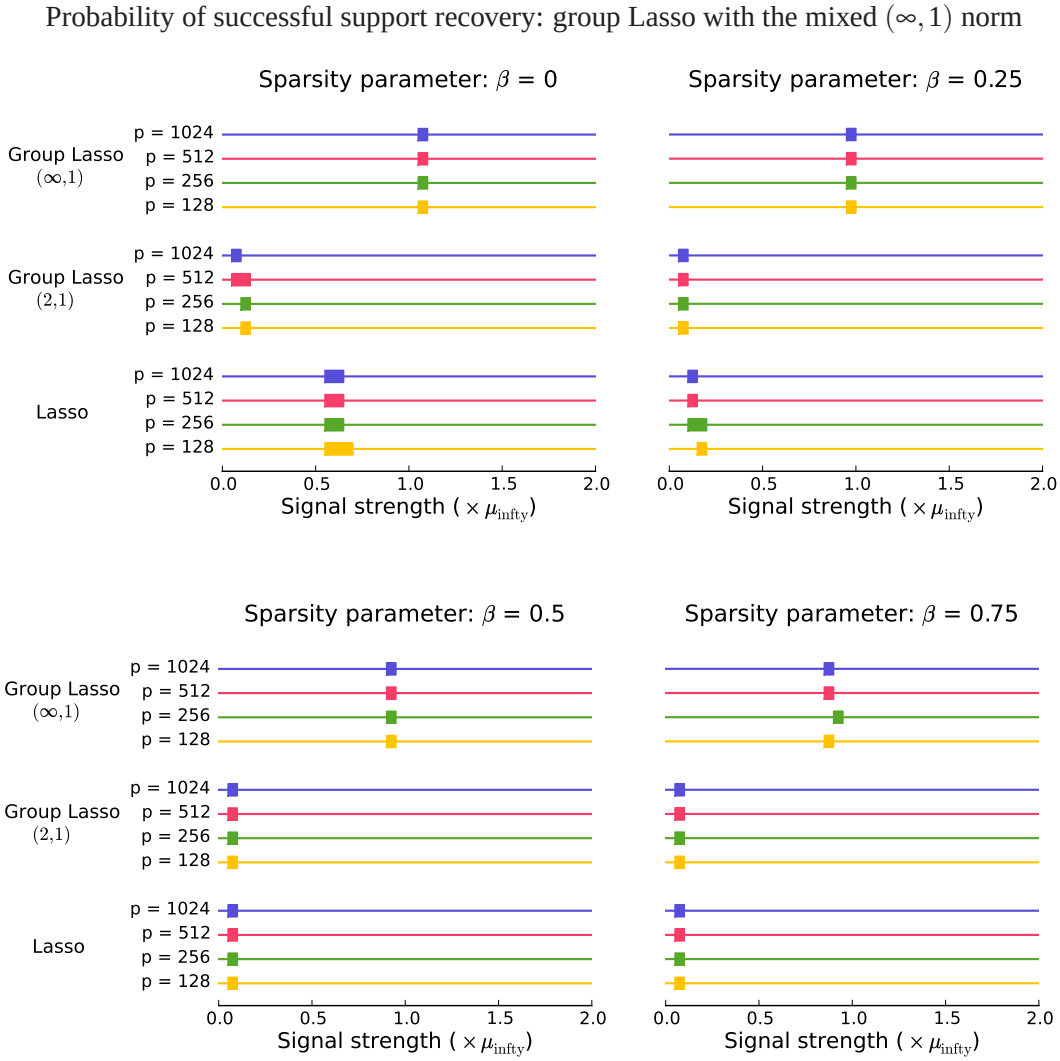


Figure 3: The probability of success for the group Lasso with mixed $(\infty, 1)$ norm for the problem of estimating $S$ plotted against the signal strength, which is varied as a multiple of $\mu_{\text{infty}}$ defined in (15). A rectangle on each horizontal line represents points at which the probability $\mathbb{P}[\widehat{S} = S]$ is between 0.05 and 0.95. To the left of the rectangle the probability is smaller than 0.05, while to the right the probability is larger than 0.95. Different subplots represent the probability of success as the sparsity parameter β changes.

performs better when each non-zero row is dense. Since in many practical situations one does not know how much overlap there is between different tasks, it would be useful to combine the Lasso and the group Lasso in order to improve the performance. For example, one can take the union of the Lasso and the group Lasso estimate, $\widehat{S} = S(\widehat{\boldsymbol{\mu}}^{\ell_1}) \cup S(\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_2})$. The suggested approach has the advantage that one does not need to know in advance which estimation procedure to use. While such a combination can be justified in the Normal means problem as a way to increase the power to detect the non-zero rows, it is not clear whether the same approach can be justified in the multi-task regression model (1).

The analysis of the Normal means model in (3) provides insights into the theoretical results we could expect in the conventional multi-task learning given in (1). However, there is no direct way to translate our results into valid results for the model in (1); a separate analysis needs to be done in order to establish sharp theoretical results.

## 6. Proofs

This section collects technical proofs of the results presented in the paper. Throughout the section we use $c_1, c_2, \ldots$ to denote positive constants whose value may change from line to line.

### 6.1 Proof of Theorem 1

Without loss of generality, we may assume $\sigma = 1$. Let $\phi(u)$ be the density of $\mathcal{N}(0,1)$ and define $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_{\mathbf{1}}$ to be two probability measures on $\mathbb{R}^k$ with the densities with respect to the Lebesgue measure given as

$$f_0(a_1,\ldots,a_k) = \prod_{j\in[k]} \phi(a_j) \tag{16}$$

and

$$f_1(a_1,\ldots,a_k) = \mathbb{E}_Z \mathbb{E}_m \mathbb{E}_\xi \prod_{j\in m} \phi(a_j - \xi_j \mu_{\min}) \prod_{j\notin m} \phi(a_j) \tag{17}$$

where $Z \sim \mathrm{Bin}(k, k^{-\beta})$, $m$ is a random variable uniformly distributed over $\mathcal{M}(Z,k)$ and $\{\xi_j\}_{j\in[k]}$ is a sequence of Rademacher random variables, independent of $Z$ and $m$. A Rademacher random variable takes values $\pm 1$ with probability $\frac{1}{2}$.

To simplify the discussion, suppose that $p-s+1$ is divisible by 2. Let $T = (p-s+1)/2$. Using $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_{\mathbf{1}}$, we construct the following three measures,

$$\widetilde{\mathbb{Q}} = \mathbb{P}_{\mathbf{1}}^{s-1} \otimes \mathbb{P}_{\mathbf{0}}^{p-s+1},$$

$$\mathbb{Q}_0 = \frac{1}{T} \sum_{\substack{j\in\{s,\ldots,p\} \\ j \text{ odd}}} \mathbb{P}_{\mathbf{1}}^{s-1} \otimes \mathbb{P}_{\mathbf{0}}^{j-s} \otimes \mathbb{P}_{\mathbf{1}} \otimes \mathbb{P}_{\mathbf{0}}^{p-j}$$

and

$$\mathbb{Q}_1 = \frac{1}{T} \sum_{\substack{j\in\{s,\ldots,p\} \\ j \text{ even}}} \mathbb{P}_{\mathbf{1}}^{s-1} \otimes \mathbb{P}_{\mathbf{0}}^{j-s} \otimes \mathbb{P}_{\mathbf{1}} \otimes \mathbb{P}_{\mathbf{0}}^{p-j}.$$

It holds that

$$\begin{aligned} \inf_{\widehat{\mu}} \sup_{M\in\mathbb{M}} \mathbb{P}_M[S(M) \neq S(\widehat{\mu})] &\geq \inf_{\Psi} \max \Big( \mathbb{Q}_0(\Psi = 1), \mathbb{Q}_1(\Psi = 0) \Big) \\ &\geq \frac{1}{2} - \frac{1}{2} ||\mathbb{Q}_0 - \mathbb{Q}_1||_1, \end{aligned}$$

where the infimum is taken over all tests $\Psi$ taking values in $\{0,1\}$ and $||\cdot||_1$ is the total variation distance between probability measures. For a readable introduction on lower bounds on the minimax probability of error, see Section 2 in Tsybakov (2009). In particular, our approach is related to the one described in Section 2.7.4. We proceed by upper bounding the total variation distance between

$\mathbb{Q}_0$ and $\mathbb{Q}_1$. Let $g = d\mathbb{P}_{\mathbf{1}}/d\mathbb{P}_{\mathbf{0}}$ and let $u_i \in \mathbb{R}^k$ for each $i \in [p]$, then

$$\begin{aligned}
&\frac{d\mathbb{Q}_0}{d\widetilde{\mathbb{Q}}}(u_1,\ldots,u_p) \\
&\quad= \frac{1}{T}\sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ even}}} \prod_{i\in\{1,\ldots,s-1\}} \frac{d\mathbb{P}_{\mathbf{1}}}{d\mathbb{P}_{\mathbf{1}}}(u_i) \prod_{i\in\{s,\ldots,j-1\}} \frac{d\mathbb{P}_{\mathbf{0}}}{d\mathbb{P}_{\mathbf{0}}}(u_i) \frac{d\mathbb{P}_{\mathbf{1}}}{d\mathbb{P}_{\mathbf{0}}}(u_j) \prod_{i\in\{j+1,\ldots,p\}} \frac{d\mathbb{P}_{\mathbf{0}}}{d\mathbb{P}_{\mathbf{0}}}(u_i) \\
&\quad= \frac{1}{T}\sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ even}}} g(u_j)
\end{aligned}$$

and, similarly, we can compute $d\mathbb{Q}_1/d\widetilde{\mathbb{Q}}$. The following holds

$$\begin{aligned}
&\|\mathbb{Q}_0 - \mathbb{Q}_1\|_1^2 \\
&\quad= \left( \int \Big| \frac{1}{T} \Big( \sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ even}}} g(u_j) - \sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ odd}}} g(u_j) \Big) \Big| \prod_{i\in\{s,\ldots,p\}} d\mathbb{P}_{\mathbf{0}}(u_i) \right)^2 \\
&\quad\le \frac{1}{T^2} \int \Big( \sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ even}}} g(u_j) - \sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ odd}}} g(u_j) \Big)^2 \prod_{i\in\{s,\ldots,p\}} d\mathbb{P}_{\mathbf{0}}(u_i) \\
&\quad= \frac{2}{T}\big(\mathbb{P}_{\mathbf{0}}(g^2) - 1\big),
\end{aligned} \tag{18}$$

where the last equality follows by observing that

$$\int \sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ even}}} \sum_{\substack{j'\in\{s,\ldots,p\}\\ j' \text{ even}}} g(u_j)g(u_{j'}) \prod_{\substack{i\in\{s,\ldots,p\}\\ i \text{ even}}} d\mathbb{P}_{\mathbf{0}}(u_i) = T\ \mathbb{P}_{\mathbf{0}}(g^2) + T^2 - T$$

and

$$\int \sum_{\substack{j\in\{s,\ldots,p\}\\ j \text{ even}}} \sum_{\substack{j'\in\{s,\ldots,p\}\\ j' \text{ odd}}} g(u_j)g(u_{j'}) \prod_{i\in\{s,\ldots,p\}} d\mathbb{P}_{\mathbf{0}}(u_i) = T^2.$$

Next, we proceed to upper bound $\mathbb{P}_{\mathbf{0}}(g^2)$, using some ideas presented in the proof of Theorem 1 in Baraud (2002). Recall definitions of $f_0$ and $f_1$ in (16) and (17) respectively. Then $g = d\mathbb{P}_{\mathbf{1}}/d\mathbb{P}_{\mathbf{0}} = f_1/f_0$ and we have

$$\begin{aligned}
g(a_1,\ldots,a_k) &= \mathbb{E}_Z \mathbb{E}_m \mathbb{E}_\xi \left[ \exp\Big( -\frac{Z\mu_{\min}^2}{2} + \mu_{\min} \sum_{j\in m} \xi_j a_j \Big) \right] \\
&= \mathbb{E}_Z \left[ \exp\Big( -\frac{Z\mu_{\min}^2}{2} \Big)\ \mathbb{E}_m \Big[ \prod_{j\in m} \cosh(\mu_{\min} a_j) \Big] \right].
\end{aligned}$$

Furthermore, let $Z' \sim \mathrm{Bin}(k, k^{-\beta})$ be independent of $Z$ and $m'$ uniformly distributed over $\mathcal{M}(Z', k)$. The following holds

$$\begin{aligned}
&\mathbb{P}_{\mathbf{0}}(g^2) \\
&= \mathbb{P}_{\mathbf{0}}\Big(\mathbb{E}_{Z',Z}\Big[\exp\Big(-\frac{(Z+Z')\mu_{\min}^2}{2}\Big)\,\mathbb{E}_{m,m'}\prod_{j\in m}\cosh(\mu_{\min}a_j)\prod_{j\in m'}\cosh(\mu_{\min}a_j)\Big]\Big) \\
&= \mathbb{E}_{Z',Z}\Big[\exp\Big(-\frac{(Z+Z')\mu_{\min}^2}{2}\Big) \\
&\qquad\qquad \mathbb{E}_{m,m'}\Big[\prod_{j\in m\cap m'}\int\cosh^2(\mu_{\min}a_j)\phi(a_j)da_j \\
&\qquad\qquad\qquad \prod_{j\in m\triangle m'}\int\cosh(\mu_{\min}a_j)\phi(a_j)da_j\Big]\Big],
\end{aligned}$$

where we use $m\triangle m'$ to denote $(m\cup m')\backslash(m\cap m')$. By direct calculation, we have that

$$\int\cosh^2(\mu_{\min}a_j)\phi(a_j)da_j = \exp(\mu_{\min}^2)\cosh(\mu_{\min}^2)$$

and

$$\int\cosh(\mu_{\min}a_j)\phi(a_j)da_j = \exp(\mu_{\min}^2/2).$$

Since $\frac{1}{2}|m\triangle m'| + |m\cap m'| = (Z+Z')/2$, we have that

$$\begin{aligned}
\mathbb{P}_{\mathbf{0}}(g^2) &= \mathbb{E}_{Z,Z'}\Big[E_{m,m'}\Big[\big(\cosh(\mu_{\min}^2)\big)^{|m\cap m'|}\Big]\Big] \\
&= \mathbb{E}_{Z,Z'}\Big[\sum_{j=0}^{k}p_j\big(\cosh(\mu_{\min}^2)\big)^j\Big] \\
&= \mathbb{E}_{Z,Z'}\Big[\mathbb{E}_X\Big[\cosh(\mu_{\min}^2)^X\Big]\Big],
\end{aligned}$$

where

$$p_j = \begin{cases} 0 & \text{if } j < Z+Z'-k \text{ or } j > \min(Z,Z') \\ \frac{\binom{Z'}{j}\binom{k-Z'}{Z-j}}{\binom{k}{Z}} & \text{otherwise} \end{cases}$$

and $P[X=j] = p_j$. Therefore, $X$ follows a hypergeometric distribution with parameters $k$, $Z$, $Z'/k$. [The first parameter denotes the total number of stones in an urn, the second parameter denotes the number of stones we are going to sample without replacement from the urn and the last parameter denotes the fraction of white stones in the urn.] Then following (Aldous, 1985, p. 173; see also Baraud 2002), we know that $X$ has the same distribution as the random variable $\mathbb{E}[\widetilde{X}|\mathcal{T}]$ where $\widetilde{X}$ is a binomial random variable with parameters $Z$ and $Z'/k$, and $\mathcal{T}$ is a suitable $\sigma$-algebra. By convexity, it follows that

$$\begin{aligned}
\mathbb{P}_{\mathbf{0}}(g^2) &\le \mathbb{E}_{Z,Z'}\Big[\mathbb{E}_{\widetilde{X}}\Big[\cosh(\mu_{\min}^2)^{\widetilde{X}}\Big]\Big] \\
&= \mathbb{E}_{Z,Z'}\left[\exp\left(Z\ln\left(1+\frac{Z'}{k}\big(\cosh(\mu_{\min}^2)-1\big)\right)\right)\right] \\
&= \mathbb{E}_{Z'}\mathbb{E}_Z\left[\exp\left(Z\ln\left(1+\frac{Z'}{k}u\right)\right)\right]
\end{aligned}$$

where $\mu_{\min}^2 = \ln(1+u+\sqrt{2u+u^2})$ with

$$u = \frac{\ln\left(1+\frac{\alpha^2 T}{2}\right)}{2k^{1-2\beta}}.$$

Continuing with our calculations, we have that

$$\begin{aligned}
\mathbb{P}_{\mathbf{0}}(g^2) &= \mathbb{E}_{Z'} \exp\left(k \ln\left(1+k^{-(1+\beta)} u Z'\right)\right) \\
&\leq \mathbb{E}_{Z'} \exp\left(k^{-\beta} u Z'\right) \\
&= \exp\left(k \ln\left(1+k^{-\beta}\left(\exp(k^{-\beta}u)-1\right)\right)\right) \\
&\leq \exp\left(k^{1-\beta}\left(\exp\left(k^{-\beta}u\right)-1\right)\right) \\
&\leq \exp\left(2k^{1-2\beta}u\right) \\
&= 1+\frac{\alpha^2 T}{2},
\end{aligned} \tag{19}$$

where the last inequality follows since $k^{-\beta}u < 1$ for all large $p$. Combining (19) with (18), we have that

$$\|\mathbb{Q}_0 - \mathbb{Q}_1\|_1 \leq \alpha,$$

which implies that

$$\inf_{\widehat{\mu}} \sup_{M \in \mathbb{M}} \mathbb{P}_M[S(M) \neq S(\widehat{\mu})] \geq \frac{1}{2} - \frac{1}{2}\alpha.$$

### 6.2 Proof of Theorem 2

Without loss of generality, we can assume that $\sigma = 1$ and rescale the final result. For $\lambda$ given in (8), it holds that $\mathbb{P}[|\mathcal{N}(0,1) \geq \lambda] = o(1)$. For the probability defined in (9), we have the following lower bound

$$\pi_k = (1-\varepsilon)\mathbb{P}[|\mathcal{N}(0,1)| \geq \lambda] + \varepsilon\mathbb{P}[|\mathcal{N}(\mu_{\min},1)| \geq \lambda] \geq \varepsilon\mathbb{P}[\mathcal{N}(\mu_{\min},1) \geq \lambda].$$

We prove the two cases separately.
**Case 1:** *Large number of tasks.* By direct calculation

$$\pi_k \geq \varepsilon\mathbb{P}[\mathcal{N}(\mu_{\min},1) \geq \lambda] = \frac{1}{\sqrt{4\pi \log k}\left(\sqrt{1+C_{k,p,s}}-\sqrt{r}\right)} k^{-\beta-\left(\sqrt{1+C_{k,p,s}}-\sqrt{r}\right)^2} =: \underline{\pi_k}.$$

Since $1-\beta > \left(\sqrt{1+C_{k,p,s}}-\sqrt{r}\right)^2$, we have that $\mathbb{P}[\mathrm{Bin}(k,\pi_k)=0] \xrightarrow{n\to\infty} 0$. We can conclude that as soon as $k\underline{\pi_k} \geq \ln(s/\delta')$, it holds that $\mathbb{P}[S(\widehat{\mu}^{\ell_1}) \neq S] \leq \alpha$.
**Case 2:** *Medium number of tasks.* When $\mu_{\min} \geq \lambda$, it holds that

$$\pi_k \geq \varepsilon\mathbb{P}[\mathcal{N}(\mu_{\min},1) \geq \lambda] \geq \frac{k^{-\beta}}{2}.$$

We can conclude that as soon as $k^{1-\beta}/2 \geq \ln(s/\delta')$, it holds that $\mathbb{P}[S(\widehat{\mu}^{\ell_1}) \neq S] \leq \alpha$.

### 6.3 Proof of Theorem 4

Using a Chernoff bound, $\mathbb{P}[\mathrm{Bin}(k,k^{-\beta}) \leq (1-c)k^{1-\beta}] \leq \delta'/2s$ for $c = \sqrt{2\ln(2s/\delta')/k^{1-\beta}}$. For $i \in S$, we have that

$$\mathbb{P}[S_k(i) \leq \lambda] \leq \frac{\delta'}{2s} + \Big(1 - \frac{\delta'}{2s}\Big)\mathbb{P}\Big[S_k(i) \leq \lambda \;\Big|\; \Big\{||\theta_i||_2^2 \geq (1-c)k^{1-\beta}\mu_{\min}^2\Big\}\Big].$$

Therefore, using lemma 3 with $\delta = \delta'/(2s-\delta')$, if follows that $\mathbb{P}[S_k(i) \leq \lambda] \leq \delta'/(2s)$ for all $i \in S$ when

$$\mu_{\min} \geq \sigma\sqrt{2(\sqrt{5}+4)}\sqrt{\frac{k^{-1/2+\beta}}{1-c}}\sqrt{\ln\frac{2e(2s-\delta')(p-s)}{\alpha'\delta'}}.$$

Since $\lambda = t_{n,\alpha'/(p-s)}\sigma^2$, $\mathbb{P}[S_k(i) \geq \lambda] \leq \alpha'/(p-s)$ for all $i \in S^c$. We can conclude that $\mathbb{P}[S(\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_2}) \neq S] \leq \alpha$.

### 6.4 Proof of Theorem 6

Without loss of generality, we can assume that $\sigma = 1$. Proceeding as in the proof of theorem 4, $\mathbb{P}[\mathrm{Bin}(k,k^{-\beta}) \leq (1-c)k^{1-\beta}] \leq \delta'/2s$ for $c = \sqrt{2\ln(2s/\delta')/k^{1-\beta}}$. Then for $i \in S$ it holds that

$$\mathbb{P}[\sum_j |Y_{ij}| \leq \lambda] \leq \frac{\delta'}{2s} + \Big(1 - \frac{\delta'}{2s}\Big)\mathbb{P}[(1-c)k^{1-\beta}\mu_{\min} + z_k \leq \lambda],$$

where $z_k \sim \mathcal{N}(0,k)$. Since $(1-c)k^{1-\beta}\mu_{\min} \geq (1+\tau)\lambda$, the right-hand side of the above display can upper bounded as

$$\frac{\delta'}{2s} + \Big(1 - \frac{\delta'}{2s}\Big)\mathbb{P}[\mathcal{N}(0,1) \geq \tau\lambda/\sqrt{k}] \leq \frac{\delta'}{2s} + \Big(1 - \frac{\delta'}{2s}\Big)\frac{\delta'}{2s-\delta'} \leq \frac{\delta'}{s}.$$

The above display gives us the desired control of the type two error, and we can conclude that $\mathbb{P}[S(\widehat{\boldsymbol{\mu}}^{\ell_1/\ell_\infty}) \neq S] \leq \alpha$.

## Acknowledgments

We would like to thank Han Liu for useful discussions. We also thank two anonymous reviewers and the associate editor for comments that have helped improve the paper. The research reported here was supported in part by NSF grant CCF-0625879, AFOSR contract FA9550-09-1-0373, a grant from Google, and a graduate fellowship from Facebook.

## References

David Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII 1983*, pages 1–198. Springer, Berlin, 1985.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems 22*, pages 46–54. 2009.

Yannick Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5): 577–606, 2002.

Larry Brown and Mark Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, 96:1348–1360, 2001.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. Technical report, Stanford, 2010. Available at arXiv:1001.0736.

Seyoung Kim, Kyung-Ah Sohn, and Eric P. Xing. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12):i204–212, 2009.

Mladen Kolar and Eric P. Xing. Ultra-high dimensional multiple output learning with simultaneous orthogonal matching pursuit: Screening approach. In *AISTATS '10: Proc. 13th Int'l Conf. on Artifical Intelligence and Statistics*, pages 413–420, 2010.

Han Liu, Mark Palatucci, and Jian Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML '09: Proc. 26th Int'l Conf. on Machine Learning*, pages 649–656, New York, NY, USA, 2009.

Karim Lounici, Massimiliano Pontil, Alexandre Tsybakov, and Sara van de Geer. Taking advantage of sparsity in Multi-Task learning. In *COLT '09: Proc. Conf. on Learning Theory*, 2009.

Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara van de Geer. Oracle inequalities and optimal inference under group sparsity. *Preprint*, 2010. Available at arXiv:1007.1771.

Sahand Negahban and Martin Wainwright. Phase transitions for high-dimensional joint support recovery. In *Advances in Neural Information Processing Systems 21*, pages 1161–1168, 2009.

Michael Nussbaum. Asymptotic equivalence of density estimation and gaussian white noise. *Ann. Statist.*, 24(6):2399–2430, 1996.

Guillaume Obozinski, Martin Wainwright, and Michael Jordan. Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, 39(1):1–47, 2011.

Alexandre Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Berwin Turlach, William Venables, and Stephen Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005. ISSN 0040-1706.

Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Elec. J. Statist.*, 3:1360–1392, 2009.

Jian Zhang. *A Probabilistic Framework for Multitask Learning*. PhD thesis, Carnegie Mellon University, 2006.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006. ISSN 1533-7928.

Hui Zou and Ming Yuan. The $F_\infty$-norm support vector machine. *Stat. Sin*, 18:379–398, 2008.