Marginal Regression For Multitask Learning

Mladen Kolar
Machine Learning Department
Carnegie Mellon University
mladenk@cs.cmu.edu

Han Liu
Biostatistics
Johns Hopkins University
hanliu@jhsph.edu

Abstract

Variable selection is an important and practical problem that arises in analysis of many high-dimensional datasets. Convex optimization procedures that arise from relaxing the NP-hard subset selection procedure, e.g., the Lasso or Dantzig selector, have become the focus of intense theoretical investigations. Although many efficient algorithms exist that solve these problems, finding a solution when the number of variables is large, e.g., several hundreds of thousands in problems arising in genome-wide association analysis, is still computationally challenging. A practical solution for these high-dimensional problems is marginal regression, where the output is regressed on each variable separately. We investigate theoretical properties of marginal regression in a multitask framework. Our contribution include: i) sharp analysis for marginal regression in a single task setting with random design, ii) sufficient conditions for the multitask screening to select the relevant variables, iii) a lower bound on the Hamming distance convergence for multitask variable selection problems. A simulation study further demonstrates the performance of marginal regression.

1 Introduction

Recent technological advances are allowing scientists in a variety of disciplines to collect data of unprecedented size and complexity. Examples include data from biology, genetics, astronomy, brain imaging and high frequency trading. These applications are often characterized by a large number of variables p, which can be much larger than the number of observations n, and are currently driving the development of statistical and machine learning procedures. The sparsity assumption has been recognized to play a critical role in effective high-dimensional inference in classification and regression problems, that is, the statistical inference is possible in under-determined problems under the assumption that only a few variables contribute to the response. Therefore, the variable selection is of fundamental importance in high-dimensional problems.

Consider a regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

with response $\mathbf{y} = (y_1, \ldots, y_m)'$, $m \times p$ design matrix $\mathbf{X}$, noise vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)'$ and coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. For simplicity of presentation, we assume that $m = 2n$ and use the first n samples to estimate the parameters and use remaining parameters to optimally select the tuning parameters. The high dimensional setting assumes $p \gg n$ and the sparsity assumption roughly states that the coefficient vector $\boldsymbol{\beta}$ has a few non-zero components or that it can be well approximated by such a vector. In the context of linear regression, there has been a lot of recent work focusing on variable selection under the sparsity assumption, such as, Tibshirani (1996), Fan and Li (2001), Candes and Tao (2007), Zou (2006), Zou and Li (2008), Zhang (2010), Cai et al. (2010), Chen et al. (1999), Donoho (2006), Wainwright (2009), Zhao and Yu (2006), and Meinshausen and Yu (2009), to name a few. Many of these methods are based on constrained or penalized optimization procedures in which solutions are biased to have many zero coefficients. One of the main tools for variable selection in a regression model is the Lasso estimator defined by

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \tag{2}$$

where $\lambda \geq 0$ is a user defined regularization parameter. Theoretical properties of the estimator $\widehat{\beta}$ are

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

now well understood and the optimization problem (2) can be efficiently solved for medium sized problems. However, finding a solution in problems involving hundreds of thousands variables, which commonly arise in genome-wide association mapping problems, still remains a computationally challenging task, even when many variables can be pruned using rules based on the KKT conditions (El Ghaoui et al., 2010; Tibshirani et al., 2010).

One computationally superior alternative to the Lasso is marginal regression, also known as correlation learning, marginal learning and sure screening. This is a very old and simple procedure, which has recently gained popularity due to its desirable properties in high-dimensional setting (Wasserman and Roeder, 2009; Fan and Lv, 2008; Fan et al., 2009, 2011). See also Kerkyacharian et al. (2009) and Alquier (2008) for related procedures. Marginal regression is based on regressing the response variable on each variable separately

$$\widehat{\mu}_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y}, \tag{3}$$

where $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})'$. Next, the values $\{|\widehat{\mu}_j|\}$ are sorted in decreasing order, with $\{\widehat{r}_j\}$ denoting the ranks, and the set of estimated variables is

$$\widehat{S}(k) := \{1 \le j \le p : \widehat{r}_j \le k\}, \qquad 1 \le k \le p. \tag{4}$$

Note that in Eq. (3) we use the first n samples only to compute $\widehat{\mu}_j$. Under a condition, related to the faithfulness conditions used in causal literature (Robins et al., 2003; Spirtes et al., 2000), it can be shown that the set $\widehat{S}(k)$ correctly estimates the relevant variables $S := \{1 \le j \le p : \beta_j \neq 0\}$, see Wasserman and Roeder (2009). The following result provides the conditions under which the exact variable selection is possible if the size of the support $s := |S|$ is known.

Theorem 1. *Consider the regression model in* (1) *with* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, $\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, *and* $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{X}$ *independent of* $\boldsymbol{\epsilon}$. *Assume that*

$$\max_{j \in S^C} |\mathbf{\Sigma}_{jS}\boldsymbol{\beta}_S| + \gamma_n(p, s, \boldsymbol{\beta}, \mathbf{\Sigma}, \delta) < \min_{j \in S} |\mathbf{\Sigma}_{jS}\boldsymbol{\beta}_S| \tag{5}$$

with $\gamma_n = \mathcal{O}(\sqrt{\log(p-s)/n})$, *then*

$$\mathbb{P}[\widehat{S}(s) = S] \ge 1 - \delta.$$

The above theorem is based on the asymptotic result in Wasserman and Roeder (2009). We provide a finite sample analysis and explicit constants for the term $\gamma_n(p, s, \boldsymbol{\beta}, \mathbf{\Sigma}, \delta)$ in Appendix. A condition like the one in Eq. (5) is essentially unavoidable for marginal regression, since it can be seen that in the noiseless setting ($\boldsymbol{\epsilon} = \mathbf{0}$) the condition (5) with $\gamma_n = 0$ is necessary and sufficient for successful recovery. See Genovese et al. (2009) for discussion of cases where the faithfulness condition is weaker than the irrepresentable condition, which is necessary and sufficient for exact recovery of the support using the Lasso (Zhao and Yu, 2006; Wainwright, 2009).

Besides computational simplicity, another practical advantage of marginal regression is that the number of relevant variables s can be estimated from data efficiently as we show below. This corresponds to choosing the tuning parameter λ in the Lasso problem (2) from data. To estimate the number of relevant variables, we will use the samples indexed by $\{n+1, \ldots, 2n\}$, which are independent from those used to estimate $\{\widehat{\mu}_j\}_j$. For a fixed $1 \le k \le p$, let j_k denote the index of the variable for which $\widehat{r}_{j_k} = k$. Let $\widehat{V}_n(k) = \text{span}\{\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_k}\}$ be the linear space spanned by k variables whose empirical correlation with the response is the highest, and let $\widehat{\mathbf{H}}(k)$ be the projection matrix from $\mathbb{R}^n$ to $\widehat{V}_n(k)$. Note that $\mathbf{X}_{j_k} = (x_{n+1,j_k}, \ldots, x_{2n,j_k})$. Define

$$\widehat{\xi}_n(k) := ||(\widehat{\mathbf{H}}(k+1) - \widehat{\mathbf{H}}(k))\mathbf{y}||_2^2, \quad 1 \le k \le p-1, \tag{6}$$

which is then used to estimate the number of relevant variables as

$$\widehat{s}_n = \max\{1 \le k \le p-1 : \widehat{\xi}_n(k) \le 2\sigma^2 \log\tfrac{4n}{\delta}\} + 1. \tag{7}$$

Using an independent sample to select the number of relevant variables is needed so that the projection matrix is independent of the noise $\boldsymbol{\epsilon}$. With these definitions, we have the following result.

Theorem 2. *Assume that the conditions of Theorem 1 are satisfied. Furthermore, assume that*

$$\min_{j \in S} |\beta_j| = \Omega(\sqrt{\log n}).$$

Then $\mathbb{P}[\widehat{S}(\widehat{s}_n) = S] \xrightarrow{n \to \infty} 1$.

The above results builds on Theorem 3 in Genovese et al. (2009). A full statement of the theorem provides a finite sample result for a random design regression model is proven in Appendix.

Motivated by successful applications to variable selection in single task problems, we study properties of marginal regression in a multitask setting. In a number of applications, ranging from genome-wide association studies (Kim and Xing, 2009) to cognitive neuroscience (Liu et al., 2009), it has been observed that learning from related tasks jointly improves performance over procedures that learn from each task independently. This has sparked a lot of interest in machine learning and statistics community, see e.g. Turlach et al. (2005), Zou and Yuan (2008), Obozinski et al. (2011), Lounici et al. (2009), Liu et al.

(2009), Kolar and Xing (2010), Lounici et al. (2010), Kolar et al. (2011) and references therein. Section 2 provides sufficient conditions for marginal regression to exactly select relevant variables in a multitask setting. We provide versions of Theorem 1 and Theorem 2 for the multitask regression problem given in (8) below. Improvements using the multitask learning are illustrated on a model with an orthogonal design. Section 3 analyzes the recovery of the relevant variables under the Hamming distance. A universal lower bound on the Hamming distance between $\widehat{S}$ and S is provided. Some illustrative simulations are given in Section 4. All proofs are deferred to Appendix.

2 Multitask Learning with Marginal Regression

In this section, we analyze properties of marginal regression in a multitask setting. We will consider the following multitask regression model

$$\mathbf{y}_t = \mathbf{X}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \qquad t = 1, \dots, T \tag{8}$$

where $\mathbf{y}_t, \boldsymbol{\epsilon} \in \mathbb{R}^m$ and $\mathbf{X} \in \mathbb{R}^{m \times p}$. Again, we assume that $m = 2n$ and use half of the samples to rank the variables and the other half to select the correct number of relevant variables. The subscript t indexes tasks and $\boldsymbol{\beta}_t \in \mathbb{R}^p$ is the unknown regression coefficient for the t-th task. We assume that there is a shared design matrix $\mathbf{X}$ for all tasks, a situation that arises, for example, in genome-wide association studies. Alternatively, one can have one design matrix $\mathbf{X}_t$ for each task. We assume that the regression coefficients are jointly sparse. Let $S_t := \{1 \le j \le p : \beta_{tj} \neq 0\}$ be the set of relevant variables for the t-th task and let $S = \cup_t S_t$ be the set of all relevant variables. Under the joint sparsity assumption $s := |S| \ll n$.

To perform marginal regression in the multitask setting, one computes correlation between each variable and each task using the first half of the samples

$$\widehat{\mu}_{tj} = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y}_t, \tag{9}$$

for each $t = 1, \dots, T,\ j = 1, \dots, p$. Let $\Phi : \mathbb{R}^T \mapsto \mathbb{R}_+$ be a scoring function, which is used to sort the values $\{\Phi(\{\widehat{\mu}_{tj}\}_t)\}_j$ in decreasing order. Let $\{\widehat{r}_{\Phi,j}\}$ denote the rank of variable j in the ordering, then the set of estimated variables is

$$\widehat{S}_\Phi(k) := \{1 \le j \le p : \widehat{r}_{\Phi,j} \le k\}, \qquad 1 \le k \le p. \tag{10}$$

For concreteness, we will use the norm $||\cdot||_1, ||\cdot||_2$ and $||\cdot||_\infty$ as our scoring functions and denote the sets of estimated variables $\widehat{S}_{\ell_1}(k)$, $\widehat{S}_{\ell_2}(k)$ and $\widehat{S}_{\ell_\infty}(k)$ respectively.

With the notation introduced, we focus on providing conditions for marginal regression to exactly select the relevant variables S. We start our analysis in the fixed design setting. Let $\boldsymbol{\Sigma} = n^{-1}\mathbf{X}'\mathbf{X}$ and assume that the variables are standardized to have zero mean and unit variance, so that the diagonal elements of $\boldsymbol{\Sigma}$ are equal to 1. Now it simply follows from (9) that

$$\widehat{\mu}_{tj} = n^{-1}\mathbf{X}_j'\mathbf{y}_t = \boldsymbol{\Sigma}_{jS_t}\boldsymbol{\beta}_{tS_t} + n^{-1}\mathbf{X}_j'\boldsymbol{\epsilon}_t.$$

In order to show that marginal regression exactly recovers the set of relevant variables, we need to have

$$\max_{j \in S^C} \Phi(\{\widehat{\mu}_{tj}\}_t) \le \min_{j \in S} \Phi(\{\widehat{\mu}_{tj}\}_t). \tag{11}$$

It is easy to see that (11) is necessary for exact recovery. The following theorem provides sufficient conditions for (11) to hold.

Theorem 3. *Consider the model* (8) *with* $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ *and* $\sigma > 0$ *known. The following three claims hold: i) Define* $\nu_j = \sigma^{-2}n\sum_{t=1}^T(\boldsymbol{\Sigma}_{jS_t}\boldsymbol{\beta}_{tS_t})^2$. *If*

$$\begin{aligned}\max_{j \in S^C} \nu_j + 2\log\frac{2(p-s)}{\delta} \\ + \max_{j \in S} 2\sqrt{(T + 2\nu_j)\log\frac{2s}{\delta}} \\ + \max_{j \in S^C} 2\sqrt{(T + 2\nu_j)\log\frac{2(p-s)}{\delta}} \\ \le \min_{j \in S} \nu_j\end{aligned} \tag{12}$$

then $\mathbb{P}[\widehat{S}_{\ell_2}(s) = S] \ge 1 - \delta$. *ii) If*

$$\begin{aligned}&\max_{j \in S^C} \sum_{t=1}^T |\boldsymbol{\Sigma}_{jS_t}\boldsymbol{\beta}_{tS_t}| \\ &+ n^{-1/2}\sigma\sqrt{T^2 + 2T\sqrt{T\log\frac{2(p-s)}{\delta}} + 2T\log\frac{2(p-s)}{\delta}} \\ &+ n^{-1/2}\sigma\sqrt{T^2 + 2T\sqrt{T\log\frac{2s}{\delta}} + 2T\log\frac{2s}{\delta}} \\ &\le \min_{j \in S} \sum_{t=1}^T |\boldsymbol{\Sigma}_{jS_k}\boldsymbol{\beta}_{kS_k}|\end{aligned} \tag{13}$$

then $\mathbb{P}[\widehat{S}_{\ell_1}(s) = S] \ge 1 - \delta$. *iii) If*

$$\begin{aligned}&\max_{j \in S^C} \max_{1 \le t \le T} |\boldsymbol{\Sigma}_{jS_t}\boldsymbol{\beta}_{tS_t}| \\ &+ n^{-1/2}\sigma\left(\sqrt{2\log\frac{2(p-s)T}{\delta}} + \sqrt{2\log\frac{2sT}{\delta}}\right) \\ &\le \min_{j \in S} \max_{1 \le t \le T} |\boldsymbol{\Sigma}_{jS_t}\boldsymbol{\beta}_{tS_t}|\end{aligned} \tag{14}$$

then $\mathbb{P}[\widehat{S}_{\ell_\infty}(s) = S] \ge 1 - \delta$.

Theorem 3 extends Theorem 1 to the multitask setting and provides sufficient conditions for marginal regression to perform exact variable selection. We will discuss how the three different scoring procedures compare to each other in the following section.

Theorem 3 assumes that the number of relevant variables is known, as in Theorem 1. Therefore, we need to estimate the number of relevant variables in a data-dependent way. This is done using the remaining n samples, indexed by $\{n+1,\dots,2n\}$. Recall the definitions from p. 2, where j_k denotes the index of the variable for which $\widehat{r}_{\Phi,j_k}=k$, $\widehat{V}_n(k)=\text{span}\{\mathbf{X}_{j_1},\dots,\mathbf{X}_{j_k}\}$ and $\widehat{\mathbf{H}}(k)$ is the projection matrix from $\mathbb{R}^n$ to $\widehat{V}_n(k)$. Define

$$\widehat{\xi}_{\ell_2,n}(k) := \sum_{t=1}^{T} ||(\widehat{\mathbf{H}}(k+1)-\widehat{\mathbf{H}}(k))\mathbf{y}_t||_2^2, \quad 1\le k\le p-1, \tag{15}$$

which is then used to estimate the number of relevant variables as

$$\widehat{s}_{\ell_2,n} = 1+\max\{1\le k\le p-1 : \widehat{\xi}_{\ell_2,n}(k) \le (T+2\sqrt{T\log(2/\delta)}+2\log(2/\delta))\sigma^2\}. \tag{16}$$

Let $V_S=\text{span}\{\mathbf{X}_j : j\in S\}$ be the subspace spanned by columns of $\mathbf{X}$ indexed by S and similarly define $V_{S,-j}=\text{span}\{\mathbf{X}_{j'} : j'\in S\backslash\{j\}\}$. Let $\mathbf{X}_j^{(2)}$ denote the projection of $\mathbf{X}_j$ to $V_S\cap V_{S,-j}^{\perp}$. With these definitions, we have the following result.

Theorem 4. *Consider the model* (8) *with* $\boldsymbol{\epsilon}_t\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}_n)$ *and* $\sigma>0$ *known. Suppose that one of the following three claims hold: i) Eq.* (12) *holds and variables are ranked as* $\{\widehat{r}_{\ell_2,j}\}_j$*, ii) Eq.* (13) *holds and variables are ranked as* $\{\widehat{r}_{\ell_1,j}\}_j$*, or iii) Eq.* (14) *holds and variables are ranked as* $\{\widehat{r}_{\ell_1,j}\}_j$*. Furthermore assume that*

$$\min_{j\in S}\sum_{t=1}^{T}||\mathbf{X}_j^{(2)}\beta_{tj}||_2^2 > \left[2\sqrt{5}\log^{1/2}\left(\frac{4}{\delta^2}\right)\sqrt{T}+8\log\left(\frac{4}{\delta^2}\right)\right]\sigma^2. \tag{17}$$

Then $\mathbb{P}[\widehat{s}_{\ell_2,n}=s]\ge 1-2\delta$ *and* $\mathbb{P}[\widehat{S}_\phi(\widehat{s}_{\ell_2,n})=S]\ge 1-2\delta$.

Theorem 4 provides a way to select the number of relevant variables in a multitask setting. It is assumed that one of the conditions given in Theorem 3 are satisfied and that the corresponding scoring procedure is used to rank features. Condition (17) is required in order to distinguish relevant variables from noise. If the signal strength is small compared to the noise, there is no hope to select the relevant variables. Comparing to Theorem 2, we can quantify improvement over applying marginal regression to each task individually. First, the minimal signal strength for each variable, quantified as $\min_{j\in S}\sum_{t=1}^{T}||\mathbf{X}_j^{(2)}\beta_{tj}||_2^2$ needs to increase only as $\mathcal{O}(\sqrt{T})$ in multitask setting compared to $\mathcal{O}(T)$ when marginal regression is applied to each task individually.

Theorem 3 and 4 assume that the design is fixed. However, given proofs of Theorem 1 and 2, extending the proofs of the multitask marginal regression is straightforward.

2.1 Comparing Different Scoring Procedures

In this section, we compare the three scoring procedures based on $||\cdot||_1$, $||\cdot||_2$ and $||\cdot||_\infty$. Theorem 3 provides sufficient conditions under which $\widehat{S}_{\ell_1}$, $\widehat{S}_{\ell_2}$ and $\widehat{S}_{\ell_\infty}$ exactly recover the set of relevant variables S. In order to provide more intuition, we will focus on conditions (12), (13) and (14) when $\boldsymbol{\Sigma}=\mathbf{I}$. Furthermore, we assume that $s=\mathcal{O}(1)$.

From (12), we have that

$$\max_{j\in S^C} T^{-1}||\beta_{\cdot j}||_2^2+\mathcal{O}(\frac{\log p}{nT}) + \mathcal{O}(\frac{\sqrt{(T+n\max_j||\beta_{\cdot j}||_2^2)\log p}}{nT}) \le \min_{j\in S} T^{-1}||\beta_{\cdot j}||_2^2$$

is sufficient for $\widehat{S}_{\ell_2}$ to recover S. Condition (13) simplifies to

$$\max_{j\in S^C} T^{-1}||\beta_{\cdot j}||_1+\mathcal{O}(\sqrt{\frac{1+T^{-1}\log p+T^{-1/2}\sqrt{\log p}}{n}}) \le \min_{j\in S} T^{-1}||\beta_{\cdot j}||_1.$$

Finally, condition (14) simplifies to

$$\max_{j\in S^C}||\beta_{\cdot j}||_\infty+\mathcal{O}\left(\sqrt{\frac{\log pT}{n}}\right)\le\min_{j\in S}||\beta_{\cdot j}||_\infty.$$

Comparing the sufficient condition in this simplified form, we can observe that the $\widehat{S}_{\ell_2}$ requires weaker conditions for exact support recovery than $\widehat{S}_{\ell_\infty}$. Furthermore, it can be seen that the estimator $\widehat{S}_{\ell_\infty}$ is the most related to the support recovered using marginal regression on each task separately. From Theorem 1, if we stack regression coefficients for different tasks into a big vector, we have that

$$\max_{j\in S^C}\max_{1\le t\le T}|\beta_{tj}|+\mathcal{O}\left(\sqrt{\frac{\log pT}{n}}\right)\le\min_{j\in S}\min_{1\le t\le T}|\beta_{tj}|$$

is sufficient for the exact support recovery. This is a stronger requirement than the one needed for $\widehat{S}_{\ell_\infty}$.

Still, from the numerical results, we observe that $\widehat{S}_{\ell_1}$ and $\widehat{S}_{\ell_2}$ perform better than $\widehat{S}_{\ell_\infty}$.

3 Universal Lower Bound for Hamming distance

So far, we have focused on the exact variable selection. Although the exact variable selection has been the focus of many studies, the exact recovery of variables is not possible in many practical applications with low signal to noise ratio. Therefore, it is more natural to measure performance using a distance between the sets of selected variables and the true set S.

In this section, let $\mathbf{X}$, $\mathbf{y}_1, \ldots, \mathbf{y}_T$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T$, $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T$ be the same as before. Here $\mathbf{X}$ could be either deterministic or random satisfying $\mathbf{X}_j'\mathbf{X}_j = 1$ for $j = 1, \ldots, p$. We are interested in studying the lower bound for variable selection problem measured by Hamming distance. To construct lower bound, we need to clearly define the model family we are studying. We use the following random coefficient model which is adapted from Genovese et al. (2009):

$$\beta_{tj} \overset{\text{i.i.d.}}{\sim} (1-\eta_p)\nu_0 + \eta_p \nu_{\tau_p}, \tag{18}$$

for all $t = 1, \ldots, T$, $j = 1, \ldots, p$, where ν_0 is the point mass at 0 and ν_{τ_p} is the point mass at τ_p. Both η_p and τ_p vary with p. We set

$$\eta_p = p^{-v}, \quad 0 < v < 1, \tag{19}$$

so that the expected number of signals is $s_p = p\eta_p = p^{1-v}$. Let $r > 0$ be some fixed constant and set $\tau_p = \sqrt{2r \log p}$ the signal strength. Such a setting has been extensively explored in the community of modern statistics to explore the theoretical limit of many problems including classification, density estimation, and multiple hypothesis testing (Donoho and Jin, 2004; Cai et al., 2007; Ji and Jin, 2010).

Let $\widehat{S}$ be the index set of selected variables for any variable selection procedure and S be the index set of true relevant variables. We define the Hamming distance

$$H_p(\widehat{S}, S \,|\, \mathbf{X}) = \mathbb{E}_{\eta_p, \pi_p}\left[\left|(\widehat{S} \setminus S) \cup (S \setminus \widehat{S})\right|\right]. \tag{20}$$

Let

$$\begin{aligned}\lambda_p &:= \frac{1}{\tau_p}\left[\log\left(\frac{1-\eta_p}{\eta_p}\right) + \frac{T\tau_p^2}{2}\right] \\ &= \frac{1}{\sqrt{2r\log p}}\log(p^v - 1) + T\sqrt{\frac{r\log p}{2}} \\ &\le \frac{(v+Tr)\sqrt{\log p}}{\sqrt{2r}}.\end{aligned}$$

Our main result in this section provides a universal lower bound of $H_p(\widehat{S}, S \,|\, \mathbf{X})$ for all sample size n and design matrix $\mathbf{X}$. Let $F(\cdot)$ and $\bar{F}(\cdot)$ be the distribution function and survival function of the standard Gaussian distribution and let $\phi(\cdot)$ denote the density function of the standard Gaussian distribution. We have the following lower bound results.

Theorem 5. (Universal lower bound) *Fix $v \in (0,1), r > 0$ and a sufficiently large p. For any n and design matrix $\mathbf{X}$ such that $\mathbf{X}'\mathbf{X}$ has unit diagonals, we have the following lower bound:*

$$\frac{H_p(\widehat{S}, S \,|\, \mathbf{X})}{s_p} \ge \left[\frac{1-\eta_p}{\eta_p}\bar{F}\left(\frac{\lambda_p}{\sqrt{T}}\right) + F\left(\frac{\lambda_p}{\sqrt{T}} - \sqrt{T}\tau_p\right)\right]. \tag{21}$$

This can be further written as

$$\frac{H_p(\widehat{S}, S \,|\, \mathbf{X})}{s_p} \ge \begin{cases} \dfrac{\sqrt{rT}}{2(v+Tr)\sqrt{\pi\log p}} \cdot p^{-(v-Tr)^2/(4rT)}, & v < rT \\ 1 + o(1), & v > rT. \end{cases} \tag{22}$$

One thing to note in the above theorem is that such a lower bound simultaneously holds for any sample size n. The main reason for this is that we constraint $\mathbf{X}_j'\mathbf{X}_j = 1$ for all $j = 1, \ldots, p$. Such a standardization essentially fixes the signal-to-noise ratio under asymptotic framework where p increases. Therefore, the lower bound does not depend on sample size n.

3.1 Comparing with Single Task Screening

It would be instructive to compare the lower bounds for multitask screening with that for single task screening. By setting $T = 1$, we can obtain from Theorem 5 that the Hamming distance lower bound for single task screening takes the form:

$$\frac{H_p^{\text{single}}(\widehat{S}, S \,|\, \mathbf{X})}{s_p} \ge \begin{cases} \dfrac{\sqrt{r}}{2(v+r)\sqrt{\pi\log p}} \cdot p^{-(v-r)^2/(4r)}, & v < r \\ 1 + o(1), & v > r. \end{cases} \tag{23}$$

If $v > r$, $H_p^{\text{single}}(\widehat{S}, S \,|\, \mathbf{X}) \ge s_p + o(1)$, which means that no procedure can recover any information of the true signal at all. On the other hand, the corresponding no recovery condition for multitask screening is strengthened to be $r > Tr$ and such a condition rarely holds when T is larger. Therefore, one effect of the

multitask setting is that the signal-to-noise ratio is improved by jointly considering multiple tasks. For the case that $r < vT$ and $r < T$ in both settings, it can be seen that the rate for multitask screening is much faster than that for single-task screening.

4 Empirical Results

We conduct an extensive number of numerical studies to evaluate the finite sample performance of marginal regression on the multitask model given in (8). We consider marginal regression using the three scoring procedures outlined in Section 2. The variables are ranked using $||\cdot||_1$, $||\cdot||_2$ and $||\cdot||_\infty$ norms and the resulting sets of variables are denoted $\widehat{S}_{\ell_1}$, $\widehat{S}_{\ell_2}$ and $\widehat{S}_{\ell_\infty}$. The number of active variables is set using the result of Theorem 4.

Let $\widehat{S}$ be an estimate obtained by one of the scoring methods. We evaluate the performance averaged over 200 simulation runs. Let $\widehat{\mathbb{E}}_n$ denote the empirical average over the simulation runs. We measure the size of the support $\widehat{S}$. Next, we estimate the probability that the estimated set contains the true set S, that is, $\widehat{\mathbb{E}}_n[\mathbb{I}\{S \subseteq \widehat{S}\}]$, which we call coverage probability. We define fraction of correct zeros $(p-s)^{-1}\widehat{\mathbb{E}}_n[|\widehat{S}^C \cap S^C|]$, fraction of incorrect zeros $s^{-1}\widehat{\mathbb{E}}_n[|\widehat{S}^C \cap S|]$ and fraction of correctly fitted $\widehat{\mathbb{E}}_n[\mathbb{I}\{S = \widehat{S}\}]$ to measure the performance of different scoring procedures.

We outline main findings using the following simulation studies. Due to space constraints, tables with detailed numerical results are given in the Appendix.

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 $t \in 1, \ldots, T$, the non-zero coefficients are given as $\beta_{tj} = (-1)^u(4n^{-1/2}\log n + |z|)$, where $u \sim$ Bernoulli(0.4) and $z \sim \mathcal{N}(0, 1)$. The number of non-zero elements in $\{\beta_{tj}\}_t$ is given as a parameter $T_{\text{non-zero}} \in \{500, 300, 100\}$. The positions of non-zero elements are chosen uniformly at random from $\{1, \ldots, T\}$. The noise is Gaussian with the standard deviation σ set to control the signal-to-noise ratio (SNR). SNR is defined as $\text{Var}(\mathbf{x}\boldsymbol{\beta})/\text{Var}(\boldsymbol{\epsilon})$ and we vary SNR $\in \{15, 10, 5, 1\}$.

Simulation 2: The following model is used to evaluate the performance of the methods as the number of non-zero elements in $\{\beta_{tj}\}_t$ varies. We set $(n, p, s) = (100, 500, 10)$ and vary the number of outputs $T \in \{500, 750, 1000\}$. For each number of outputs T, we vary $T_{\text{non-zero}} \in \{0.8T, 0.5T, 0.2T\}$. The samples $\mathbf{x}_i$ and regression coefficients are given as in Simulation 1, that is, $\mathbf{x}_i$ is drawn from a multivariate standard normal distribution and the non-zero coefficients are given as $\beta_{tj} = (-1)^u(4n^{-1/2}\log n + |z|)$, where $u \sim$ Bernoulli(0.4) and $z \sim \mathcal{N}(0, 1)$. The noise is Gaussian, with the standard deviation defined through the SNR, which varies in $\{10, 5, 1\}$.

Simulation 3: The following model is borrowed from Wang (2009). We assume a correlation structure between variables given as $\text{Var}(\mathbf{X}_{j_1}, \mathbf{X}_{j_2}) = \rho^{|j_1 - j_2|}$, where $\rho \in \{0.2, 0.5, 0.7\}$. This correlation structure appears naturally among ordered variables. We set $(n, p, s, T) = (100, 5000, 3, 150)$ and $T_{\text{non-zero}} = 80$. The relevant variables are at positions $(1, 4, 7)$ and non-zero coefficients are given as $3, 1.5$ and 2 respectively. The SNR varies in $\{10, 5, 1\}$.

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 is set to 0 otherwise. We set $(n, p, s, T) = (150, 4000, 8, 150)$, $T_{\text{non-zero}} = 80$ and the parameter $\rho \in \{0.2, 0.5\}$. 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 are given in Simulation 1. The noise is Gaussian and the SNR varies in $\{10, 5, 1\}$.

Simulation 5: This model represents a difficult setting. It is modified from Wang (2009). We set $(n, p, s, T) = (200, 10000, 5, 500)$. The number of non-zero elements in each row varies is $T_{\text{non-zero}} \in \{400, 250, 100\}$. For $j \in [s]$ and $t \in [T]$, the non-zero elements equal $\beta_{tj} = 2j$. 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 than $\text{Corr}(x_{i,j}, y_{t,i})$ for $j \in [p]\backslash[s]$, so that it becomes difficult to select variable 1. The variable 1 is masked with the noisy variables. This setting is difficult for screening procedures as they take into consideration only marginal information. The noise is Gaussian with standard deviation $\sigma \in \{1.5, 2.5, 4.5\}$.

Our simulation setting transitions from a simple scenario considered in Simulation 1 towards a challenging one in Simulation 5. Simulation 1 represents a toy model, where variables are independent. Simulation 2 examines the influence of the number of non-zero elements in the set $\{\beta_{tj}\}_t$. Simulations 3 and 4 represent more challenging situations with structured correlation that naturally appears in many data sets, for example, a correlation between gene measurements that

	$\widehat{S}$	Prob. (%) of $S \subseteq \widehat{S}$	Fraction (%) of Correct zeros	Fraction (%) of Incorrect zeros	Fraction (%) of $S = \widehat{S}$	$\lvert\widehat{S}\rvert$
Simulation 1: $(n, p, s, T) = (500, 20000, 18, 500)$, $T_{\text{non-zero}} = 300$						
SNR $= 5$	$\widehat{S}_{\ell_\infty}$	100.0	100.0	0.0	76.0	18.3
	$\widehat{S}_{\ell_1}$	100.0	100.0	0.0	91.0	18.1
	$\widehat{S}_{\ell_2}$	100.0	100.0	0.0	92.0	18.1
Simulation 2.a: $(n, p, s, T) = (200, 5000, 10, 500)$, $T_{\text{non-zero}} = 400$						
SNR $= 5$	$\widehat{S}_{\ell_\infty}$	100.0	100.0	0.0	82.0	10.2
	$\widehat{S}_{\ell_1}$	100.0	100.0	0.0	91.0	10.1
	$\widehat{S}_{\ell_2}$	100.0	100.0	0.0	91.0	10.1
Simulation 3: $(n, p, s, T) = (100, 5000, 3, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.7$						
SNR $= 5$	$\widehat{S}_{\ell_\infty}$	96.0	100.0	1.3	95.0	3.0
	$\widehat{S}_{\ell_1}$	99.0	100.0	0.3	97.0	3.0
	$\widehat{S}_{\ell_2}$	97.0	100.0	1.0	95.0	3.0
Simulation 4: $(n, p, s, T) = (150, 4000, 8, 150)$, $T_{\text{non-zero}} = 80$, $\rho = 0.5$						
SNR $= 5$	$\widehat{S}_{\ell_\infty}$	100.0	100.0	0.0	84.0	8.2
	$\widehat{S}_{\ell_1}$	100.0	100.0	0.0	87.0	8.1
	$\widehat{S}_{\ell_2}$	100.0	100.0	0.0	87.0	8.1
Simulation 5: $(n, p, s, T) = (200, 10000, 5, 500)$, $T_{\text{non-zero}} = 250$						
$\sigma = 2.5$	$\widehat{S}_{\ell_\infty}$	87.0	100.0	2.6	39.0	5.9
	$\widehat{S}_{\ell_1}$	0.0	99.9	90.6	0.0	14.8
	$\widehat{S}_{\ell_2}$	0.0	99.9	55.0	0.0	12.5

Table 1: Results of simulations. Tables with all results are given in the Appendix.

are closely located on a chromosome. Finally, Simulation 5 is constructed in such a way such that an irrelevant variable is more correlated with the output than a relevant variable. Tables giving detailed results of the above described simulations are given in the Appendix. Table 1 reproduces some of the results. We observe that the sets $\widehat{S}_{\ell_1}$ and $\widehat{S}_{\ell_2}$ perform similarly across different simulation settings. Except for the simulation 5, $\widehat{S}_{\ell_\infty}$ has worse performance than the other two estimators. The performance difference is increased as the signal to noise ratio decreases. However, when the signal to noise ratio is large, there is little difference between the procedures.

5 Discussion

This paper has focused on the analysis of marginal regression in the multitask setting. Due to its simplicity and computational efficiency, marginal regression is often applied in practice. Therefore, it is important to understand under what assumptions it can be expected to work well. Using multiple related tasks, the signal in data can be more easily detected and the estimation procedure is more efficient. Our theoretical results support this intuition. One open question still remains. It is still not clear how to match the lower bound on the Hamming distance given in Section 3, but we suspect that recent developments in Ji and Jin (2010) could provide tools to match the lower bound.

Acknowledgements

We would like to thank anonymous reviewers whose comments helped improve the manuscript. Han Liu is a supported by NSFgrant IIS-1116730.

References

P. Alquier. Lasso, iterative feature selection and the correlation selector: Oracle inequalities and numerical performances. *Electronic Journal of Statistics*, 2:1129–1152, 2008.

T. Cai, L. Wang, and G. Xu. Shifting inequality and recovery of sparse signals. *Signal Processing*, 58(3): 1300–1308, 2010.

T.T. Cai, J. Jin, and M.G. Low. Estimation and confidence sets for sparse normal mixtures. *The Annals of Statistics*, 35(6):2421–2449, 2007.

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.

S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

D.L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. Technical Report UC/EECS-2010-126, EECS Dept., University of California at Berkeley, September 2010.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA*, 96:1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *JRSS: B*, 70 (5):849–911, 2008.

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: beyond the linear model. *JMLR*, 10:2013–2038, 2009.

J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *JASA*, 106(495):544–557, 2011.

C. Genovese, J. Jin, and L. Wasserman. Revisiting marginal regression. *arXiv:0911.4080*, 2009.

P. Ji and J. Jin. UPS Delivers Optimal Phase Diagram in High Dimensional Variable Selection. *ArXiv e-prints*, October 2010.

G. Kerkyacharian, M. Mougeot, D. Picard, and K. Tribouley. Learning out of leaders. *Multiscale, Nonlinear and Adaptive Approximation*, pages 295–324, 2009.

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

M. Kolar and E. P. Xing. Ultra-high dimensional multiple output learning with simultaneous orthogonal matching pursuit: Screening approach. In *AISTATS*, pages 413–420, 2010.

M. Kolar, J. Lafferty, and L. Wasserman. Union support recovery in multi-task learning. *J. Mach. Learn. Res.*, 12:2415–2435, July 2011. ISSN 1532-4435.

H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In *ICML*, pages 649–656, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in Multi-Task learning. In *COLT*, 2009.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *arXiv 1007.1771*, 2010.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

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.

J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000. ISBN 0-262-19440-6.

R. Tibshirani. Regression shrinkage and selection via the lasso. *JRSS: B*, 58:267–288, 1996.

R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R.J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Arxiv preprint arXiv:1011.2234*, 2010.

B.A. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Technometrics*, 47(3): 349–363, 2005. ISSN 0040-1706.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

H. Wang. Forward regression for ultra-high dimensional variable screening. *JASA*, 104(488):1512–1524, 2009.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38 (2):894–942, 2010.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *JASA*, 101:1418–1429, 2006.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.

H. Zou and M. Yuan. The $\mathtt{F}_\infty$-norm support vector machine. *Stat. Sin*, 18:379–398, 2008.