Ultra-high Dimensional Multiple Output Learning With Simultaneous Orthogonal Matching Pursuit: Screening Approach

Mladen Kolar
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA

Eric P. Xing
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

We propose a novel application of the Simultaneous Orthogonal Matching Pursuit (S-OMP) procedure to perform variable selection in ultra-high dimensional multiple output regression problems, which is the first attempt to utilize multiple outputs to perform fast removal of the irrelevant variables. As our main theoretical contribution, we show that the S-OMP can be used to reduce an ultra-high number of variables to below the sample size, without losing relevant variables. We also provide formal evidence that the modified Bayesian information criterion (BIC) can be used to efficiently select the number of iterations in the S-OMP. Once the number of variables has been reduced to a manageable size, we show that a more computationally demanding procedure can be used to identify the relevant variables for each of the regression outputs. We further provide evidence on the benefit of variable selection using the regression outputs jointly, as opposed to performing variable selection for each output separately. The finite sample performance of the S-OMP has been demonstrated on extensive simulation studies.

1 Introduction

Multiple output ultra-high dimensional regression problems commonly arise in a genome-wide association mapping studies. These studies aim to find a small set of causal single-nucleotide polymorphisms (SNP) (*variables*) that account for genetic variations of a large number of genes (*regression outputs*). However, this is a very challenging problem for current statistical methods since the number of variables is likely to reach millions. Genes in a biological pathway are co-expressed as a module and it is often assumed that a causal SNP affects multiple genes in one pathway, but not all of the genes in the pathway. In order to effectively reduce the dimensionality of the problem and to detect the causal SNPs, it is very important to look at how SNPs affect all genes in a biological pathway. Since the experimentally collected data is usually very noisy, regressing genes individually onto SNPs may not be sufficient to identify the relevant SNPs that are only weakly correlated with each gene. However, once the whole biological pathway is examined, it is much easier to find the causal SNPs. In this paper, we demonstrate that the Simultaneous Orthogonal Matching Pursuit (S-OMP) (Tropp et al., 2006) can be used to quickly reduce the dimensionality of the problem, without losing any of the relevant variables.

As the dimensionality of the problem and the number of outputs increase, it becomes computationally hard to solve the commonly used convex programs used to identify relevant variables in multiple output regression problems. Previous work Liu et al. (2009); Lounici et al. (2009); Kim et al. (2009), do not scale well to settings when the number of variables exceeds $\gtrsim 10000$ and the number of outputs exceeds $\gtrsim 1000$ as in genome-wide association studies. Furthermore, estimation error of the regression coefficients depends on the number of variables in the problem, so that the variable selection can improve convergence rates of estimation procedures. These concerns motivate us to propose and study the S-OMP as a fast way to remove many of the irrelevant variables.

Formally, the association mapping problem can be cast as a variable selection problem in a multiple output regression model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{W} \tag{1}$$

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T] \in \mathbb{R}^{n \times T}$ is a matrix of outputs, whose column $\mathbf{y}_t$ is an n-vector for the t-th output, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a random design matrix, with each row $\mathbf{x}_i$ denoting a p-dimensional input, $\mathbf{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T] \in \mathbb{R}^{p \times T}$ is the matrix of regression coefficients and $\mathbf{W} = [\epsilon_1, \ldots, \epsilon_T] \in \mathbb{R}^{n \times T}$ is a matrix of IID random noise, independent of $\mathbf{X}$. We are interested in estimating the regression coefficients, under the assumption that they share a common structure, e.g., there exists a subset of variables with non-zero coefficients for more than one regression output. We informally refer to such outputs as related. Two interesting questions, commonly asked in this setting, are: i) how can information be shared between related outputs in order to improve the efficiency over the independent estimation on each output separately; ii) how to improve the variable selection based on information from related outputs. To address these two questions, one line of research (e.g., Zhang, 2006; Liu et al., 2009; Lounici et al., 2009) has looked into the following estimation procedure; $\hat{\mathbf{B}}$ is a minimizer of

$$\min_{\boldsymbol{\beta}_t \in \mathbb{R}^p, t \in [T]} \sum_{t=1}^{T} ||\mathbf{y}_t - \mathbf{X}\boldsymbol{\beta}_t||_2^2 + \lambda \sum_{j=1}^{p} \text{pen}(\beta_{1,j}, \ldots, \beta_{T,j}), \tag{2}$$

with $\text{pen}(\mathbf{a}) = \max_{t \in [T]} |a_t|$ or $\text{pen}(\mathbf{a}) = ||a||_2$ for a vector $\mathbf{a} \in \mathbb{R}^T$. Under an appropriate choice of the penalty parameter λ, the estimator $\hat{\mathbf{B}}$ has many rows equal to zero, which correspond to irrelevant variables. However, solving (2) can be computationally prohibitive.

In the current work, we consider the ultra-high dimensional setting where the number of variables p is much higher than the sample size n, e.g. $p = \mathcal{O}(\exp(n^{\delta_p}))$ for a positive constant δ_p, but the regression coefficients $\boldsymbol{\beta}_t$ are sparse, i.e., for each t, there exist a small number of variables that are relevant to the output. Under the sparsity assumption, it is highly important to efficiently select the relevant variables in order to improve the accuracy of the estimation and prediction, and to facilitate the understanding of the underlying phenomenon for domain experts. In the seminal paper of Fan and Lv (2008), the concept of *sure screening* was introduced, a property of the variable selection procedure that keeps all the relevant variables with high probability. We show that the S-OMP, has the sure screening property for the multiple output regression problem in (1). To the best of our knowledge, this is the first attempt to analyze the sure screening property in the ultra-high dimensional space using the shared information from the multiple regression outputs.

The variable selection in the model (1) can be formalized in two ways: (1) the *union support* recovery of $\mathbf{B}$, where a subset of variables is selected that affect at least one output; (2) the *exact support* recovery of $\mathbf{B}$, where the exact positions of non-zero elements in $\mathbf{B}$ are estimated. We address the problem of the exact support recovery, which is of particular importance in problems like genome-wide association mapping (Kim and Xing, 2009) or biological network estimation (Peng et al., 2008), in two steps. In the first step, the S-OMP is used to screen the variables, i.e., select a subset of variables that contain all the true variables. In the second step, we use the adaptive Lasso (ALasso) (Zou, 2006) to further select a subset of screened variables for each task.

The model in (1) has been used in many different domains ranging from multivariate regression (Obozinski et al., 2009; Negahban and Wainwright, 2009) and sparse approximation (Tropp et al., 2006) to neural science (Liu et al., 2009), multi-task learning (Lounici et al., 2009; Argyriou et al., 2008) and biological network estimation (Peng et al., 2008). A number of authors has provided theoretical understanding of the estimation in the model using the convex program (2) to estimate $\hat{\mathbf{B}}$. Lounici et al. (2009) showed the benefits of the joint estimation, when there is a small set of variables common to all outputs and the number of outputs is large. Obozinski et al. (2009) and Negahban and Wainwright (2009) analyzed the consistent recovery of the union support. Negahban and Wainwright (2009) provided the analysis of the exact support recovery for a special case with two outputs.

The Orthogonal Matching Pursuit (OMP) has been analyzed before in the literature (see, e.g., Zhang, 2009; Lozano et al., 2009; Wang, 2009; Barron et al., 2008). In particular, our work should be contrasted to Wang (2009), which showed that the OMP has the sure screening property in a linear regression with a single output, and to the exact variable selection property of the OMP analyzed in Zhang (2009) and Lozano et al. (2009). The exact variable selection requires much stronger assumptions on the design, such as the irrepresentable condition, that are hard to satisfy in the ultra-high dimensional setting. On the other hand, the sure screening property can be shown to hold under much weaker assumptions.

In this paper, we make the following novel contributions: i) we prove that the S-OMP can be used for the ultra-high dimensional variable screening in multiple output regression problems and demonstrate its performance on extensive numerical studies; ii) we show that a two step procedure can be used to select exactly the relevant variables for each task; and iii) we prove that a modification of the BIC score (Chen and Chen, 2008) can be used to select the number of steps in the S-OMP.

2 Methodology

2.1 The model and notation

We will consider a slightly more general model

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \quad t = 1, \ldots, T \tag{3}$$

than the one given in (1). The model in (1) is a special case of the model in (3), with all the design matrices $\{\mathbf{X}_t\}_{t\in[T]}$ equal. Assume that for all $t \in [T]$, $\mathbf{X}_t \in \mathbb{R}^{n\times p}$. For the design $\mathbf{X}_t$, we denote $\mathbf{X}_{t,j}$ the j-th column, $\mathbf{x}_{t,i}$ the i-th row and $x_{t,ij}$ the element at (i,j). Denote $\boldsymbol{\Sigma}_t = \mathrm{Cov}(\mathbf{x}_{t,i})$. Without loss of generality, we assume that $\mathrm{Var}(y_{t,i}) = 1$, $\mathbb{E}(x_{t,ij}) = 0$ and $\mathrm{Var}(x_{t,ij}) = 1$. The noise $\boldsymbol{\epsilon}_t$ is zero mean and $\mathrm{Cov}(\boldsymbol{\epsilon}_t) = \sigma^2\mathbf{I}_{n\times n}$. We assume that the number of variables $p \gg n$ and that the regression coefficients $\boldsymbol{\beta}_t$ are jointly sparse. Let $\mathcal{M}_{*,t}$ denote the set of non-zero coefficients of $\boldsymbol{\beta}_t$ and $\mathcal{M}_* = \cup_{t=1}^T \mathcal{M}_{*,t}$ denote the set of all relevant variables. For an arbitrary set $\mathcal{M} = \{j_1, \ldots, j_k\}$, $\mathbf{X}_{t,\mathcal{M}}$ denotes the design with columns indexed by $\mathcal{M}$, $\mathbf{B}_{\mathcal{M}}$ denotes the rows of $\mathbf{B}$ indexed by $\mathcal{M}$ and $\mathbf{B}_j = (\boldsymbol{\beta}_{1,j}, \ldots, \boldsymbol{\beta}_{T,j})'$. The cardinality of the set $\mathcal{M}$ is denoted as $|\mathcal{M}|$. Let $s := |\mathcal{M}_*|$ denote the total number of relevant variables, so under the sparsity assumption we have $s < n$. For a square matrix $\mathbf{A}$, $\Lambda_{\min}(\mathbf{A})$ and $\Lambda_{\max}(\mathbf{A})$ are used to denote the minimum and the maximum eigenvalue, respectively. Lastly, we use $[p]$ to denote the set $\{1, \ldots, p\}$.

Before we continue, we give a few definitions that will facilitate the presentation of the algorithm and theoretical results.

Definition 1. *The union support recovery deals with estimation of $\mathcal{M}_*$, the set of all relevant variables.*

Definition 2. *The exact support recovery deals with estimation of $\{\mathcal{M}_{*,t}\}_{t\in[T]}$, the exact set of non-zero elements of $\mathbf{B}$.*

Definition 3. *An estimation procedure is said to have the sure screening property if it is able to find an estimator $\hat{\mathcal{M}}$ of the union support that satisfies $\mathbb{P}[\mathcal{M}_* \subseteq \hat{\mathcal{M}}] \to 1$ as $n \to \infty$.*

2.2 Simultaneous Orthogonal Matching Pursuit: Screening

The Simultaneous Orthogonal Matching Pursuit is outlined in Algorithm 1. Before describing the algorithm, we introduce some additional notation. For a model $\mathcal{M}$, let $\mathbf{H}_{t,\mathcal{M}}$ be the orthogonal projection onto $\mathrm{Span}(\mathbf{X}_{t,\mathcal{M}})$, i.e., $\mathbf{H}_{t,\mathcal{M}} = \mathbf{X}_{t,\mathcal{M}}(\mathbf{X}'_{t,\mathcal{M}}\mathbf{X}_{t,\mathcal{M}})^{-1}\mathbf{X}'_{t,\mathcal{M}}$, and define the residual sum of squares (RSS) as $\mathrm{RSS}(\mathcal{M}) = \sum_{t=1}^T \mathbf{y}'_t(\mathbf{I}_{n\times n} - \mathbf{H}_{t,\mathcal{M}})\mathbf{y}_t$.

The algorithm starts with an empty model $\mathcal{M}^{(0)} = \emptyset$. We recursively define the model $\mathcal{M}^{(k)}$ based on the model $\mathcal{M}^{(k-1)}$. The model $\mathcal{M}^{(k)}$ is obtained by adding a variable $\hat{j}_k$, which minimizes $\mathrm{RSS}(\mathcal{M}^{(k-1)} \cup j)$ over the set $[p]\backslash\mathcal{M}^{(k-1)}$, to the model $\mathcal{M}^{(k-1)}$. Repeating the algorithm for $n-1$ steps, a sequence of nested models $\{\mathcal{M}^{(k)}\}_{k=0}^{n-1}$ is obtained, with $\mathcal{M}^{(k)} = \{\hat{j}_1, \ldots, \hat{j}_k\}$.

Algorithm 1 Group Forward Regression

Input: Dataset $\{\mathbf{X}_t, \mathbf{y}_t\}_{t=1}^T$
Output: Sequence of selected models $\{\mathcal{M}^{(k)}\}_{k=0}^{n-1}$

1: Set $\mathcal{M}^{(0)} = \emptyset$
2: **for** $k = 1$ to $n-1$ **do**
3: **for** $j = 1$ to p **do**
4: $\tilde{\mathcal{M}}_j^{(k)} = \mathcal{M}^{(k-1)} \cup \{j\}$
5: $\mathbf{H}_{t,j} = \mathbf{X}_{t,\tilde{\mathcal{M}}_j^{(k)}}(\mathbf{X}'_{t,\tilde{\mathcal{M}}_j^{(k)}}\mathbf{X}_{t,\tilde{\mathcal{M}}_j^{(k)}})^{-1}\mathbf{X}'_{t,\tilde{\mathcal{M}}_j^{(k)}}$
6: $\mathrm{RSS}(\tilde{\mathcal{M}}_j^{(k)}) = \sum_{t=1}^T \mathbf{y}'_t(\mathbf{I}_{n\times n} - \mathbf{H}_{t,j})\mathbf{y}_t$
7: **end for**
8: $\hat{j}_k = \mathrm{argmin}_{j\in\{1,\ldots,p\}\backslash\mathcal{M}^{(k-1)}} \mathrm{RSS}(\tilde{\mathcal{M}}_j^{(k)})$
9: $\mathcal{M}^{(k)} = \mathcal{M}^{(k-1)} \cup \{\hat{j}_k\}$
10: **end for**

To practically select one of the models from $\{\mathcal{M}^{(k)}\}_{k=0}^{n-1}$, we minimize the modified BIC criterion (Chen and Chen, 2008), which is defined as

$$\mathrm{BIC}(\mathcal{M}) = \log\left(\frac{\mathrm{RSS}(\mathcal{M})}{nT}\right) + \frac{|\mathcal{M}|(\log(n) + 2\log(p))}{n} \tag{4}$$

with $|\mathcal{M}|$ denoting the number of elements of the set $\mathcal{M}$. Let $\hat{s} = \mathrm{argmin}_{k\in\{0,\ldots,n-1\}} \mathrm{BIC}(\mathcal{M}^{(k)})$, so that the selected model is $\mathcal{M}^{(\hat{s})}$. Observe that $\mathcal{M}^{(\hat{s})}$ estimates only the union support and that further subselection is needed to estimate the exact support.

Remark: The S-OMP algorithm is outlined only conceptually in this section. The steps 5 and 6 of the algorithm can be implemented efficiently using the progressive Cholesky decomposition see, e.g., Cotter et al. (1999).

2.3 Exact variable selection

After the dimensionality of the original problem has been reduced to the size of the model $\mathcal{M}^{(\hat{s})}$, which is less than the sample size n, one can address the problem of estimating the regression coefficients and recovering the exact support of $\mathbf{B}$ using a lower dimensional selection procedure. In this paper, we use the adaptive Lasso as a lower dimensional selection procedure, which was shown to have oracle properties Zou (2006). The ALasso solves the penalized least square problem

$$\hat{\boldsymbol{\beta}}_t = \underset{\boldsymbol{\beta}_t\in\mathbb{R}^{\hat{s}}}{\mathrm{argmin}} ||\mathbf{y}_t - \mathbf{X}_{t,\mathcal{M}^{(\hat{s})}}\boldsymbol{\beta}_t||_2^2 + \lambda \sum_{j\in\mathcal{M}^{(\hat{s})}} w_j|\beta_{t,j}|, \tag{5}$$

where $(w_j)_{j\in\mathcal{M}^{(\hat{s})}}$ is a vector of known weight and λ is a tuning parameter. Usually, the weights are defined

as $w_j = 1/|\hat{\beta}_{t,j}|$ for a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_t$. Observe that the adaptive Lasso in (5) is defined for each output separately, so that the exact support of $\mathbf{B}$ can be recovered under an assumption that the number of outputs does not diverge too quickly. We point out that solving the multi-task problem defined in (2) can be efficiently done on the reduced set of variables, but it is not obvious how to obtain the estimate of the exact support using (2). In Section 4, our numerical studies show that the ALasso applied to the reduced set of variables can be used to estimate the exact support of $\mathbf{B}$.

3 Theory

In this section we state conditions under which Algorithm 1 has the sure screening property. We also show that the model selected using the modified BIC criterion contains all the relevant variables.

3.1 Assumptions

Before we state the theorem characterizing the performance of the S-OMP, we give some technical conditions that are needed for our analysis.

A1: The random noise vectors $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T$ are independent Gaussian with zero mean and covariance matrix $\sigma^2 \mathbf{I}_{n\times n}$.

A2: Each row of the design matrix $\mathbf{X}_t$ is IID Gaussian with zero mean and covariance matrix $\boldsymbol{\Sigma}_t$. Furthermore, there exist two positive constants $0 < \phi_{\min} < \phi_{\max} < \infty$ such that

$$\phi_{\min} \leq \min_{t\in[T]} \Lambda_{\min}(\boldsymbol{\Sigma}_t) \leq \max_{t\in[T]} \Lambda_{\max}(\boldsymbol{\Sigma}_t) \leq \phi_{\max}. \quad (6)$$

A3: The true regression coefficients are bounded, i.e., there exists a positive constant $C_{\boldsymbol{\beta}}$ such that $||\boldsymbol{\beta}||_{2,1} \leq C_{\boldsymbol{\beta}}$. Define $\mathbf{B}_{\min}$ as a T-vector that satisfies $||\mathbf{B}_{\min}||_2^2 = \min_{j\in\mathcal{M}_*} \sum_{t\in[T]} \beta_{t,j}^2$, i.e., the norm of the vector $\mathbf{B}_{\min}$ lower bounds the norm of a row of $\mathbf{B}$ for any relevant variable. There exist positive constants $c_{\boldsymbol{\beta}}$ and $\delta_{\min}$ such that $T^{-1}||\mathbf{B}_{\min}||_2^2 \geq c_{\boldsymbol{\beta}} n^{-\delta_{\min}}$.

A4: There exist positive constants C_s, C_p, δ_s and δ_p such that $|\mathcal{M}_*| \leq C_s n^{\delta_s}$ and $\log(p) \leq C_p n^{\delta_p}$.

The normality condition **A1** is assumed here only to facilitate presentation of theoretical results, as is commonly assumed in literature, (e.g., Zhang and Huang, 2008; Fan and Lv, 2008). The normality assumption can be avoided at the cost of more technical proofs, e.g., Lounici et al. (2009), where the main technical difficulty is showing that the concentration properties still hold. Under the condition **A2** we will be able to show that the empirical covariance matrix satisfies the sparse eigenvalue condition (see Lemma 6) with probability tending to one. The assumption that the rows of the design are Gaussian can be easily relaxed to the case when the rows are sub-Gaussian, without any technical difficulties in proofs, since we would still obtain exponential bounds on the tail probabilities. The condition **A3** states that the regression coefficients are bounded, which is a technical condition likely to be satisfied in practice. Furthermore, it is assumed that the row norms of $\mathbf{B}_{\mathcal{M}_*}$ do not decay to zero too fast or, otherwise, they would not be distinguishable from noise. The condition is not too restrictive, e.g., if every non-zero coefficient is bounded away from zero by a constant, the condition **A3** is trivially satisfied with $\delta_{\min} = 0$. However, we allow for the coefficients of the relevant variables to get smaller as the sample size increases and still guarantee that the relevant variable will be identified. The condition **A4** sets the upper bound on the number of relevant variables and the total number of variables. While the total number of variables can diverge to infinity much faster than the sample size, the number of relevant variables needs to be smaller than the sample size. It can be seen from **A3** and **A4** that many outputs should share the same non-zero coefficients. Otherwise, some coefficients would be too weak to be detected.

3.2 The screening consistency

Theorem 4. *Assume the model in* (3) *and that the conditions* **A1**-**A4** *are satisfied. Furthermore, assume that*

$$\frac{n^{1-6\delta_s-6\delta_{\min}}}{\max\{\log(p), \log(T)\}} \to \infty, \ \textit{as } n \to \infty. \quad (7)$$

Then there exists a number $m^*_{\max} = m^*_{\max}(n)$, so that in $m^*_{\max}$ all the relevant variables are included in the model, i.e., as $n \to \infty$

$$\mathbb{P}[\mathcal{M}_* \subseteq \mathcal{M}^{(m^*_{\max})}] \geq 1 - C_1 \exp\left(-C_2 \frac{n^{1-6\delta_s-6\delta_{\min}}}{\max\{\log p, \log T\}}\right), \quad (8)$$

for some constants C_1, C_2. The exact value of $m^*_{\max}$ is given as

$$m^*_{\max} = \lfloor 2^4 \phi_{\min}^{-2} \phi_{\max} C_{\boldsymbol{\beta}}^2 C_s^2 c_{\boldsymbol{\beta}}^{-2} n^{2\delta_s+2\delta_{\min}} \rfloor. \quad (9)$$

Remarks: Under the assumptions of Theorem 4, $m^*_{\max} \leq n-1$, so that the procedure effectively reduces the dimensionality below the sample size. From the proof of the theorem, it is clear how multiple outputs help to identify the relevant variables. The crucial quantity in identifying all relevant variables is the

minimum non-zero row norm of $\mathbf{B}$, which allows us to identify weak variables if they are relevant for a large number of outputs even though individual coefficients may be small.

Proof. We outline the proof here, while the details are given in the supplementary materials. The proof uses ideas from Zhang (2009) and Wang (2009).

Under the assumptions of the theorem, the number of relevant variables s is relatively small compared to the sample size n. The proof strategy can be outlined as follows: i) we are going to show that, with high probability, at least one relevant variable is going to be identified within the following m^*_{one} steps, conditioning on the already selected variables $\mathcal{M}^{(k)}$ and this holds uniformly for all k; ii) we can conclude that all the relevant variables are going to be selected within $m^*_{\max} = sm^*_{\text{one}}$ steps. Exact values for m^*_{one} and $m^*_{\max}$ are given below. Without loss of generality, we analyze the first step of the algorithm, i.e., we show that the first relevant variable is going to be selected within the first m^*_{one} steps.

Assume that in the first $m^*_{\text{one}} - 1$ steps, there were no relevant variables selected. Assuming that the m^*_{one}-th selected variable is still an irrelevant one, we will arrive to a contradiction, which shows that at least one relevant variable has been selected in the first m^*_{one} steps. For any step k, the squared error reduction is given as $\Delta(k) := \text{RSS}(k-1) - \text{RSS}(k)$, that is

$$\Delta(k) = \sum_t ||\mathbf{H}^{(k)}_{t,\hat{j}_k}(\mathbf{I}_{n\times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}})\mathbf{y}_t||_2^2 \qquad (10)$$

with $\mathbf{H}^{(k)}_{t,j} = \mathbf{X}^{(k)}_{t,j}\mathbf{X}^{(k)'}_{t,j}||\mathbf{X}^{(k)}_{t,j}||^{-2}$ and $\mathbf{X}^{(k)}_{t,j} = (\mathbf{I}_{n\times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}})\mathbf{X}_{t,j}$. We are interested in the quantity $\sum_{k=1}^{m^*_{\text{one}}}\Delta(k)$, when all the selected variables $\hat{j}_k$ (see Algorithm 1) belong to $[p]\backslash\mathcal{M}_*$.

In what follows, we will derive a lower bound for $\Delta(k)$. We perform our analysis on the event $\mathcal{E}$ where $\hat{\mathbf{\Sigma}}$ satisfies the sparse eigenvalue condition in Lemma 6 with $m^*_{\max}$. From the definition of $\hat{j}_k$, $\Delta(k)$ is lower bounded as

$$\begin{aligned}\Delta(k) \geq &\max_{j\in\mathcal{M}_*}\sum_t ||\mathbf{H}^{(k)}_{t,j}(\mathbf{I}_{n\times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}})\mathbf{X}_{t,\mathcal{M}_*}\boldsymbol{\beta}_{t,\mathcal{M}_*}||_2^2 \\ &- \max_{j\in\mathcal{M}_*}\sum_t ||\mathbf{H}^{(k)}_{t,j}(\mathbf{I}_{n\times n} - \mathbf{H}_{t,\mathcal{M}^{(k)}})\boldsymbol{\epsilon}_t||_2^2 \\ &= (I) - (II).\end{aligned}$$

We deal with these two terms separately. One can show that

$$\begin{aligned}(I) &\geq 2^{-3}\phi^2_{\min}\phi^{-1}_{\max}C_{\boldsymbol{\beta}}^{-2}ns^{-1}T^{-1}||\mathbf{B}_{\min}||_2^4 \\ &\geq 2^{-3}\phi^2_{\min}\phi^{-1}_{\max}C_{\boldsymbol{\beta}}^{-2}C_s^{-1}n^{1-\delta_s}T^{-1}||\mathbf{B}_{\min}||_2^4,\end{aligned}$$

and that

$$\begin{aligned}(II) &\leq 2^3\phi^{-1}_{\min}\phi_{\max}T(m^*_{\max}+2)\log p \\ &\leq 9\phi^{-1}_{\min}\phi_{\max}C_p n^{\delta_p}Tm^*_{\max},\end{aligned}$$

with high probability. Combining (I) and (II), one obtains a lower bound on $\Delta(k)$ that does not depend on k, so that

$$n^{-1}T^{-1}\sum_{t\in[T]}||\mathbf{y}_t||_2^2 \geq 2(1 - Cn^{3\delta_s+4\delta_{\min}+\delta_p-1}) \to 2,$$

under the conditions of the theorem. We have arrived to a contradiction, since under the assumptions $\text{Var}(y_{t,i}) = 1$ and by the weak law of large numbers $n^{-1}T^{-1}\sum_{t\in[T]}||\mathbf{y}_t||_2^2 \to 1$ in probability. Therefore, at least one relevant variable will be selected in m^*_{one} steps.

To complete the proof, one needs to lower bound the probability of the event $\mathcal{E}$. For that we can invoke Lemma 6. □

The following theorem guarantees that the model $\mathcal{M}^{(\hat{s})}$, selected with the modified BIC criterion, is screening consistent.

Theorem 5. *Assume that the conditions of Theorem 4 are satisfied. Let*

$$\hat{s} = \operatorname*{argmin}_{k\in\{0,\ldots,n-1\}} \text{BIC}(\mathcal{M}^{(k)}) \qquad (11)$$

be the index of the model selected by optimizing the modified BIC criterion. Then, as $n \to \infty$

$$\mathbb{P}[\mathcal{M}_* \subseteq \mathcal{M}^{(\hat{s})}] \to 1. \qquad (12)$$

Proof. We outline the proof here, while the details are given in the supplementary materials. To prove the theorem, we use the same strategy as in Wang (2009). From Theorem 4 we have that $\mathbb{P}[\exists k \in \{0,\ldots,n-1\} : \mathcal{M}_* \subseteq \mathcal{M}^{(k)}] \to 1$, so $k_{\min} := \min_{k\in\{0,\ldots,n-1\}}\{k : \mathcal{M}_* \subseteq \mathcal{M}^{(k)}\}$ is well defined and $k_{\min} \leq m^*_{\max}$, for $m^*_{\max}$ defined in (9). We show that

$$\mathbb{P}[\min_{k\in\{0,\ldots,k_{\min}-1\}}(\text{BIC}(\mathcal{M}^{(k)}) - \text{BIC}(\mathcal{M}^{(k+1)})) > 0] \to 1, \qquad (13)$$

so that $\mathbb{P}[\hat{s} < k_{\min}] \to 0$ as $n \to \infty$. We proceed by lower bounding the difference in the BIC scores as

$$\begin{aligned}&\text{BIC}(\mathcal{M}^{(k)}) - \text{BIC}(\mathcal{M}^{(k+1)}) \\ &\quad\geq \log\left(1 + \frac{\text{RSS}(\mathcal{M}^{(k)}) - \text{RSS}(\mathcal{M}^{(k+1)})}{\text{RSS}(\mathcal{M}^{(k+1)})}\right) \qquad (14) \\ &\quad\quad - 3n^{-1}\log(p),\end{aligned}$$

where we have assumed $p > n$. Define the event $\mathcal{A} := \{n^{-1}T^{-1}\sum_{t\in[T]}||\mathbf{y}_t||_2^2 \leq 2\}$. Note that

$\mathrm{RSS}(\mathcal{M}^{(k+1)}) \leq \sum_{t\in[T]} ||\mathbf{y}_t||_2^2$, so on the event $\mathcal{A}$ the difference in the BIC scores is lower bounded as

$$\log(1 + 2n^{-1}T^{-1}\Delta(k)) - 3n^{-1}\log(p), \qquad (15)$$

where $\Delta(k)$ is defined in (10). Using the fact that $\log(1+x) \geq \min(\log(2), 2^{-1}x)$ and the lower bound on $\Delta(k)$ from the proof of Theorem 4, we have

$$\begin{aligned} &\mathrm{BIC}(\mathcal{M}^{(k)}) - \mathrm{BIC}(\mathcal{M}^{(k+1)}) \\ &\quad \geq \min(\log 2, Cn^{-\delta_s - 2\delta_{\min}}) - 3n^{-1}\log p, \end{aligned} \qquad (16)$$

for some positive constant C. It is easy to check that $\log 2 - 3n^{-1}\log p > 0$ and $Cn^{-\delta_s-2\delta_{\min}} - 3n^{-1}\log p > 0$ under the conditions of the theorem. The lower bound in (16) is uniform for $k \in \{0, \ldots, k_{\min}\}$, so the proof is complete if we show that $\mathbb{P}[\mathcal{A}] \to 1$. But this easily follows from the tail bounds on the central chi-squared random variable. □

4 Simulation studies

We conduct a number of numerical studies to evaluate the finite sample performance of the S-OMP. We consider three procedures that perform estimation on individuals outputs: Sure Independence Screening (SIS) and Iterative SIS (ISIS) (Fan and Lv, 2008), and the OMP, for comparison purposes. The evaluation is done on the model in (1). SIS and ISIS are used to select a subset of variables and then the ALasso is used to further refine the selection. We denote this combination as SIS-ALasso and ISIS-ALasso. The size of the model selected by SIS is fixed as $n-1$, while the ISIS selects $\lfloor n/\log(n) \rfloor$ variables in each of the $\lfloor \log(n) - 1 \rfloor$ iterations. From the screened variables, the final model is selected using the ALasso, together with the BIC criterion (4) to select the penalty parameter λ. We use the OMP without further refinement using the ALasso, since it was observed from the numerical studies in Wang (2009) that the combination does not gain much improvement. The S-OMP is used to reduce the dimensionality below the sample size jointly using the regression outputs. Next, the ALasso is used on each of the outputs to further perform the estimation. This combination is denoted SOMP-ALasso.

Let $\hat{\mathbf{B}} = [\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_T] \in \mathbb{R}^{p\times T}$ be an estimate obtained by one of the estimation procedures. We evaluate the performance averaged over 200 simulation runs. Let $\hat{\mathbb{E}}_n$ denote the empirical average over the simulation runs. We measure the size of the union support $\hat{S} = S(\hat{\mathbf{B}}) := \{j \in [p] : ||\hat{\mathbf{B}}_j||_2^2 > 0\}$. Next, we estimate the probability that the screening property is satisfied $\hat{\mathbb{E}}_n[\mathbb{1}\{\mathcal{M}_* \subseteq S(\hat{\mathbf{B}})\}]$, which we call coverage probability. The fraction of incorrect zeros is measured as $s^{-1}\hat{\mathbb{E}}_n[|S(\hat{\mathbf{B}})^C \cap \mathcal{M}_*|]$. Similar quantities are defined for the exact support recovery. The following simulation studies are used to comparatively access the numerical performance of the procedures. Additional simulations are given in the Supplementary material.

Simulation 1: The following toy model is based on the simulation I in Fan and Lv (2008) with $(n, p, s, T) = (400, 20000, 18, 500)$. Each $\mathbf{x}_i$ is drawn independently from a standard multivariate normal distribution, so that the variables are mutually independent. For $j \in [s]$ and $t \in [T]$, the non-zero coefficients of $\mathbf{B}$ are given as $\beta_{t,j} = (-1)^u(4n^{-1/2}\log n + |z|)$, where $u \sim$ Bernoulli(0.4) and $z \sim \mathcal{N}(0,1)$. The number of non-zero elements in $\mathbf{B}_j$ is given as a parameter $T_{\mathrm{non-zero}} = 300$, i.e., a variable is shared across 300 outputs. The positions of non-zero elements are chosen uniformly at random from $[T]$. The noise is Gaussian with the standard deviation σ set to control the signal-to-noise ratio (SNR). SNR is defined as $\mathrm{Var}(\mathbf{x}\boldsymbol{\beta})/\mathrm{Var}(\boldsymbol{\epsilon})$ and we set it to $\mathrm{SNR} = 5$.

Simulation 2: The following scenario is used to evaluate the performance of the methods as the number of non-zero elements in a row of $\mathbf{B}$ varies. We set $(n, p, s) = (100, 500, 10)$ and the number of outputs $T = 1000$. We vary $T_{\mathrm{non-zero}} \in \{0.8T, 0.2T\}$. $\mathbf{x}_i$ and $\mathbf{B}$ are given as in Simulation 1, i.e., $\mathbf{x}_i$ is drawn from a multivariate standard normal distribution and the non-zero coefficients $\mathbf{B}$ are given as $\beta_{t,j} = (-1)^u(4n^{-1/2}\log n + |z|)$, where $u \sim$ Bernoulli(0.4) and $z \sim \mathcal{N}(0,1)$. The noise is Gaussian, with the standard deviation defined through the $\mathrm{SNR} = 5$.

Simulation 3: The following model is borrowed from Wang (2009). We assume a correlation structure between variables given as $\mathrm{Var}(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) = \rho^{|j_1 - j_2|}$, where $\rho = 0.5$. This correlation structure appears naturally among ordered variables. We set $(n, p, s, T) = (100, 5000, 3, 150)$ and $T_{\mathrm{non-zero}} = 80$. The relevant variables are at positions $(1, 4, 7)$ and non-zero coefficients are given as $3, 1.5$ and 2 respectively. The SNR is set to 5.

Simulation 4: The following model assumes a block compound correlation structure. For a parameter ρ, the correlation between two variables $\mathbf{X}_{j_1}$ and $\mathbf{X}_{j_2}$ is given as ρ, ρ^2 or ρ^3 when $|j_1 - j_2| \leq 10$, $|j_1 - j_2| \in (10, 20]$ or $|j_1 - j_2| \in (20, 30]$ and it is set to 0 otherwise. We set $(n, p, s, T) = (150, 4000, 8, 150)$, $T_{\mathrm{non-zero}} = 80$ and the parameter $\rho = 0.5$. The relevant variables are located at positions 1, 11, 21, 31, 41, 51, 61, 71 and 81, so that each block of highly correlated variables has exactly one relevant variable. The values of relevant coefficients is given as in Simulation 1. The noise is Gaussian and the SNR is set to 5.

Simulation 5: This model represents a difficult setting. It is modified from Wang (2009). We set $(n, p, s, T) =$

Table 1: Simulation 1 $(n, p, s, T) = (400, 20000, 18, 500)$ $T_{\text{non-zero}} = 300$, SNR $= 5$

Method Name	Coverage (%)	Incorrect Zeros(%)	Size of Support
	Union Support		
SIS-ALASSO	100.0	0.0	21.1
ISIS-ALASSO	100.0	0.0	20.8
OMP	100.0	0.0	23.0
S-OMP	100.0	0.0	18.0
S-OMP-ALASSO	100.0	0.0	18.0
	Exact Support		
SIS-ALASSO	24.0	0.0	5400.9
ISIS-ALASSO	99.0	0.0	5402.8
OMP	100.0	0.0	5405.0
S-OMP-ALASSO	100.0	0.0	5400.0

Table 2: Simulation 2 $(n, p, s, T) = (100, 500, 10, 1000)$ $T_{\text{non-zero}} = 800$, SNR $= 5$

Method Name	Coverage (%)	Incorrect Zeros(%)	Size of Support
	Union Support		
SIS-ALASSO	100.0	0.0	11.9
ISIS-ALASSO	100.0	0.0	11.7
OMP	100.0	0.0	33.0
S-OMP	100.0	0.0	10.0
S-OMP-ALASSO	100.0	0.0	10.0
	Exact Support		
SIS-ALASSO	0.0	65.5	2759.0
ISIS-ALASSO	0.0	62.7	2984.0
OMP	100.0	0.0	8023.1
S-OMP-ALASSO	0.0	48.1	4152.9

Table 3: Simulation 2 $(n, p, s, T) = (100, 500, 10, 1000)$ $T_{\text{non-zero}} = 200$, SNR $= 5$

Method Name	Coverage (%)	Incorrect Zeros(%)	Size of Support
	Union Support		
SIS-ALASSO	100.0	0.0	10.0
ISIS-ALASSO	100.0	0.0	10.0
OMP	100.0	0.0	139.6
S-OMP	100.0	0.0	10.0
S-OMP-ALASSO	100.0	0.0	10.0
	Exact Support		
SIS-ALASSO	100.0	0.0	2000.0
ISIS-ALASSO	100.0	0.0	2000.0
OMP	100.0	0.0	2131.6
S-OMP-ALASSO	100.0	0.0	2000.0

$(200, 10000, 5, 500)$, $T_{\text{non-zero}} = 250$. For $j \in [s]$ and $t \in [T]$, the non-zero elements equal $\beta_{t,j} = 2j$. Each row of $\mathbf{X}$ is generated as follows. Draw independently $\mathbf{z}_i$ and $\mathbf{z}'_i$ from a p-dimensional standard multivariate normal distribution. Now, $x_{ij} = (z_{ij} + z'_{ij})/\sqrt{(2)}$ for $j \in [s]$ and $x_{ij} = (z_{ij} + \sum_{j' \in [s]} z_{ij'})/2$ for $j \in [p] \backslash [s]$. Now, $\text{Corr}(x_{i,1}, y_{t,i})$ is much smaller then $\text{Corr}(x_{i,j}, y_{t,i})$ for $j \in [p] \backslash [s]$, so that it becomes difficult to select variable 1. The noise is Gaussian with standard deviation $\sigma = 1.5$.

The simulation results are summarized in Tables 1-3, from which we have the following observations. When the variables are independent, it is relatively easy for all methods to cover the union support and the exact support. In this setting, it was previously shown in Fan and Lv (2008) and Wang (2009) that procedures which estimate independently from different outputs do well. Additional simulations (given in the Supplementary materials) suggest that the joint estimation with S-OMP-ALasso has some advantages when the SNR is low and when a variable is relevant for a large number of outputs. On the other hand, as the number of non-zero elements $T_{\text{non-zero}}$ in a row of $\mathbf{B}$ decreases, SIS-ALasso, ISIS-ALasso and OMP start to perform comparably and sometimes even better. The theory suggests (but does not prove) that the crucial parameter for a variable to be selected is the minimum non-zero row norm in the case of the S-OMP and the minimum absolute value of a non-zero regression coefficient in the other cases. When the correlation between variables start to increase, as in the simulations 3, 4 and 5, the S-OMP outperforms the other procedures, which tend to under-fit the model.

5 Conclusions

In this work, we analyze the Simultaneous Orthogonal Matching Pursuit as a method for variable selection in an ultra-high dimensional space. We prove that the S-OMP is screening consistent and provide a practical way to select the number of steps in the procedure using the modified Bayesian information criterion. Our limited numerical experience shows that the method performs well in practice and that the joint estimation from multiple outputs often outperforms methods that use one regression output at the time.

Acknowledgements

EPX is supported by grants NSF DBI-0546594, NSF DBI-0640543 and an Alfred P. Sloan Research Fellowship.

Appendix

The following Lemma (proved in supplement) is used to prove Theorem 4.

Lemma 6. *Let* $\mathbf{x} \sim \mathcal{N}(0, \mathbf{\Sigma})$ *and* $\hat{\mathbf{\Sigma}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}'_i$ *be the empirical estimate from n independent realizations of* $\mathbf{x}$. *Denote* $\mathbf{\Sigma} = [\sigma_{ab}]$ *and* $\hat{\mathbf{\Sigma}} = [\hat{\sigma}_{ab}]$. *Assume* $\phi_{min} \leq \Lambda_{\min}(\mathbf{\Sigma}) \leq \Lambda_{\max}(\mathbf{\Sigma}) \leq \phi_{\max}$. *Then as* $n \to \infty$,

$$\mathbb{P}[\max_{\mathcal{M} \subseteq [p], |\mathcal{M}| < s} \Lambda_{\max}(\hat{\mathbf{\Sigma}}_{\mathcal{M}}) \geq 2\phi_{\max}] = \mathcal{O}(\exp(-\frac{n}{s^3 \log p}))$$

and

$$\mathbb{P}[\min_{\mathcal{M} \subseteq [p], |\mathcal{M}| < s} \Lambda_{\min}(\hat{\mathbf{\Sigma}}_{\mathcal{M}}) \leq \phi_{\min}/2] = \mathcal{O}(\exp(-\frac{n}{s^3 \log p})).$$

Table 4: Simulation 3 $(n, p, s, T) = (100, 5000, 3, 150)$ $T_{\text{non-zero}} = 80$, SNR $= 5$, $\rho = 0.5$

Method Name	Coverage (%)	Incorrect Zeros(%)	Size of Support
	Union Support		
SIS-ALASSO	100.0	0.0	3.0
ISIS-ALASSO	100.0	0.0	3.0
OMP	100.0	0.0	19.6
S-OMP	100.0	0.0	3.00
S-OMP-ALASSO	100.0	0.0	3.00
	Exact Support		
SIS-ALASSO	60.0	0.2	239.5
ISIS-ALASSO	84.0	0.1	239.8
OMP	100.0	0.0	256.6
S-OMP-ALASSO	100.0	0.0	240.0

Table 5: Simulation 4 $(n, p, s, T) = (150, 4000, 8, 150)$ $T_{\text{non-zero}} = 80$, SNR $= 5$, $\rho = 0.5$

Method Name	Coverage (%)	Incorrect Zeros(%)	Size of Support
	Union Support		
SIS-ALASSO	100.0	0.0	8.4
ISIS-ALASSO	100.0	0.0	8.9
OMP	100.0	0.0	12.3
S-OMP	100.0	0.0	8.0
S-OMP-ALASSO	100.0	0.0	8.0
	Exact Support		
SIS-ALASSO	0.0	23.8	487.8
ISIS-ALASSO	0.0	7.6	592.5
OMP	99.0	0.0	644.4
S-OMP-ALASSO	7.0	2.8	622.2

Table 6: Simulation 5 $(n, p, s, T) = (200, 10000, 5, 500)$ $T_{\text{non-zero}} = 250$, $\sigma = 1.5$

Method Name	Coverage (%)	Incorrect Zeros(%)	Size of Support
	Union Support		
SIS-ALASSO	100.0	0.0	31.5
ISIS-ALASSO	100.0	0.0	14.3
OMP	100.0	0.0	30.8
S-OMP	100.0	0.0	5.8
S-OMP-ALASSO	100.0	0.0	5.0
	Exact Support		
SIS-ALASSO	0.0	45.9	768.9
ISIS-ALASSO	0.0	5.3	1200.7
OMP	100.0	0.0	1287.6
S-OMP-ALASSO	100.0	0.0	1250.0

References

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

Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A. Devore. Approximation and learning by greedy algorithms. *The Annals of Statistics*, 36(1):64–94, 2008.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

S.F. Cotter, R. Adler, R.D. Rao, and K. Kreutz-Delgado. Forward sequential algorithms for best basis selection. *Vision, Image and Signal Processing, IEE Proceedings* -, 146(5):235–244, 1999. ISSN 1350-245X.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal Of The Royal Statistical Society Series B*, 70(5):849–911, 2008.

Seyoung Kim and Eric P. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet*, 5(8):e1000587, 2009.

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, June 2009.

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*, 2009.

Karim Lounici, Massimiliano Pontil, Alexandre B. Tsybakov, and Sara van de Geer. Taking advantage of sparsity in Multi-Task learning. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

A. Lozano, G. Swirszcz, and N. Abe. Grouped orthogonal matching pursuit for variable selection and prediction. In *NIPS 22*. 2009.

Sahand Negahban and Martin Wainwright. Phase transitions for high-dimensional joint support recovery. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1161–1168. 2009.

Guillaume Obozinski, Martin Wainwright, and Michael Jordan. High-dimensional support union recovery in multivariate regression. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1217–1224. 2009.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, 2008.

Joel A. Tropp, Anna C. Gilbert, and Martin J. Strauss. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing*, 86(3):572 – 588, 2006. ISSN 0165-1684. Sparse Approximations in Signal and Image Processing.

Hansheng Wang. Forward Regression for Ultra-High Dimensional Variable Screening. *SSRN eLibrary*, 2009.

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

J. Zhang. *A probabilistic framework for multitask learning (Technical Report CMU-LTI-06-006)*. PhD thesis, Ph. D. thesis, Carnegie Mellon University, 2006.

Tong Zhang. On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, 10(Mar):555–568, 2009.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006.