Block Regularized Lasso for Multivariate Multi-Response Linear Regression

Weiguang Wang
Syracuse University

Yingbin Liang
Syracuse University

Eric P. Xing
Carnegie Mellon University

Abstract

The multivariate multi-response (MVMR) linear regression problem is investigated, in which design matrices are Gaussian with covariance matrices $\Sigma^{(1:K)} = \left(\Sigma^{(1)}, \ldots, \Sigma^{(K)}\right)$ for K linear regressions. The support union of K p-dimensional regression vectors (collected as columns of matrix B^*) are recovered using l_1/l_2-regularized Lasso. Sufficient and necessary conditions to guarantee successful recovery of the support union are characterized via a threshold. More specifically, it is shown that under certain conditions on the distributions of design matrices, if $n > c_{p1}\psi(B^*, \Sigma^{(1:K)})\log(p-s)$ where c_{p1} is a constant, and s is the size of the support set, then l_1/l_2-regularized Lasso correctly recovers the support union; and if $n < c_{p2}\psi(B^*, \Sigma^{(1:K)})\log(p-s)$ where c_{p2} is a constant, then l_1/l_2-regularized Lasso fails to recover the support union. In particular, $\psi(B^*, \Sigma^{(1:K)})$ captures the impact of the sparsity of K regression vectors and the statistical properties of the design matrices on the threshold for support recovery. Numerical results are provided to demonstrate the advantages of joint support union recovery using multi-task Lasso over individual support recovery using single-task Lasso.

1 Introduction

Linear regression is a simple but practically very useful statistical model, in which an n sample response vector $\overrightarrow{Y}$ can be modeled as

$$\overrightarrow{Y} = X\overrightarrow{\theta} + \overrightarrow{W}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix containing n samples of feature vectors, $\overrightarrow{\theta} = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ contains regression coefficients, and $\overrightarrow{W} \in \mathbb{R}^n$ is the noise vector. The goal is to find the regression coefficients $\overrightarrow{\theta}$ such that the linear relationship is as accurate as possible with regard to a certain performance criterion. The problem is more interesting in high dimensional regime with a sparse regression vector, in which the sample size n can be much smaller than the dimension p of the regression vector.

In order to estimate the sparse regression vector, it is natural to construct an optimization problem with an l_0-constraint on $\overrightarrow{\theta}$, i.e., the number of nonzero components of $\overrightarrow{\theta}$. However, such an optimization problem is nonconvex and in general very difficult to solve in an efficient manner [1]. More recently, the convex relaxation (referred to as Lasso) has been studied with an l_1-constraint on $\overrightarrow{\theta}$ based on the idea in the seminal work [2] by Tibshirani, [3] by Chen, Donoho and Saunders, and [4] by Donoho and Huo. More specifically, the regression problem can be formulated as:

$$\min_{\overrightarrow{\theta}\in\mathbb{R}^p} \frac{1}{n}\|\overrightarrow{Y} - X\overrightarrow{\theta}\|_{l_2}^2 + \lambda_n\|\overrightarrow{\theta}\|_{l_1}.$$

The l_1-regularized estimator has been proved to be equivalent to Dantzig Selector [5], which was proposed in [6]. Various efficient algorithms have been developed to solve the above convex problem efficiently (see a review monograph [7]), although the objective function is not differentiable everywhere due to l_1-regularization. Moreover, the l_1-regularization is critical to force the minimizer to have sparse components as shown in [2–4].

A vast amount of recent work has studied the high dimensional linear regression problem via l_1-regularized Lasso under various assumptions. For example, the studies in [8–11] investigated the noiseless scenario and showed that recovery of true coefficients could be guaranteed with certain conditions on design matrices and sparsity. A number of studies developed l_1-regularized quadratic programming to achieve sparsity

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

recovery for noisy scenarios. Some work (e.g., [12–14]) focused on the problem with deterministic design matrices, while other work (e. g., [15, 16]) studied the problem with random design matrices. Lasso has also been proved to be useful for generalized linear models (GLMs) such as logistic regression [17, 18] and some specific exponential families [19].

Inspired by the success of Lasso in the single-task problem, block-regularized Lasso for the high-dimensional multivariate linear regression problem (i.e., the multi-task linear regression problem) was intensively studied (see, e.g., [20–23] and references therein). The model of the problem is given by

$$Y = XB^* + W \tag{1}$$

where $Y \in \mathbb{R}^{n\times K}$ has each column corresponding to the output of one task, $X \in \mathbb{R}^{n\times p}$ is the design matrix, the regression matrix $B^* \in \mathbb{R}^{p\times K}$ has each column corresponding to the regression vector for one task, and $W \in \mathbb{R}^{n\times K}$ has each column corresponding to the noise vector of one task. For each column $\overrightarrow{Y}^{(k)}$ of the matrix Y, it is clear that $\overrightarrow{Y}^{(k)} = X\overrightarrow{\theta}^{*(k)} + \overrightarrow{W}^{(k)}$, where $\overrightarrow{\theta}^{*(k)}$ and $\overrightarrow{W}^{(k)}$ are the corresponding columns in B^* and W. Then each column is a single-task linear regression problem and can be solved individually. However, the K individual problems (i.e., tasks) can also be coupled together via a block regularized Lasso and solved jointly in one problem.

Various types of block regularization have been proposed and studied. In [24], the l_1/l_2-regularization was adopted to recover the support union of the regression vectors. More specifically, the following problem was studied

$$\min_{B\in\mathbb{R}^{p\times K}} \frac{1}{2n}|||Y - XB|||_F^2 + \lambda_n \|B\|_{l_1/l_2},$$

where $\|\cdot\|_{l_a/l_b}$ is defined in (5) in section 2.1. Sufficient and necessary conditions for correct recovery of the support union (i.e., the union of the supports of all columns of B^*) have been characterized. The l_1/l_q-regularized Lasso was adopted for learning structured linear regression model in [25]. Block regularized Lasso has also been applied to study various other models and problems in [26–36].

In the multivariate linear regression problem given in (1), the design matrix is identical for all tasks, namely, i.e., X is the same for all column vectors of Y and B^*. However, in many applications, it is often the case that different output variables may depend on design variables that are different or distributed differently. Thus, the resulting model includes K linear regression models with different design matrices and is given by:

$$\overrightarrow{Y}^{(k)} = X^{(k)}\overrightarrow{\theta}^{*(k)} + \overrightarrow{W}^{(k)} \tag{2}$$

for $k = 1, \ldots, K$, where $\overrightarrow{Y}^{(k)} \in \mathbb{R}^n$, $X^{(k)} \in \mathbb{R}^{n\times p}$, $\overrightarrow{\theta}^{*(k)} \in \mathbb{R}^p$, and $\overrightarrow{W}^{(k)} \in \mathbb{R}^n$. We refer to the above problem as the *multivariate multi-response (MVMR) linear regression model*, and the goal is to recover $\overrightarrow{\theta}^{*(k)}$ for $k = 1, \ldots, K$ jointly. This problem has been studied in [37] via the l_1/l_2 regularized Lasso for fixed matrices $X^{(1)}, \ldots, X^{(K)}$. For random design matrices, this model has been studied via l_1/l_∞-regularized Lasso in [38] and via $l_1/l_1 + l_1/l_\infty$-regularized Lasso in [22] for incorporating both row sparsity and individual sparsity.

In this paper, we study the MVMR problem for random design matrices via l_1/l_2-regularized Lasso. It is assumed that the design matrices are Gaussian distributed, and are independent but not identical across k. For each task k, the row vector of $X^{(k)}$ is Gaussian with mean zero and the covariance matrix $\Sigma^{(k)}$ for $k = 1, \ldots, K$. The noise vectors and hence the output vectors are also Gaussian distributed and independent across tasks. We are interested in joint recovery of the union of the support sets (i.e., the support union) of regression vectors $\overrightarrow{\theta}^{*(1)}, \ldots, \overrightarrow{\theta}^{*(K)}$. We collect these vectors together as a matrix $B^* = \left[\overrightarrow{\theta}^{*(1)}, \ldots, \overrightarrow{\theta}^{*(K)}\right]$.

We adopt the l_1/l_2-regularized Lasso problem for recovery of the support union via the following optimization problem:

$$\min_{B\in\mathcal{R}^{p\times K}} \frac{1}{2n}\sum_{k=1}^{K}\left\|\overrightarrow{Y}^{(k)} - X^{(k)}\overrightarrow{\theta}^{(k)}\right\|_2^2 + \lambda_n \|B\|_{l_1/l_2}$$

where $B = \left[\overrightarrow{\theta}^{(1)}, \ldots, \overrightarrow{\theta}^{(K)}\right]$. In this way, the K linear regression problems are coupled together via the regularization constraint. We show that this approach is advantageous as opposed to individual recovery of the support set for each linear regression problem. This is because the K regression models may share their samples in joint support recovery so that the total number of samples needed can be significantly reduced compared to performing each task individually.

1.1 Main Results and Contributions

In the following, we summarize the main contributions of this work. Our results contain two parts: the achievability and the converse, corresponding respectively to sufficient and necessary conditions under which the l_1/l_2- regularized Lasso recovers the support union for the MVMR linear regression problem. Our proof adapts the techniques developed by Wainwright in [15] and by Obozinski, Wainwright, and Jordan in [24], but involves nontrivial development to deal with the differently distributed design matrices across tasks. This also leads to interesting generalization of the results in [24] as we articulate in section 1.2.

More specifically, we show that under certain conditions that the distributions of the design matrices satisfy, if $n > c_{p1}\psi(B^*, \Sigma^{(1:K)})\log(p-s)$, where $\psi(\cdot)$ is defined in (6) in Section 2.1 and c_{p1} is a constant, then the l_1/l_2-regularized Lasso recovers the support union for the MVMR linear regression problem; and if $n < c_{p2}\psi(B^*, \Sigma^{(1:K)})\log(p-s)$, where c_{p2} is a constant, then the l_1/l_2-regularized Lasso fails to recover the support union. Thus, $\psi(B^*, \Sigma^{(1:K)})\log(p-s)$ serves as a sharp threshold on the sample size.

In particular, $\psi(B^*, \Sigma^{(1:K)})$ captures the sparsity of B^* and the statistical properties of the design matrices, which are important in determining the sufficient and necessary conditions for successful recovery of the support union. The property of $\psi(B^*, \Sigma^{(1:K)})$ also captures the advantages of the multi-task Lasso over solving each problem individually via the single-task Lasso. We show that when the K tasks share the same support sets (although the design matrices can be differently distributed), $\psi(B^*, \Sigma^{(1:K)}) = \frac{1}{K}\max_{1\le k\le K}\psi(\overrightarrow{\theta}^*_k, \Sigma^{(k)})$. This means that the number of samples needed per task for the multi-task Lasso to jointly recover the support union is reduced by K compared to that of the single-task Lasso to recover each support set individually. On the other hand, if the K tasks have disjoint support sets, then $\psi(B^*, \Sigma^{(1:K)}) = \max_{1\le k\le K}\psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)})$. This implies that the number of samples needed per task to correctly recover the support union is almost the same as that of the single-task Lasso to recover each support individually. Between these two extreme cases, tasks can have overlapped support sets with different overlapping levels, and the impact of these properties on the sample size for recovery of the support union is precisely captured by $\psi(B^*, \Sigma^{(1:K)})$.

1.2 Comparison to Previous Results

The MVMR model (with differently distributed design matrices across tasks) can be viewed as generalization of the multivariate model (with an identical design matrix across tasks) studied in [24]. It is thus interesting to compare our results to the results in [24]. For the scenario when the tasks share the same regression vector, it is shown in [24] that the major advantage of jointly solving a multi-task Lasso problem over solving each single-task Lasso problem individually is reduction of effective noise variance by the factor K. But the sample size needed per task for recovery of the support union via multi-task Lasso is the same as that needed for recovery of each support set individually via single-task Lasso. This implies that multi-task Lasso does not offer benefit in reducing the sample size (in the order sense) for this case. Our result, on the other hand, shows that the benefit in the sample size of multi-task Lasso appears when the design matrices are differently distributed across tasks. For such a case, although the design matrices are different across K tasks, we show that the same total sample size (i.e., as the case with the same design matrix) is still sufficient for correct recovery of the support union. Hence, the sample size needed per task is reduced by K via multi-task Lasso compared to recovery of each support set individually via single-task Lasso. Consequently, our result is a nontrivial generalization of the result in [24]. For the scenario when the tasks have disjoint support sets, our result is consistent with the result in [24], which suggests that there is no advantage of performing multi-task Lasso as opposed to performing single-task Lasso for each task.

As we mentioned before, the MVMR model was also studied in [22, 38], in which l_1/l_∞ and $l_1/l_1 + l_1/l_\infty$-regularization were adopted for support union recovery, respectively. We study the same model but under l_1/l_2-regularization. More importantly, we characterize the gain in saving the sample size due to multi-task Lasso. Furthermore, we characterize the conditions on the sample size for recovery of support union as a threshold for the general model. This type of results were given in [22, 38] only for specific scenarios.

1.3 Relationship to Jointly Learning Multiple Markov Networks

One application of the MVMR linear regression model is to jointly learning multiple Gaussian Markov network structures. In this context, it solves a multi-task neighbor selection problem. This is also a natural scenario where features and their distributions vary across tasks of linear regression.

We consider K Gaussian Markov networks, each with $p+1$ nodes represented by $X_1^{(k)}, \ldots, X_{p+1}^{(k)}$ for $k = 1, \ldots, K$. The distribution of the Gaussian vector for graph k is given by $\mathcal{N}\left(0, \Sigma_{p+1}^{(k)}\right)$, where $\Sigma_{p+1}^{(k)} \in \mathbb{R}^{(p+1)\times(p+1)}$. Assume for each graph, there are n i.i.d. samples generated based on the joint distribution of the nodes. The objective is to estimate the connection relationship of nodes based on the samples. We denote n samples of each variable $X_j^{(k)}$ by a column vector $\overrightarrow{X}_j^{(k)} \in \mathbb{R}^n$ for $j = 1, \ldots, p+1$ and $k = 1, \ldots, K$. For each graph k and each node with index a, the sample vector $\overrightarrow{X}_a^{(k)}$ can be expressed as:

$$\overrightarrow{X}_a^{(k)} = X_{-a}^{(k)}\overrightarrow{\theta}^{(k)} + \overrightarrow{W}_a^{(k)} \tag{3}$$

where $X_{-a}^{(k)}$ is an $n \times p$ matrix that groups all column vectors $\overrightarrow{X}_j^{(k)}$ for $j \neq a$, $\overrightarrow{\theta}^{(k)}$ is a p-dimensional vector consisting of the estimation parameters of $X_a^{(k)}$

from $X_j^{(k)}$ with $j \neq a$, and $\overrightarrow{W}_a^{(k)}$ is the n-dimensional Gaussian vector containing i.i.d. components with zero mean and variance given by

$$\begin{aligned}{\sigma_W^{(k)}}^2 &= Var(X_{1a}) \\ &- Cov(X_{1a}, X_{1,-a})Cov^{-1}(X_{1,-a})Cov(X_{1,-a}, X_{1a}).\end{aligned}$$

It has been shown that the nonzero components of the vector $\overrightarrow{\theta}^{(k)}$ represent existence of the edges between the corresponding nodes and node a in graph k. Hence, estimation of the support set of $\overrightarrow{\theta}^{(k)}$ provides an estimation of the graph structure, which is referred to as the *neighbor selection problem* [39].

Therefore, multi-task Lasso for the MVMR linear regression problem provides an useful approach for joint neighbor selection over K graphs. It is clear that in this case, the design matrices $X_{-a}^{(k)}$ in general have different distributions across k, and hence the multi-feature model is well justified. We note that jointly learning multiple graphs has also been studied in [40,41], which adopted a different objective function of the precision matrix Σ^{-1}. Via the MVMR linear regression model, we characterize the threshold-based sufficient and necessary conditions for joint recovery of the graphs.

2 Problem Formulation and Notations

In this paper, we study the MVMR linear regression problem given by (2), which contains K linear regressions. Here, the design matrices $X^{(1)}, \ldots, X^{(K)}$ and noise vectors $\overrightarrow{W}^{(1)}, \ldots, \overrightarrow{W}^{(K)}$ are Gaussian distributed, and are independent but not identical across k. For each task k, $X^{(k)}$ has independent and identically distributed (i.i.d.) row vectors with each being Gaussian with mean zero and covariance matrix $\Sigma^{(k)}$, and the noise vector $\overrightarrow{W}^{(k)}$ has i.i.d. components with each being Gaussian with mean zero and variance ${\sigma_W^{(k)}}^2$. We let $\sigma_{max} = \max_{1 \leq k \leq K} {\sigma_W^{(k)}}^2$.

In (2), $\overrightarrow{\theta}^{*(k)}$ denotes the true regression vector for each task k. We define the support set for each $\overrightarrow{\theta}^{*(k)}$ as $S_k := \{j \in \{1, \ldots, p\} | \overrightarrow{\theta}_j^{*(k)} \neq 0\}$. The support union over K tasks is defined to be $S := \cup_{k=1}^K S_k$. In this paper, we are interested in estimating the support union jointly for K tasks.

We adopt the l_1/l_2-regularized Lasso to jointly recover the support union for the MVMR linear regression model. More specifically, we solve the following multi-task Lasso problem rewritten below:

$$\min_{B \in \mathcal{R}^{p \times K}} \frac{1}{2n} \sum_{k=1}^{K} \left\| \overrightarrow{Y}^{(k)} - X^{(k)} \overrightarrow{\theta}^{(k)} \right\|_2^2 + \lambda_n \|B\|_{l_1/l_2} \tag{4}$$

where $B = \left[\overrightarrow{\theta}^{(1)}, \ldots, \overrightarrow{\theta}^{(K)}\right]$. In this way, the K linear regression problems are coupled together via the regularization constraint. In this paper, we characterize conditions under which the solution to the above multi-task Lasso problem correctly recover the support union of the true regression vectors for K tasks.

2.1 Notations

We introduce some notations that we use in this paper. For a matrix $A \in \mathbb{R}^{p \times K}$, we define the l_a/l_b block norm as

$$\|A\|_{l_a/l_b} := \left[\sum_{i=1}^{p} \left(\sum_{j=1}^{K} |A_{ij}|^b \right)^{a/b} \right]^{1/a}. \tag{5}$$

We also define the operator norm for a matrix as

$$|||A|||_{a,b} := \sup_{\|x\|_b = 1} \|Ax\|_a.$$

In particular, we define the spectral norm as $|||A|||_2 = |||A|||_{2,2}$ and the l_∞-operator norm as $|||A|||_\infty = |||A|||_{\infty,\infty}$, which are special cases of the operator norm.

For matrix $B = \left[\overrightarrow{\theta}^{(1)}, \ldots, \overrightarrow{\theta}^{(K)}\right]$ that appears in (4), $\overrightarrow{\theta}^{(k)}$ denotes its kth columns for $k = 1, \ldots, K$. We further let B_i to be the ith row of B. Similarly, for $B^* = \left[\overrightarrow{\theta}^{*(1)}, \ldots, \overrightarrow{\theta}^{*(K)}\right]$ that contains true regression vectors, its kth column is denoted by $\overrightarrow{\theta}^{*(k)}$ and the ith row is denoted by B_i^*. We next define the normalized row vectors of B^* as

$$Z_i^* = \begin{cases} \frac{B_i^*}{\|B_i^*\|_{l_2}} & \text{if } B_i^* \neq 0 \\ 0 & \text{otherwise,} \end{cases}$$

and define the matrix Z^* to contain Z_i^* as its ith row for $i = 1, \ldots, p$. To avoid confusion, we use $\widehat{B}$ to denote the solution to the multi-task Lasso problem (4).

The support union $S(B)$ for a matrix $B \in \mathbb{R}^{p \times K}$ is denoted as $S(B) = \{i \in \{1, \ldots, p\} | B_i \neq 0\}$, which includes indices of the nonzero rows of the matrix B. We use S to represent $S(B^*)$ (i.e., the true support union) for convenience and use S^c to denote the complement of the set S. We let $s = |S|$ denote the size of the set S. For any matrix $X^{(k)} \in \mathbb{R}^{n \times p}$, the matrix $X_S^{(k)}$ contains the columns of matrix $X^{(k)}$ with column indices

in the set S, and $X^{(k)}_{S^c}$ contains the columns of matrix $X^{(k)}$ with column indices in the set S^c. Similarly, B^*_S and Z^*_S respectively contain rows of B^* and Z^* with indices in S.

As each row of matrix $X^{(k)}$ is Gaussian distributed as $\mathcal{N}(0, \Sigma^{(k)})$, we use $\Sigma^{(k)}_{SS}$ to denote the covariance matrix for each row of $X^{(k)}_S$, and use $\Sigma^{(k)}_{S^cS}$ to denote the cross covariance between rows of $X^{(k)}_{S^c}$ and $X^{(k)}_S$.

For convenience, we use $\Sigma^{(1:K)}$ to denote a set of matrices $\Sigma^{(1)}, \ldots, \Sigma^{(K)}$. We also define the following functions of matrices $Q^{(1:K)}$ to simply our notations:

$$\rho_u\left(Q^{(1:K)}\right) := \max_{j\in S^c} \max_{1\le k\le K} Q^{(k)}_{jj},$$

$$\rho_l\left(Q^{(1:K)}\right) := \min_{i,j\in S^c, j\neq i} \min_{1\le k\le K} \left[Q^{(k)}_{jj} + Q^{(k)}_{ii} - 2Q^{(k)}_{ji}\right].$$

In particular, our results contain the functions $\rho_u\left(\Sigma^{(1:K)}_{S^cS^c|S}\right)$ and $\rho_l\left(\Sigma^{(1:K)}_{S^cS^c|S}\right)$, where $\Sigma^{(k)}_{S^cS^c|S}$ is the covariance matrix of each row of $X^{(k)}_{S^c}$ with $X^{(1:K)}_S$ given.

For matrix B^*, we define $b^*_{min} = \min_{j\in S} \left\|B^*_j\right\|_{l_2}$. We define the following function that captures the sparsity of B^* and the statistical properties of the design matrices $X^{(1:K)}_S$:

$$\psi(B^*, \Sigma^{(1:K)}) := \max_{1\le k\le K} \overrightarrow{Z}^{*T}_{Sk}\left(\Sigma^{(k)}_{SS}\right)^{-1}\overrightarrow{Z}^*_{Sk}, \quad (6)$$

where $\overrightarrow{Z}^*_{Sk}$ is the kth column of Z^*_S. We note that this definition of $\psi(\cdot)$ function is different from the previous work [24]. Here, due to different design matrices, $\psi(\cdot)$ depends on K quantities $\overrightarrow{Z}^{*T}_{Sk}\left(\Sigma^{(k)}_{SS}\right)^{-1}\overrightarrow{Z}^*_{Sk}$ with each depending on a column vector $\overrightarrow{Z}^*_{Sk}$.

3 Main Results

In this section, we provide our main results on using multi-task Lasso to recover the support union for the MVMR linear regression model. Our results contain two parts: one is the achievability, i.e., sufficient conditions for the l_1/l_2-regularized Lasso to recover the support union; and the other is the converse, i.e., conditions under which the l_1/l_2-regularized Lasso fails to recover the support union. We then discuss implications of our results by considering a few representative scenarios, and compare our results with those for the multivariate linear regression with an identical design matrix across tasks.

3.1 Achievability and Converse

We first introduce a number of conditions on covariance matrices $\Sigma^{(k)}$ for $k = 1, \ldots, K$, which are useful for the statements of our results.

(C1). There exists a real number $\gamma \in (0, 1]$ such that $\left|\left|\left|A\right|\right|\right|_\infty \le 1 - \gamma$, where $A_{js} = \max_{1\le k\le K}\left|\left(\Sigma^{(k)}_{S^cS}\left(\Sigma^{(k)}_{SS}\right)^{-1}\right)_{js}\right|$ for $j \in S^c$ and $s \in S$.

(C2). There exist constants $0 < C_{min} \le C_{max} < +\infty$ such that all eigenvalues of the matrix $\Sigma^{(k)}_{SS}$ are contained in the interval $[C_{min}, C_{max}]$ for $k = 1, \ldots, K$.

(C3). There exists a constant $D_{max} < +\infty$ such that $\max_{1\le k\le K}\left|\left|\left|\left(\Sigma^{(k)}_{SS}\right)^{-1}\right|\right|\right|_\infty \le D_{max}$.

In this paper, we consider the asymptotic regime, in which $p \to \infty$, $s \to \infty$, and $\log(p - s) \to +\infty$. In such a regime, we introduce the conditions on the regularization parameter and the sample size n as follows:

(P1). Regularization parameter $\lambda_n = \sqrt{\frac{f(p)\log p}{n}}$, where the function $f(p)$ is chosen such that $f(p) \to +\infty$ as $p \to +\infty$, and $\frac{f(p)\log p}{n} \to 0$ as $n \to \infty$, i.e., $\lambda_n \to 0$ as $n \to +\infty$.

(P2). Define $\rho(n, s, \lambda_n)$ as

$$\rho(n, s, \lambda_n) := \sqrt{\frac{8\sigma^2_{max} s\log s}{nC_{min}}} + \lambda_n\left(D_{max} + \frac{12s}{C_{min}\sqrt{n}}\right)$$

and require $\frac{\rho(n,s,\lambda_n)}{b^*_{min}} = o(1)$.

The following theorem characterizes sufficient conditions for recovery of the support union via the multi-task Lasso.

Theorem 1. *Consider the MVMR problem in the asymptotic regime, in which $p \to \infty$, $s \to \infty$ and $\log(p - s) \to \infty$. We assume that the parameters $\left(n, p, s, B^*, \Sigma^{(1:K)}\right)$ satisfy the conditions (C1)-(C3), and (P1)-(P2). If for some small constant $v > 0$,*

$$n > 2(1 + v)\psi\left(B^*, \Sigma^{(1:K)}\right)\log(p - s)\frac{\rho_u\left(\Sigma^{(1:K)}_{S^cS^c|S}\right)}{\gamma^2}, \quad (7)$$

then the problem (4) has a unique solution $\widehat{B}$, the support union $S(\widehat{B})$ is the same as the true support union $S(B^*)$, and $\|\widehat{B} - B^*\|_{l_\infty/l_2} = o(b^*_{min})$ with the probability greater than

$$1 - K\exp(-c_0\log s) - \exp(-c_1\log(p - s)) \quad (8)$$

where c_0 and c_1 are constants.

Theorem 1 provides sufficient conditions on the sample size such that the solution to the multi-task Lasso problem correctly recovers the support union of the MVMR linear regression model. We next provides a

theorem about the conditions on the sample size under which the solution to the multi-task Lasso problem fails to recover the support union.

Theorem 2. *Consider the MVMR problem in the asymptotic regime, in which $p \to \infty$, $s \to \infty$ and $\log(p-s) \to \infty$. We assume that the parameters $\left(n, p, s, B^*, \Sigma^{(1:K)}\right)$ satisfy the conditions (C1)-(C2) and the conditions: $s/n = o(1)$ and $\frac{1}{\lambda_n^2 s} \to 0$. If for some small constant $v > 0$,*

$$n < 2(1-v)\psi(B^*, \Sigma^{(1:K)}) \log(p-s) \frac{\rho_l\left(\Sigma^{(1:K)}_{(S^c S^c|S)}\right)}{(2-\gamma)^2}, \tag{9}$$

then with the probability greater than

$$1 - \exp(-c_2 s) - c_3 \exp\left(-c_4 \frac{n}{s}\right) \tag{10}$$

for some positive constants c_2, c_3 and c_4, no solution $\widehat{B}$ to the multi-task Lasso problem recovers the true support union and achieves $\|\widehat{B} - B^*\|_{l_\infty/l_2} = o(b^*_{min})$.

The proofs of Theorems 1 and 2 adapt the techniques developed by Obozinski, Wainwright, and Jordan in [24], but involve novel development to deal with the differently distributed design matrices across tasks.

Combining Theorems 1 and 2, the quantity $\psi(B^*, \Sigma^{(1:K)}) \log(p-s)$ serves as a threshold on the sample size n, which is tight in the order sense. As the sample size is above the threshold, the multi-task Lasso recovers the true support union, and as the sample size is below the threshold, the multi-task Lasso fails to recover the true support union. The following proposition provides bounds on the scaling behavior of the function $\psi(B^*, \Sigma^{(1:K)})$ in the asymptotic regime.

Proposition 1. *Consider the MVMR linear regression model with the regression matrix B^* and the covariance matrices $\Sigma^{(1:K)}$ satisfying the condition (C2), the function $\psi(B^*, \Sigma^{(1:K)})$ is bounded:*

$$\frac{s}{KC_{min}} \leq \psi(B^*, \Sigma^{(1:K)}) \leq \frac{s}{C_{min}}.$$

In the next subsection, we explore the properties of the quantity $\psi(B^*, \Sigma^{(1:K)})$ in order to understand the impact of sparsity of B^* and covariance matrices $\Sigma^{(1:K)}$ on the conditions for recovering the support union.

3.2 Implications

The quantity $\psi(B^*, \Sigma^{(1:K)})$ captures sparsity of B^* and statistical properties of design matrices (i.e., $\Sigma^{(1:K)}$), and hence plays an important role in determining the conditions on the sample size for recovery of the support union as shown in Theorems 1 and 2.

In this section, we analyze $\psi(B^*, \Sigma^{(1:K)})$ for a number of representative cases in order to understand advantages of multi-task Lasso which solves multiple linear regression problems jointly over single-task Lasso which solves each linear regression problem individually.

We denote $\psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)})$ as the function corresponding to a single linear regression problem, where $\overrightarrow{\theta}^{*(k)}$ represents the kth column of B^*. It is clear that $\psi(B^*, \Sigma^{(1:K)})$ captures the threshold on the sample size for the multi-task Lasso problem. Comparison of $\psi(B^*, \Sigma^{(1:K)})$ and $\psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)})$ provides comparison between multi-task Lasso and single-task Lasso in terms of the number of samples needed for recovery of the support union/set. We explicitly express $\psi(B^*, \Sigma^{(1:K)})$ and $\psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)})$ as follows:

$$\psi(B^*, \Sigma^{(1:K)}) = \max_{1 \leq k \leq K} \sum_{i \in S} \sum_{j \in S} \frac{B^*_{ik} B^*_{jk}}{\|B^*_i\|_{l_2} \left\|B^*_j\right\|_{l_2}} \left(\left(\Sigma^{(k)}_{SS}\right)^{-1}\right)_{ij} \tag{11}$$

$$\psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)}) = \sum_{i \in S} \sum_{j \in S} \frac{\overrightarrow{\theta}^{*(k)}_i \overrightarrow{\theta}^{*(k)}_j}{\left|\overrightarrow{\theta}^{*(k)}_i\right| \left|\overrightarrow{\theta}^{*(k)}_j\right|} \left(\left(\Sigma^{(k)}_{SS}\right)^{-1}\right)_{ij} \tag{12}$$

where B^*_{ik} denotes the (i,k)th entry of the matrix B^* and $\overrightarrow{\theta}^{*(k)}_i$ denotes the ith entry of the vector $\overrightarrow{\theta}^{*(k)}$.

We first study the scenario, in which all K tasks have the same regression vectors, and hence have the same support sets.

Corollary 1. *(Identical Regression Vectors) If B^* has identical column vectors, i.e., $\overrightarrow{\theta}^{*(k)} = \overrightarrow{\theta}^*$ for $k = 1, \ldots, K$, then*

$$\psi(B^*, \Sigma^{(1:K)}) = \frac{1}{K} \max_{1 \leq k \leq K} \psi(\overrightarrow{\theta}^*, \Sigma^{(k)}). \tag{13}$$

Remark 1. *Corollary 1 implies that the number of samples per task needed to correctly recover the support union via multi-task Lasso is reduced by a factor of K compared to single-task Lasso that recovers each support set individually.*

It can be seen that although the K tasks involve design matrices that have different covariances, as long as dependence of the output variables on the feature variables is the same for all tasks, the tasks share samples in multi-task Lasso to recover the support union so that the sample size needed per task is reduced by a factor of K. Hence, there is a significant advantage of grouping tasks with similar regression vectors together for multi-task learning.

Corollary 1 can be viewed as a generalization of the result in [24], in which the design matrices for the tasks are the same. The result in [24] suggests that if the tasks share the same regression vector, there is no benefit in terms of the number of samples needed for support recovery using multi-task Lasso compared to single-task Lasso. Our result suggests that the benefit of multi-task Lasso in fact appears when the design matrices are differently distributed. For such a case, we show that the total number of samples needed for recovery of the support union is not increased for the case with differently distributed design matrices compared to the case with the same design matrix (studied in [24]). Hence, sample size needed per design matrix (i.e., per task) is reduced by the factor K. Moreover, compared to recovery of each support set individually via single-task Lasso, multi-task Lasso also reduces sample size per task by the factor K. However, such an advantage does not appear if the K tasks have the same design matrix and regression vectors as in [24].

We next study a more general case when regression vectors are also different across tasks (but the support sets of tasks are the same) in addition to varying design matrices across tasks.

Corollary 2. *(Varying Regression Vectors with Same Supports) Suppose all entries $B^*_{jk} > 0$ for $j \in S$ and $k = 1, \ldots, K$, and all coefficients are bounded, i.e., $\bar{B}_k - \Delta_k \leq B^*_{jk} \leq \bar{B}_k + \Delta_k$, where $\Delta_k > 0$ is a small perturbation constant with $\bar{B}_k > \Delta_k$. Then,*

$$\frac{\psi(B^*, \Sigma^{(1:K)})}{\max_{1\leq k\leq K} \psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)})} \leq \frac{1}{K} \max_{1\leq k\leq K} \frac{\left(\bar{B}_k + \Delta_k\right)^2}{\left(\bar{B}_k - \Delta_k\right)^2}.$$

Corollary 2 is a strengthened version of Corollary 1 in that Corollary 2 allows both the regression vectors and design matrices to be different across tasks and still shows that the number of samples needed is reduced by a factor of K compared to single-task Lasso, as long as the support sets across tasks are the same.

Corollary 3. *(Disjoint Support Sets) Suppose the distribution of all design matrices are the same, i.e., $\Sigma^{(k)} = \Sigma$ and $\Sigma^{(k)}_{SS} = \Sigma_{SS}$ for $k = 1, \ldots, K$, and suppose that the support sets S_k of all tasks are disjoint. Let $s_k = |S_k|$, and hence $s = \sum_{k=1}^K s_k$. Then,*

$$\psi(B^*, \Sigma^{(1:K)}) = \max_{1\leq k\leq K} \psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)}).$$

We note that

$$\max_{1\leq k\leq K} \psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)}) \log(p-s) \leq \max_{1\leq k\leq K} \psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)}) \log(p-s_k).$$

Since the number of samples needed per task for multi-task Lasso is proportional to $\max_{1\leq k\leq K} \psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)}) \log(p-s)$, and the number of samples needed for single-task Lasso for task k is proportional to $\psi(\overrightarrow{\theta}^{*(k)}, \Sigma^{(k)}) \log(p-s_k)$, the above equation implies that the required number of samples for multi-task Lasso is smaller than (in fact almost the same as) that for single-task Lasso.

Corollary 3 suggests that if the tasks have disjoint support sets for regression vectors, the advantage of the multi-task Lasso disappear. This is reasonable because the tasks do not benefit from sharing the samples for recovering the supports if their supports are disjoint. The essential message of Corollary 3 should not change if the tasks have different design matrices and/or different regression vectors. The critical assumption in Corollary 3 is the disjoint support sets.

Corollaries 1 and 3 provide two extreme cases when the tasks share the same support sets and have disjoint support sets, respectively. The number of samples needed per task for recovery of the support union goes from $1/K$ of to the same as the sample size needed for single-task Lasso. It is conceivable that between these two extreme cases, tasks may have overlapped support sets with various overlapping levels. Correspondingly, the number of samples needed for recovering the support union should depend on the overlapping levels of the support sets and is captured precisely by the quantify $\psi(B^*, \Sigma^{(1:K)})$. We demonstrate such behavior via our numerically results in the next section.

4 Numerical Results

In this section, we provide numerical simulations to demonstrate our theoretical results on using block-regularized multi-task Lasso for recovery of the support union for the MVMR linear regression model. We study how the sample size needed for correct recovery of the support union depends on sparsity of the regression vectors, on the distributions of the design matrices, and on the number of tasks.

We first study the scenario considered in Corollary 1 when the K tasks have the same regression vectors, i.e., $B^* = \overrightarrow{\theta}^* \vec{1}_K^T$. We set $\overrightarrow{\theta}^* = \frac{1}{\sqrt{K}} \vec{1}_S$, where S is the common support set across K tasks. We set the covariance matrix $\Sigma^{(k)}$ different across K tasks as follows. For $k = 1, \ldots, K$, we set $Cov(X_a, X_b) > 0$ (where $a, b \in \{1, 2, \ldots, p\}$) if $a = b \pm 1$, otherwise $Cov(X_a, X_b) = 0$. $Cov(X_a, X_b) = 1 + 1/k$ if $a = b \pm 1$ and a is odd. $Cov(X_a, X_b) = 1 - 0.8/k$ if $a = b \pm 1$ and a is even. The sparsity of linear regression vectors is linearly proportional to the dimension p, i.e., $s = \alpha p$, with the parameter α controlling the sparsity of the model. We set $\alpha = 1/8$. We choose the dimension $p = 128, 256, 512$. We solve multi-task Lasso for recovery

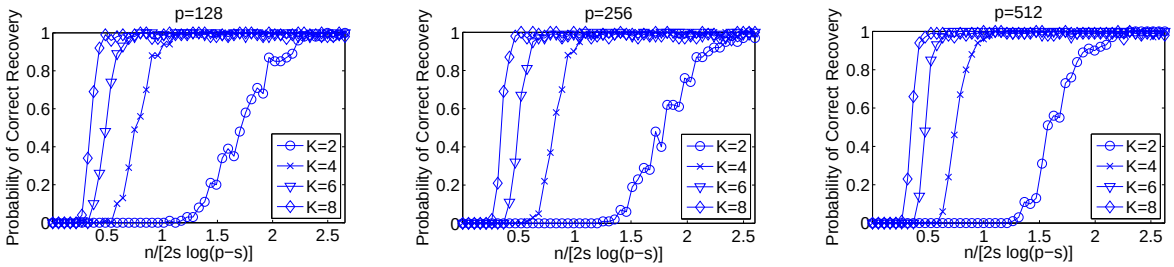

Figure 1: Impact of number of tasks on the sample size for scenarios with identical regression vectors and varying distributions for design matrices across tasks

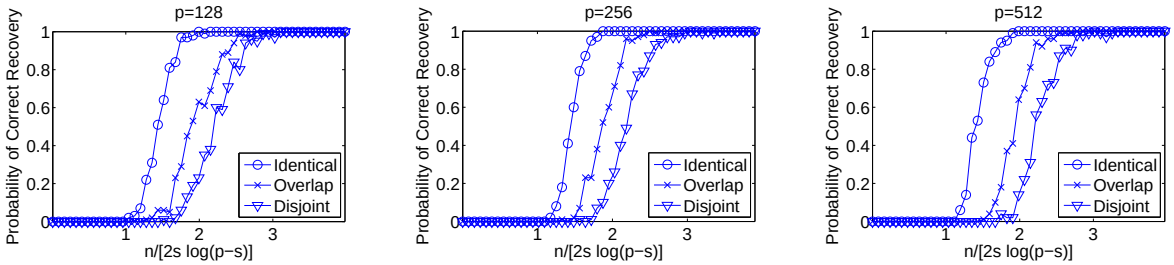

Figure 2: Impact of overlapping levels of support sets on the sample size with same regression values for overlapping entries and identical distributions for design matrices across tasks

of the support union for $K = 2, 4, 6, 8$. We set the regularization parameter $\lambda_n = 3.5 \times \sqrt{\log(p-s)\log s/n}$.

Fig. 1 plots the probability of correct recovery of the support union as a function of the scaled sample size for $p = 128, 256, 512$. It can be seen that the sample size for guaranteeing correct recovery scales in the order of $s\log(p-s)$ for all plots. Moreover, as the number of tasks K increases, the sample size (per task) needed for correct recovery decreases inversely proportionally with K, which is consistent with Corollary 1. These results demonstrate that when the regression vectors are the same across tasks, multi-task Lasso has a great advantage compared to single-task Lasso in terms of reduction in the sample size needed per task.

We next study how the overlapping levels of the support sets across tasks affect the sample size for correct recovery of the support union. We set $K = 2$, i.e., two tasks, and study three overlapping models for the two tasks: (1) same support sets $S_1 = S_2 = \{j \le p : 8t_{pe}+1\}$, where $t_{pe} \ge 0$ is an integer; (2) disjoint support sets $S_1 \bigcap S_2 = \phi$ in which $S_1 = \{j \le p : 16t_{pe}+1\}$ and $S_2 = \{j \le p : 16t_{pe}+2\}$; (3) overlapping support sets in which $S_1 = \{j : j = 24t_{pe}+1 \text{ or } j = 24t_{pe}+2\}$, and $S_2 = \{j : j = 24t_{pe}+2 \text{ or } j = 24t_{pe}+3\}$. We choose the linear sparsity model with $\alpha = 1/8$. We set $p = 128, 256, 512$, and $\Sigma^{(k)} = I_p$ for $k = 1$ and 2. We also set $\lambda_n = 3.5 \times \sqrt{\log(p-s)\log s/n}$.

Fig. 2 compares the probability of correct recovery of the support union as a function of the scaled sample size for the three overlapping models. It can be seen that the model with the same support set requires the smallest sample size, and the model with disjoint support sets requires the largest sample size. The model with overlapping support sets needs the sample size between the two extreme models. This is reasonable because as the support sets overlap more, tasks share more information in samples for support recovery and hence need less number of samples for correct recovery.

5 Conclusions

In this paper, we have investigated the Gaussian MVMR linear regression model. We have characterized sufficient and necessary conditions under which the multi-task Lasso guarantees successful recovery of the support union of K linear regression vectors. The two conditions are characterized by a threshold and hence are tight in the order sense. Our numerical results have demonstrated the advantage of joint recovery of the support union compared to using single-task Lasso to recover the support set of each task individually. Further studying the MVMR model under other block-constrains is an interesting topic in the future. Applications of the approach here to structure learning problems based on real data sets such as social network data are also interesting.

Acknowledgements

The work of W. Wang and Y. Liang was supported by a National Science Foundation CAREER Award under Grant CCF-10-26565 and by the National Science Foundation under Grant CCF-10-26566.

References

[1] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

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

[3] S. Chen, D. Donaho, and M. Saunders. Atomic decomposition by basis persuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[4] D. Donoho and X. Huo. Uncertainty principles an ideal atomic decomposition. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, 2001.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[6] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[7] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, Now Publishers, Hanover, MA, USA, 2012.

[8] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 48(9):2558–2567, 2002.

[9] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Trans. Inform. Theory*, 49(6):1579–1581, 2003.

[10] D. M. Malioutov, M. Cetin, and A. S. Willsky. Optimal sparse representations in general overcomplete bases. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.

[11] J. Tropp. Greedy is good: Algorithm resutlts for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[12] J. J. Fuch. Recovery of exact sparse representations in the presence of noise. *IEEE Trans. Inform. Theory*, 51(10):3601–3608, 2005.

[13] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

[14] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics*, 37(1):246–270, 2009.

[15] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.

[16] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[17] L. Meier, S. Van De Geer, and P. Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Stastistical Society*, 70:53–71, 2008.

[18] E. Yang, P. Ravikumar, G. Allen, and Z. Liu. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pages 1367–1375, 2012.

[19] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *AI and Statistics*, volume 9, pages 381–388, 2010.

[20] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B*, 68:49–67, 2006.

[21] J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38(4):1978–2004, 2010.

[22] A. Jalali, P. Ravikumara, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[23] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. to appear in *Statistical Science*, 2012.

[24] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

[25] H. Liu and J. Zhang. On the $l_1 - l_q$ regularized regression. arXiv:0802.1517v1, 2008.

[26] B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

[27] G. Obozinski, B. Tarskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252, 2010.

[28] M. Kolar, J. Lafferty, and L. Wasserman. Union support recovery in multi-task learning. *Journal of Machine Learning Research*, 12:2415–2435, 2011.

[29] M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[30] F. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.

[31] L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *International Conference on Machine Learning (ICML)*, 2009.

[32] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Stastistical Society*, 67:91–108, 2005.

[33] Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems (NIPS)*, volume 25, pages 1268–1276, 2012.

[34] Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16:375–390, 2006.

[35] C. H. Zhang and J. Huang. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1576–1594, 2008.

[36] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Demension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society*, 69(3):329–346, 2007.

[37] K. Lounici, M. Pontil, S. Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics*, 39:2164–2204, 2011.

[38] S. Negahban and M. J. Wainwright. Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization. *IEEE Trans. Inform. Theory*, 57(6):3841–3863, 2011.

[39] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[40] P. Danaher, P. Wang, and D. M. Witten. The joint graphical Lasso for inverse covariance estimation across multiple classes. under review, 2011.

[41] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.