Journal of Machine Learning Research 1 (2011) xx-xx Submitted xx/xx; Published xx/xx

A Comparison of the Lasso and Marginal Regression

Christopher Genovese GENOVESE@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Jiashun Jin JIASHUN@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Larry Wasserman LARRY@STAT.CMU.EDU
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Zhigang Yao ZHY16@PITT.EDU
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260, USA

Editor: xxxx

Abstract

The lasso is an important method for sparse, high-dimensional regression problems, with efficient algorithms available, a long history of practical success, and a large body of theoretical results supporting and explaining its performance. But even with the best available algorithms, finding the lasso solutions remains a computationally challenging task in cases where the number of covariates vastly exceeds the number of data points.

Marginal regression, where each dependent variable is regressed separately on each covariate, offers a promising alternative in this case because the estimates can be computed roughly two orders faster than the lasso solutions. The question that remains is how the statistical performance of the method compares to that of the lasso in these cases.

In this paper, we study the relative statistical performance of the lasso and marginal regression for sparse, high-dimensional regression problems. We consider the problem of learning which coefficients are non-zero. Our main results are as follows: (i) we compare the conditions under which the lasso and marginal regression guarantee exact recovery in the fixed design, noise free case; (ii) we establish conditions under which marginal regression provides exact recovery with high probability in the fixed design, noise free, random coefficients case; and (iii) we derive rates of convergence for both procedures, where performance is measured by the number of coefficients with incorrect sign, and characterize the regions in the parameter space recovery is and is not possible under this metric.

In light of the computational advantages of marginal regression in very high dimensional problems, our theoretical and simulations results suggest that the procedure merits further study.

Keywords: high-dimensional regression, lasso, phase diagram, regularization

1. Introduction

Consider a regression model,

$$Y = X\beta + z, \tag{1}$$

with response $Y = (Y_1, \ldots, Y_n)^T$, $n \times p$ design matrix X, coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$, and noise variables $z = (z_1, \ldots, z_n)^T$. A central theme in recent work on regression is that sparsity plays a critical role in effective high-dimensional inference. Loosely speaking, we call the model *sparse* when most of β's components equal 0, and we call it *high dimensional* when $p \gg n$.

An important problem in this context is variable selection: determining which components of β are non-zero. For general β, the problem is underdetermined, but recent results have demonstrated that under particular conditions on X, to be discussed below, sufficient sparsity of β allows (i) exact recovery of β in the noise-free case [26] and (ii) consistent selection of the non-zero coefficients in the noisy-case [5, 2, 4, 6, 8, 15, 16, 18, 20, 26, 28, 33, 34]. Many of these results are based on showing that under sparsity constraints, the lasso—a convex optimization procedure that controls the ℓ^1 norm of the coefficients—has the same solution as an (intractable) combinatorial optimization problem that controls the number of non-zero coefficients.

Recent years, the lasso [25, 5] has become one of the main practical and theoretical tools for sparse high-dimensional variable selection problems. In the regression problem, the lasso estimator is defined by

$$\widehat{\beta}_{\text{lasso}} = \operatorname*{argmin}_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{2}$$

where $\|\beta\|_1 = \sum_j |\beta_j|$ and $\lambda \geq 0$ is a tuning parameter that must be specified. The lasso gives rise to a convex optimization problem and thus is computationally tractable even for moderately large problems. Indeed, the LARS algorithm [12] can compute the entire solution path as a function of λ in $O(p^3 + np^2)$ operations. Gradient descent algorithms for the lasso are faster in practice, but have the same computational complexity. The motivation for our study is that when p is very large, finding the lasso solutions is computationally demanding.

Marginal regression (also called correlation learning, simple thresholding [6], and sure screening [15]) is an older and computationally simpler method for variable selection in which the outcome variable is regressed on each covariate separately and the resulting coefficient estimates are screened. To compute the marginal regression estimates for variable selection, we begin by computing the marginal regression coefficients which, assuming X has been standardized, are

$$\widehat{\alpha} \equiv X^T Y. \tag{3}$$

Then, we threshold $\widehat{\alpha}$ using the tuning parameter $t > 0$:

$$\widehat{\beta}_j = \widehat{\alpha}_j 1\{|\widehat{\alpha}_j| \geq t\}. \tag{4}$$

The procedure requires $O(np)$ operations, two orders faster than the lasso for $p \gg n$. This is a decided advantage for marginal regression because the procedure is tractable for much larger problems than is the lasso. The question that remains is how the *statistical*

performance of marginal regression compares to that of the lasso. In this paper, we begin to address this question.

We study the relative performance of the lasso and marginal regression for variable selection in three regimes: (a) exact variable selection in the noise-free and noisy cases with fixed design and coefficients, (b) exact variable selection in the noise-free case with fixed design and random coefficients, and (c) statistical variable selection in the noisy case where performance is measured by the number of coefficients with incorrect sign. Our goal is to reopen the case for marginal regression as a plausible alternative to the lasso for large problems. If marginal regression exhibits comparable statistical performance, theoretically and empirically, then its computational advantages make it a good choice in practice. Put another way: for very high dimensional problems, marginal regression only needs to tie to win.

Our main results are as follows:

- In the fixed design (X fixed), noise free ($z = 0$), and fixed effects (β fixed) case, both procedures guarantee exact reconstruction of $|\operatorname{sgn}\beta|$ under distinct but generally overlapping conditions.

  We analyze these conditions and give examples where each procedure fails while the other succeeds. The lasso has the advantage of providing exact reconstruction for a somewhat larger class of coefficients, but marginal regression has a better tolerance for collinearity and is easier to tune. These results are discussed in Sections 2 and 2.4.

- In the fixed design, noise free, and random effects (β random) case, we give conditions under which marginal regression gives exact reconstruction of $|\operatorname{sgn}\beta|$ with overwhelming probability.

  Our condition is closely related to both the Faithfulness condition [24, 20] and the Incoherence condition [8]. The latter depends only on X, making it easy to check in practice, but in controlling the worst case it is quite conservative. The former depends on the unknown β but is less stringent. Our condition strikes a compromise between the two. These results are discussed in Section 3.

- In the fixed design, noisy, fixed effects case, we obtain convergence rates of the two procedures in Hamming distance between the sign vectors $\operatorname{sgn}\beta$ and $\operatorname{sgn}\widehat{\beta}$.

  Under a stylized family of signals, we derive a new "phase diagram" that partitions the parameter space into regions in which (i) exact variable selection is possible (asymptotically); (ii) reconstruction of most relevant variables, but not all, is possible; and (iii) successful variable selection is impossible. We show that both the lasso and marginal regression, properly calibrated, perform similarly in each region. These results are described in Section 4.

To support these theoretical results, we also present simulation studies in Section 5. Our simulations show that marginal regression and the lasso perform similarly over a range of parameters in realistic models. Section 6 gives the proofs of all theorems and lemmas in the order they appear.

Notation. For a real number x, let $\operatorname{sgn}(x)$ be -1, 0, or 1 when $x < 0$, $x = 0$, and $x > 0$; and for vector $u, v \in \mathrm{R}^k$, define $\operatorname{sgn}(u) = (\operatorname{sgn}(u_1), \ldots, \operatorname{sgn}(u_k))^T$, and let (u, v) be the inner

product. We will use $\|\cdot\|$, with various subscripts, to denote vector and matrix norms, and $|\cdot|$ to represent absolute value, applied component-wise when applied to vectors. With some abuse of notation, we will write $\min u$ ($\min |u|$) to denote the minimum (absolute) component of a vector u. Inequalities between vectors are to be understood component-wise as well.

Consider a sequence of noiseless regression problems with deterministic design matrices, indexed by sample size n,

$$Y^{(n)} = X^{(n)}\beta^{(n)}. \tag{5}$$

Here, $Y^{(n)}$ is an $n \times 1$ response vector, $X^{(n)}$ is an $n \times p^{(n)}$ matrix and $\beta^{(n)}$ is a $p^{(n)} \times 1$ vector, where we typically assume $p^{(n)} \gg n$. We assume that $\beta^{(n)}$ is sparse in the sense that it has $s^{(n)}$ nonzero components where $s^{(n)} \ll p^{(n)}$. By rearranging $\beta^{(n)}$ without loss of generality, we can partition each $X^{(n)}$ and $\beta^{(n)}$ into "signal" and "noise" pieces, corresponding to the non-zero or zero coefficients, as follows:

$$X^{(n)} = \left(X_S^{(n)}, X_N^{(n)}\right) \qquad \beta^{(n)} = \begin{pmatrix} \beta_S \\ \beta_N \end{pmatrix}. \tag{6}$$

Finally, define the Gram matrix $C^{(n)} = (X^{(n)})^T X^{(n)}$ and partition this as

$$C^{(n)} = \begin{pmatrix} C_{SS}^{(n)} & C_{SN}^{(n)} \\ C_{NS}^{(n)} & C_{NN}^{(n)} \end{pmatrix}, \tag{7}$$

where of course $C_{NS}^{(n)} = (C_{SN}^{(n)})^T$. Except in Sections 4–5, we suppose $X^{(n)}$ is normalized so that all diagonal coordinates of $C^{(n)}$ are 1.

These $^{(n)}$ superscripts become tedious, so for the remainder of the paper, we suppress them unless necessary to show variation in n. The quantities X, C, p, s, ρ, as well as the tuning parameters λ (for the lasso; see (2)) and t (for marginal regression; see (4)) are all thus implicitly dependent on n.

2. Noise-Free Conditions for Exact Variable Selection

We restrict our attention to a sequence of regression problems in which the signal (non-zero) components of the coefficient vector have large enough magnitude to be distinguishable from zero. Specifically, assume that $\beta_S \in \mathcal{M}_\rho^s$ for a sequence $\rho(\equiv \rho^{(n)}) > 0$ (and not converging to zero too quickly) with

$$\mathcal{M}_a^k = \left\{ x = (x_1, \ldots, x_k)^T \in \mathrm{R}^k : \ |x_j| \geq a \text{ for all } 1 \leq j \leq k \right\}, \tag{8}$$

for positive integer k and $a > 0$. (We use $\mathcal{M}_\rho$ to denote the space $\mathcal{M}_\rho^s \equiv \mathcal{M}_{\rho^{(n)}}^{s^{(n)}}$.)

We will begin by specifying conditions on C, ρ, λ, and t such that in the noise-free case, exact reconstruction of β is possible for the lasso or marginal regression, for all coefficient vectors for which the (non-zero) signal coefficients $\beta_S \in \mathcal{M}_\rho$. These in turn lead to conditions on C, p, s, ρ, λ, and t such that in the case of homoscedastic Gaussian

noise, the non-zero coefficients can be selected consistently, meaning that for all sequences $\beta_S^{(n)} \in \mathcal{M}_{\rho^{(n)}}^{s^{(n)}} \equiv \mathcal{M}_\rho$,

$$P\left(\left|\operatorname{sgn}(\widehat{\beta}^{(n)})\right| = \left|\operatorname{sgn}(\beta^{(n)})\right|\right) \to 1, \tag{9}$$

as $n \to \infty$. (This property was dubbed *sparsistency* by Pradeep Ravikumar [22].) Our goal in this section is to compare the conditions for the two procedures. We focus on the noise-free case, although we comment on the noisy case briefly.

2.1 Exact reconstruction conditions for the lasso in the noise-free case

We begin by reviewing three conditions in the noise-free case that are now standard in the literature on the lasso:

Condition E. *The minimum eigenvalue of C_{SS} is positive.*

Condition I. $\max \left|C_{NS}C_{SS}^{-1} \operatorname{sgn}(\beta_S)\right| \le 1$.

Condition J. $\min \left|\beta_S - \lambda C_{SS}^{-1} \operatorname{sgn}(\beta_S)\right| > 0$.

Because C_{SS} is symmetric and non-negative definite, Condition E is equivalent to C_{SS} being invertible. Later we will strengthen this condition. Condition I is sometimes called the *irrepresentability* condition; note that it depends only on $\operatorname{sgn}\beta$, a fact that will be important later.

For the noise-free case, Wainwright [28, Lemma 1] proves that Conditions E, I, and J are necessary and sufficient for exact reconstruction of the sign vector, i.e., for the existence of a lasso solution $\widehat{\beta}$ such that $\operatorname{sgn}\widehat{\beta} = \operatorname{sgn}\beta$. (See also [33]). Note that this result is stronger than correctly selecting the non-zero coefficients, as it gets the signs correct as well.

It will be useful in what follows to give strong forms of these conditions. Maximizing the left side of Condition I over all 2^s sign patterns gives $\|C_{NS}C_{SS}^{-1}\|_\infty$, the maximum-absolute-row-sum matrix norm. It follows that Condition I holds for all $\beta_S \in \mathcal{M}_\rho$ if and only if $\|C_{NS}C_{SS}^{-1}\|_\infty \le 1$. Similarly, one way to ensure that Condition J holds over $\mathcal{M}_\rho$ is to require that every component of $\lambda C_{SS}^{-1} \operatorname{sgn}(\beta_S)$ be less than ρ. The maximum component of this vector over $\mathcal{M}_\rho$ equals $\lambda\|C_{SS}^{-1}\|_\infty$, which must be less than ρ. A simpler relation, in terms of the smallest eigenvalue of C_{SS} is

$$\frac{\sqrt{s}}{\operatorname{eigen}_{\min}(C_{SS})} = \sqrt{s}\|C_{SS}^{-1}\|_2 \ge \|C_{SS}^{-1}\|_\infty \ge \|C_{SS}^{-1}\|_2 = \frac{1}{\operatorname{eigen}_{\min}(C_{SS})}, \tag{10}$$

where the inequality follows from the symmetry of C_{SS} and standard norm inequalities. This yields the following:

Condition E'. *The minimum eigenvalue of C_{SS} is no less than $\lambda_0 > 0$, where λ_0 does not depend on n.*

Condition I'. $\|C_{NS}C_{SS}^{-1}\|_\infty \le 1 - \eta$, *for $0 < \eta < 1$ small and independent of n.*

Condition J'. $\lambda < \dfrac{\rho}{\|C_{SS}^{-1}\|_\infty}$. *(Under Condition E', we can instead use $\lambda < \rho\frac{\lambda_0}{\sqrt{s}}$.)*

Theorem 1 *In the noise-free case, Conditions E' (or E), I' (or I), and J' imply that for all $\beta_S \in \mathcal{M}_\rho$, there exists a lasso solution $\widehat{\beta}$ with* $\mathrm{sgn}(\widehat{\beta}) = \mathrm{sgn}(\beta)$.

These conditions can be weakened in various ways, but we chose these because they transition nicely to the noisy case. For instance, [28] shows that a slight extension of Conditions E', I', and J' gives sparsistency in the case of homoscedastic Gaussian noise.

2.2 Exact reconstruction conditions for marginal regression in the noise-free case

As above, define $\widehat{\alpha} = X^T Y$ and $\widehat{\beta}_j = \widehat{\alpha}_j 1\{|\widehat{\alpha}_j| \geq t\}$, $1 \leq j \leq p$. Exact reconstruction for variable selection requires that $\widehat{\beta}_j \neq 0$ whenever $\beta_j \neq 0$, or equivalently $|\widehat{\alpha}_j| \geq t$ whenever $\beta_j \neq 0$. In the literature on causal inference [24], this assumption has been called *faithfulness* and is also used in [1, 15]. The usual justification for this assumption is that if β is selected at random from some distribution, then faithfulness holds with high probability. [23] has criticized this assumption because results which hold under faithfulness cannot hold in any uniform sense. We feel that despite the lack of uniformity, it is still useful to investigate results that hold under faithfulness, since as we will show, it holds with high probabilty under weak conditions.

By simple algebra, we have that

$$\widehat{\alpha} = \begin{pmatrix} \widehat{\alpha}_S \\ \widehat{\alpha}_N \end{pmatrix} = \begin{pmatrix} X_S^T X_S \beta_S \\ X_N^T X_S \beta_S \end{pmatrix}.$$

The following condition is thus required to correctly identify the non-zero coefficients:

Condition F. $$\max |C_{NS}\beta_S| < \min |C_{SS}\beta_S|. \tag{11}$$

Because this is reminiscent of (although distinct from) the faithfulness condition mentioned above, we will refer to Condition F as the *Faithfulness Condition*.

Theorem 2 *Condition F is necessary and sufficient for exact reconstruction to be possible for some $t > 0$ with marginal regression.*

Unfortunately, as the next theorem shows, Condition F cannot hold for all $\beta_S \in \mathcal{M}_\rho$. Applying the theorem to C_{SS} shows that for any $\rho > 0$, there exists a $\beta_S \in \mathcal{M}_\rho$ that violates equation (11).

Proposition 3 *Let D be an $s \times s$ positive definite, symmetric matrix that is not diagonal. Then for any $\rho > 0$, there exists a $\beta \in \mathcal{M}_\rho^s$ such that* $\min |D\beta| = 0$.

Despite the seeming pessimism of Theorem 3, the result is not as grim as it seems. Since $C\beta \equiv X^T Y$, the theorem says that if we fix X and let $Y = X\beta$ range through all possible $\beta \in \mathcal{M}_\rho^s$, there exists a Y such that $\min |X^T Y| = 0$. However, to mitigate the pessimism, note that once X and Y are are observed, if we see that $\min |X^T Y| > 0$, we can rule out the result of Theorem 3.

2.3 Comparison of the conditions for exact recovery of sign vector in the noise-free case

In this subsection, we use several simple examples to get insight into how the exact-recovery conditions for the lasso and marginal regression compare. The examples illustrate the following points:

- (Examples 1 and 2) The conditions for the two procedures are generally overlapping.
- (Example 3) When $C_{SS} = I$, the lasso conditions are relatively weaker.
- (Example 4) Although the conditions for marginal regression do not hold uniformly over any $\mathcal{M}_\rho$, they have the advantage that they do not require invertibility of C_{SS} and hence are less sensitive to small eigenvalues.

The bottom line is that the two conditions appear to be closely related, and that there are cases where each succeeds while the other fails.

Example 1. For $s = 2$, assume that

$$C_{SS} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \beta_S = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

For $a = (a_1, a_2)$ a row of C_{NS}, Conditions I and J require that we choose $\lambda > 0$ small enough so that $|a_1 + a_2| \leq 1 + \rho$, while Condition F requires $|2a_1 + a_2| \leq \min\{(2+\rho), |1+2\rho|\}$. For many choices of ρ, both of these inequalities are satisfied (e.g., $\rho = -0.75$). Figure 1 shows the sets of (a_1, a_2) for which the respective conditions are satisfied. The two regions show significant overlap, and to a large extent, the conditions continue to overlap as ρ and β_S vary. Examples for larger s can be constructed by letting C_{SS} be a block diagonal matrix, where the size of each main diagonal block is small. For each row of C_{NS}, the conditions for the lasso and marginal regression are similar to those above, though more complicated.

Example 2. In the special case where $\beta_S \propto 1_S$, Condition I for the lasso becomes $|C_{NS}C_{SS}^{-1}1_N| \leq 1_N$, where the inequality should be interpreted as holding component-wise, and the condition for marginal regression (Condition F) is $\max\{|C_{NS}1_S|\} \leq \min\{|C_{SS}1_S|\}$. Note that if in addition 1_S is an eigen-vector of C_{SS}, then the two conditions are equivalent to each other. This includes but is not limited to the case of $s = 2$.

Example 3. Fix n and consider the special case in which $C_{SS} = I$. For the lasso, Condition E' (and thus E) is satisfied, Condition J' reduces to $\lambda < \rho$, and Condition I becomes $\|C_{NS}\| \leq 1$. Under these conditions, the lasso gives exact reconstruction, but Condition F can fail. To see how, let $\widetilde{\beta} \in \{-1, 1\}^s$ be the vector such that $\max|C_{NS}\widetilde{\beta}| = \|C_{NS}\|_\infty$ and let ℓ be the index of the row at which the maximum is attained, choosing the row with the biggest absolute element if the maximum is not unique. Let u be the maximum absolute element of row ℓ of C_{NS} with index j. Define a vector δ to be zero except in component j, which has the value $\rho\widetilde{\beta}_j/(u\|C_{NS}\|_\infty)$. Let $\beta = \rho\widetilde{\beta} + \rho\delta$. Then,

$$|(C_{NS}\beta)_\ell| = \rho\left(\|C_{NS}\|_\infty + \frac{1}{\|C_{NS}\|_\infty}\right) > \rho = \min|\beta_S|, \tag{12}$$

so Condition F fails.

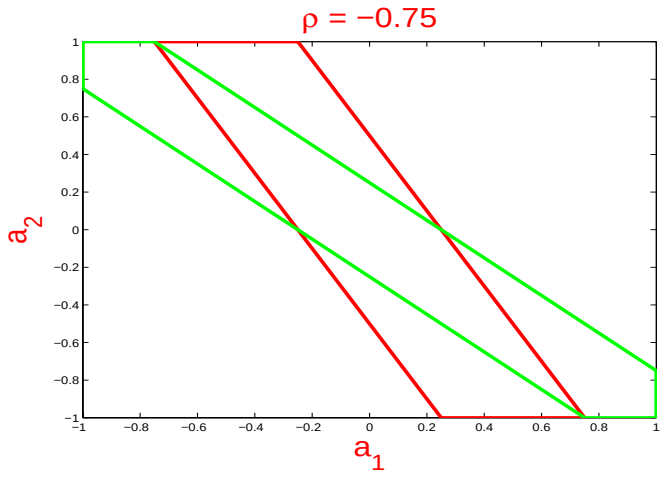

Figure 1: Let C_{SS} and β_S be as in Section 2.3 Example 2, where $\rho = -0.75$. For a row of C_{NS}, say, (a_1, a_2). The interior of the red box and that of the green box are the regions of (a_1, a_2) satisfying the respective conditions in Example 1.

On the other hand, suppose Condition F holds for all $\beta_S \in \{-1, 1\}^s$. (It cannot hold for all $\mathcal{M}_\rho$ by Theorem 3). Then, for all $\beta_S \in \{-1, 1\}^s$, $\max |C_{NS}\beta_S| \leq 1$, which implies that $\|C_{NS}\|_\infty \leq 1$. Choosing $\lambda < \rho$, we have Conditions E', I, and J' satisfied, showing by Theorem 1 that the lasso gives exact reconstruction. It follows that the conditions for the lasso are weaker in this case.

Example 4. For simplicity, assume that $\beta_S \propto 1_S$, although the phenomenon to be described below is not limited to this case. For $1 \leq i \leq s$, let λ_i and ξ_i be the i-th eigenvalue and eigenvector of C_{SS}. Without loss of generality, we can take ξ_i to have unit ℓ^2 norm. By elementary algebra, there are constants $c_1, \ldots, c_s$ such that $1_S = c_1\xi_1 + c_2\xi_2 + \ldots + c_s\xi_s$. It follows that

$$C_{SS}^{-1} \cdot 1_S = \sum_{i=1}^{s} \frac{c_i}{\lambda_i}\xi_i \qquad \text{and} \qquad C_{SS} \cdot 1_S = \sum_{i=1}^{s} (c_i\lambda_i)\xi_i.$$

Fix a row of C_{NS}, say, $a = (a_1, \ldots, a_s)$. Respectively, the conditions for the lasso and marginal regression require

$$|(a, \sum_{i=1}^{s} \frac{c_i}{\lambda_i}\xi_i)| \leq 1 \qquad \text{and} \qquad |(a, 1_S)| \leq |\sum_{i=1}^{s} c_i\lambda_i\xi_i|. \tag{13}$$

Without loss of generality, we assume that λ_1 is the smallest eigenvalue of C_{SS}. Consider the case where λ_1 is small, while all other eigenvalues have a magnitude comparable to 1. In this case, the smallness of λ_1 has a negligible effect on $\sum_{i=1}^{s}(c_i\lambda_i)\xi_i$, and so has a negligible effect on the condition for marginal regression. However, the smallness of λ_1 may have an adverse effect on the performance of the lasso. To see the point, we note that $\sum_{i=1}^{s} \frac{c_i}{\lambda_i}\xi_i \approx \frac{c_1}{\lambda_1}\xi_1$. Compare this with the first term in (13). The condition for the lasso is roughly $|(a, \xi_1)| \leq c_1\lambda_1$, which is rather restrictive since λ_1 is small.

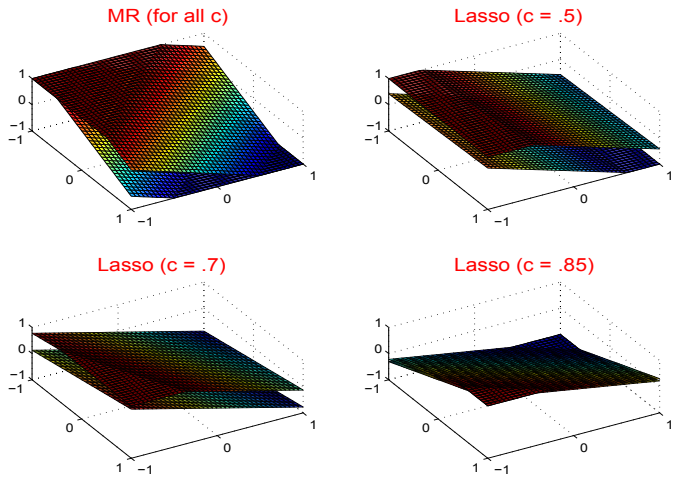

Figure 2: The regions sandwiched by two hyper-planes are the regions of $a = (a_1, a_2, a_3)$ satisfying the respective exact-recovery conditions for marginal regression (MR, panel 1) and for the lasso (panels 2–4). See Section 2.3 Example 4. Here, $c = 0.5, 0.7, 0.85$ and the smallest eigenvalues of C_{SS} are $\lambda_1(c) = 0.29, 0.14, 0.014$. As c varies, the regions for marginal regression remain the same, while the regions for the lasso get substantially smaller.

Figure 2 shows the regions in $a = (a_1, a_2, a_3)$, a row of C_{NS}, where the respective exact recover sequences hold for

$$C_{SS} = \begin{pmatrix} 1 & -1/2 & c \\ -1/2 & 1 & 0 \\ c & 0 & 1 \end{pmatrix}.$$

To better visualize these regions, we display their 2-D section (i.e., setting the first coordinate of a to 0). The Figures suggest that as λ_1 gets smaller, the region corresponding to the lasso shrinks substantially, while that corresponding to marginal regression remains the same.

While the examples in this subsection are low dimensional, they shed light on the high dimensional setting as well. For instance, the approach here can be extended to the following high-dimensional model: (a) $|\mathrm{sgn}(\beta_j)| \overset{iid}{\sim} \mathrm{Bernoulli}(\epsilon)$, (b) each row of the design matrix X are iid samples from $N(0, \Omega/n)$, where Ω is a $p \times p$ correlation matrix that is sparse in the sense that each row of Ω has relatively few coordinates, and (c) $1 \ll p\epsilon \ll n \ll p$ (note that $p\epsilon$ is the expected number of signals). Under this model, it can be shown that (1) C_{SS} is approximately a block-wise diagonal matrix where each block has a relatively small size, and outside these blocks, all coordinates of C_{SS} are uniformly small and have a negligible effect, and (2) each row of C_{NS} has relatively few large coordinates. As a result, the study on the exact reconstruction conditions for the lasso and marginal regression in this more complicated model can be reduced to a low dimensional setting, like those discussed here.

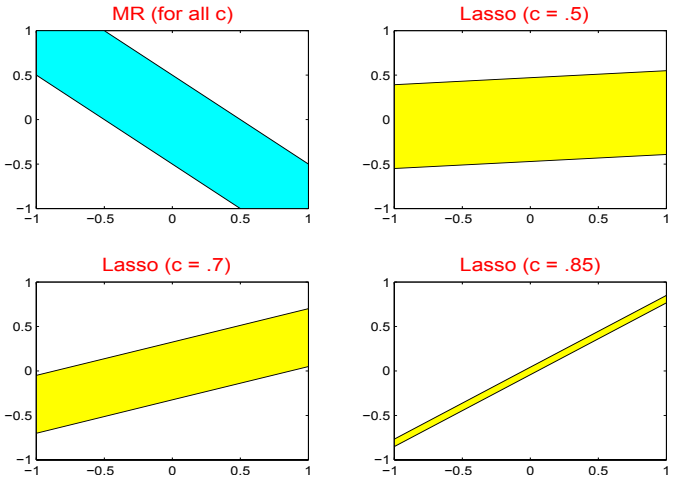

Figure 3: Displayed are the 2-D sections of the regions in Figure 2, where we set the first coordinate of a to 0. As c varies, the regions for marginal regression remain the same, but those for the lasso get substantially smaller as $\lambda_1(c)$ decrease. x-axis: a_2. y-axis: a_3.

And the point that there is no clear winner between the to procedures continues to hold in greater generality.

2.4 Exact reconstruction conditions for marginal regression in the noisy case

We now turn to the noisy case of model (1), taking z to be $N(0, \sigma_n^2 \cdot I_n)$, where we assume that the parameter σ_n^2 is known. The exact reconstruction condition for the lasso in the noisy case has been studied extensively in the literature (see for example [25]). So in this section, we focus on marginal regression. First, we consider a natural extension of Condition F to the noisy case:

Condition F′. $$\max |C_{NS}\beta_S| + 2\sigma_n\sqrt{2\log p} < \min |C_{SS}\beta_S|. \tag{14}$$

Second, when Condition F' holds, we show that with an appropriately chosen threshold t (see (4)), marginal regression fully recovers the support with high probability. Finally, we discuss how to determine the threshold t empirically.

Condition F' implies that it is possible to separate relevant variables from irrelevant variables with high probability. To see this, let $X = [x_1, x_2, \ldots, x_p]$, where x_i denotes the i-th column of X. Sort $|(Y, x_i)|$ in the descending order, and let $r_i = r_i(Y, X)$ be the ranks of $|(Y, x_i)|$ (assume no ties for simplicity). Introduce

$$\widehat{S}_n(k) = \widehat{S}_n(k; X, Y, p) = \{i : \ r_i(X, Y) \le k\}, \qquad k = 1, 2, \ldots, p.$$

Recall that $S(\beta)$ denotes the support of β and $s = |S|$. The following lemma says that, if s is known and Condition F' holds, then marginal regression is able to fully recover the support S with high probability.

Lemma 4 *Consider a sequence of regression models as in (5). If for sufficiently large n, Condition F' holds and $p^{(n)} \geq n$, then*

$$\lim_{n\to\infty} P\left(\widehat{S}_n(s^{(n)}; X^{(n)}, Y^{(n)}, p^{(n)}) \neq S(\beta^{(n)})\right) = 0.$$

Lemma 4 is proved in the appendix. We remark that if both s and $(p-s)$ tend to ∞ as n tends to ∞, then Lemma 4 continues to hold if we replace $2\sigma_n\sqrt{2\log p}$ in (14) by $\sigma_n(\sqrt{\log(p-s)} + \sqrt{\log s})$. See the proof of the lemma for details.

The key assumption of Lemma 4 is that s is known so that we know how to set the threshold t. Unfortunately, s is generally unknown. We propose the following procedure to estimate s. Fix $1 \leq k \leq p$, let i_k be the unique index satisfying $r_{i_k}(X,Y) = k$. Let $\widehat{V}_n(k) = \widehat{V}_n(k; X, Y, p)$ be the linear space spanned by $x_{i_1}, x_{i_2}, \ldots, x_{i_k}$, and let $\widehat{H}_n(k) = \widehat{H}_n(k; X, Y, p)$ be the projection matrix from R^n to $\widehat{V}_n(k)$ (here and below, the $\widehat{\ }$ sign emphasizes the dependence of indices i_k on the data). Define

$$\widehat{\delta}_n(k) = \widehat{\delta}_n(k; X, Y, p) = \|(\widehat{H}_n(k+1) - \widehat{H}_n(k))Y\|, \qquad 1 \leq k \leq p-1.$$

The term $\widehat{\delta}_n^2(k)$ is closely related to the F-test for testing whether $\beta_{i_{k+1}} \neq 0$. We estimate s by

$$\widehat{s}_n = \widehat{s}_n(X, Y, p) = \max\left\{1 \leq k \leq p : \ \widehat{\delta}_n(k) \geq \sigma_n\sqrt{2\log n}\right\} + 1$$

(in the case where $\widehat{\delta}_n(k) < \sigma_n\sqrt{2\log n}$ for all k, we define $\widehat{s}_n = 1$).

Once $\widehat{s}_n$ is determined, we estimate the support S by

$$\widehat{S}(\widehat{s}_n, X, Y, p) = \{i_k : \ k = 1, 2, \ldots, \widehat{s}_n\}.$$

It turns out that under mild conditions, $\widehat{s}_n = s$ with high probability. In detail, suppose that the support $S(\beta)$ consists of indices $j_1, j_2, \ldots, j_s$. Fix $1 \leq k \leq s$. Let $\widetilde{V}_S$ be the linear space spanned by $x_{j_1}, \ldots, x_{j_s}$, and let $\widetilde{V}_{S,(-k)}$ be the linear space spanned by $x_{j_1}, \ldots, x_{j_{k-1}}, x_{j_{k+1}}, \ldots, x_{j_s}$. Project $\beta_{j_k}x_{j_k}$ to the linear space $\widetilde{V}_S \cap \widetilde{V}_{S,(-k)}^{\perp}$. Let $\Delta_n(k, \beta, X, p)$ be the ℓ^2 norm of the resulting vector (which can be interpreted as the strength of the k-th signal after the collinearity between the k-th predictor and other predictors removed), and let

$$\Delta_n^*(\beta, X, p) = \min_{1\leq k\leq s} \Delta_n(k, \beta, X, p).$$

The following theorem says that if $\Delta_n^*(\beta, X, p)$ is slightly larger than $\sigma_n\sqrt{2\log n}$, then $\widehat{s}_n = s$ and $\widehat{S}_n = S$ with high probability. In other words, marginal regression fully recovers the support with high probability. Theorem 5 is proved in the appendix,

Theorem 5 *Consider a sequence of regression models as in (5). Suppose that for sufficiently large n, Condition F' holds, $p^{(n)} \geq n$, and*

$$\lim_{n\to\infty}\left(\frac{\Delta_n^*(\beta^{(n)}, X^{(n)}, p^{(n)})}{\sigma_n} - \sqrt{2\log n}\right) = \infty.$$

Then

$$\lim_{n\to\infty} P\left(\widehat{s}_n(X^{(n)}, Y^{(n)}, p^{(n)}) \neq s^{(n)}\right) \to 0,$$

and

$$\lim_{n\to\infty}\left(\widehat{S}_n(\widehat{s}_n(X^{(n)}, Y^{(n)}, p^{(n)}); X^{(n)}, Y^{(n)}, n, p^{(n)}) \neq S(\beta^{(n)})\right) \to 0.$$

Theorem 5 says that the tuning parameter for marginal regression (i.e. the threshold t) can be set successfully in a data driven fashion. In comparison, how to set the tuning parameter λ for the lasso has been a longstanding open problem in the literature.

We briefly discuss the case where the noise variance σ_n^2 is unknown. The topic is addressed in some of recent literature (e.g. [4, 32]). It is noteworthy that in some applications, σ_n^2 can be calibrated during data collection and so it can be assumed as known [4, Rejoinder]. It is also noteworthy that in [32], Sun and Zhang proposed a procedure to jointly estimate β and σ_n^2 using scaled lasso. The estimator was show to be consistent with σ_n^2 in rather general situations, but unfortunately it is computationally more expensive than either the lasso or marginal regression. How to find an estimator that is consistent with σ_n^2 in general situations and has low computational cost remains an open problem, and we leave the study to the future.

With that being said, we conclude this section by mentioning that both the lasso and marginal regression have their strengths and weakness, and there is no clear winner between these two in general settings. For a given data set, whether to use one or the other is a case by case decision, where a close investigation of (X, β) is usually necessary.

3. The Deterministic Design, Random Coefficient Regime

In this section, we study how generally the Faithfulness Condition holds. We approach this question by modeling the coefficients β as random (the matrix X remains fixed) and deriving a condition (F$''$) under which the Faithfulness Condition holds with high probability. The discussion in this section is closely related to the work by Donoho and Elad [8] on the Incoherence condition. Compared to the Faithfulness Condition, the advantage of the Incoherence Condition is that it does not involve the unknown support of β, so it is checkable in practice. The downside of the Incoherence Condition is that it aims to control the worst case so it is conservative. In this section, we derive a condition—Condition F"—which can be viewed as a middle ground between the Faithfulness Condition and the Incoherence Condition: it is not tied to the unknown support so it is more tractable than the Faithfulness Condition, and it is also much less stringent than the Incoherence Condition.

In detail, we model β as follows. Fix $\epsilon \in (0, 1)$, $a > 0$, and a distribution π, where

$$\text{the support of } \pi \subset (-\infty, -a] \cup [a, \infty). \tag{15}$$

For each $1 \leq i \leq p$, we draw a sample B_i from Bernoulli(ϵ). When $B_i = 0$, we set $\beta_i = 0$. When $B_i = 1$, we draw $\beta_i \sim \pi$. Marginally,

$$\beta_i \overset{iid}{\sim} (1 - \epsilon)\nu_0 + \epsilon\pi, \tag{16}$$

where ν_0 denotes the point mass at 0. This models the case where we have no information on the signals, so they appear at locations generated randomly. In the literature, it is known

that the least favorable distribution for variable selection has the form as in (16), where π is in fact degenerate. See [3] for example.

We study for which quadruples (X, ϵ, π, a) the Faithfulness Condition holds with high probability. Recall that the design matrix $X = [x_1, \ldots, x_p]$, where x_i denotes the i-th column. Fix $t \geq 0$ and $\delta > 0$. Introduce

$$g_{ij}(t) = E_\pi[e^{tu\cdot(x_i,x_j)}] - 1, \qquad \bar{g}_i(t) = \sum_{j \neq i} g_{ij}(t),$$

where the random variable $u \sim \pi$ and (x_i, x_j) denotes the inner product of x_i and x_j. As before, we have suppressed the superscript $^{(n)}$ for $g_{ij}(t)$ and $\bar{g}_i(t)$. Define

$$A_n(\delta, \epsilon, \bar{g}) = A_n(\delta, \epsilon, \bar{g}; X, \pi) = \min_{t>0}\left(e^{-\delta t}\sum_{i=1}^{p}[e^{\epsilon\bar{g}_i(t)} + e^{\epsilon\bar{g}_i(-t)}]\right),$$

where $\bar{g}$ denotes the vector $(\bar{g}_1, \ldots, \bar{g}_p)^T$. Note that $1 + g_{ij}(t)$ is the moment generating function of π evaluated at the point $(x_i, x_j)t$. In the literature, it is conventional to use moment generating function to derive sharp inequalities on the tail probability of sums of independent random variables. The following lemma is proved in the appendix.

Lemma 6 *Fix n, X, $\delta > 0$, $\epsilon \in (0, 1)$, and distribution π. Then*

$$P(\max|C_{NS}\beta_S| \geq \delta) \leq (1 - \epsilon)A_n(\delta, \epsilon, \bar{g}; X, \pi), \tag{17}$$

and

$$P(\max|(C_{SS} - I_S)\beta_S| \geq \delta) \leq \epsilon A_n(\delta, \epsilon, \bar{g}; X, \pi). \tag{18}$$

Now, suppose the distribution π satisfies (15) for some $a > 0$. Take $\delta = a/2$ on the right hand side of (17)-(18). Except for a probability of $A_n(a/2, \epsilon, \bar{g})$,

$$\max|C_{NS}\beta_S| \leq a/2, \qquad \min|C_{SS}\beta_S| \geq \min|\beta_S| - \max|(C_{SS} - I)\beta_S| \geq a/2,$$

so $\max|C_{NS}\beta_S| \leq \min|C_{SS}\beta_S|$ and the Faithfulness Condition holds. This motivates the following condition, where (a, ϵ, π) may depend on n.

Condition F$''$. $\lim_{n\to\infty} A_n(a_n/2, \epsilon_n, \bar{g}^{(n)}; X^{(n)}, \pi_n) = 0.$ (19)

The following theorem says that if Condition F" holds, then Condition F holds with high probability.

Theorem 7 *Consider a sequence of noise-free regression models as in (5), where the noise component $z^{(n)} = 0$ and $\beta^{(n)}$ is generated as in (16). Suppose Condition F" holds. Then as n tends to ∞, except for a probability that tends to 0,*

$$\max|C_{NS}\beta_S| \leq \min|C_{SS}\beta_S|.$$

Theorem 7 is the direct result of Lemma 6 so we omit the proof.

3.1 Comparison of Condition F" with the Incoherence Condition

Introduced in Donoho and Elad [8] (see also [9]), the *Incoherence* of a matrix X is defined as

$$\max_{i \neq j} |C_{ij}|,$$

where $C = X^T X$ is the Gram matrix as before. The notion is motivated by the study in recovering a sparse signal from an over-complete dictionary. In the special case where X is the concatenation two orthonormal bases (e.g. a Fourier basis and a wavelet basis), $\max_{i \neq j} |C_{ij}|$ measures how coherent two bases are and so the term of incoherence; see [8, 9] for details. Consider Model (1) in the case where both X and β are deterministic, and the noise component $z = 0$. The following results are proved in [5, 8, 9].

- The lasso yields exact variable selection if $s < \frac{1+\max_{i \neq j} |C_{ij}|}{2 \max_{i \neq j} |C_{ij}|}$.
- Marginal regression yields exact variable selection if $s < \frac{c}{2 \max_{i \neq j} |C_{ij}|}$ for some constant $c \in (0, 1)$, and that the nonzero coordinates of β have comparable magnitudes (i.e. the ratio between the largest and the smallest nonzero coordinate of β is bounded away from ∞).

In comparison, the Incoherence Condition only depends on X so it is checkable. Condition F depends on the unknown support of β. Checking such a condition is almost as hard as estimating the support S. Condition F" provides a middle ground. It depends on β only through (ϵ, π). In cases where we either have a good knowledge of (ϵ, π) or we can estimate them, Condition F" is checkable (for literature on estimating (ϵ, π), see [17, 29] for the case where we have an orthogonal design, and [19, Section 2.6] for the case where $X^T X$ is sparse in the sense that each row of $X^T X$ has relatively few large coordinates. We note that even when successful variable selection is impossible, it may be still possible to estimate (ϵ, π) well).

At the same time, the Incoherence Condition is conservative, especially when s is large. In fact, in order for either the lasso or marginal regression to have an exact variable selection, it is required that

$$\max_{i \neq j} |C_{ij}| \leq O\left(\frac{1}{s}\right), \tag{20}$$

In other words, all coordinates of the Gram matrix C need to be no greater than $O(1/s)$. This is much more conservative than Condition F.

However, we must note that the Incoherence Condition aims to control the worst case: it sets out to guarantee *uniform* success of a procedure across all β under minimum constraints. In comparison, Condition F aims to control a single case, and Condition F" aims to control almost all the cases in a specified class. As such, Condition F" provides a middle ground between Condition F and the Incoherence Condition, applying more broadly than the former, while being less conservative than the later.

Below, we use two examples to illustrate that Condition F" is much less conservative than the Incoherence Condition. In the first example, we consider a weakly dependent case where $\max_{i \neq j} |C_{ij}| \leq O(1/\log(p))$. In the second example, we suppose the matrix C is sparse, but the nonzero coordinates of C may be large.

3.1.1 The weakly dependent case

Suppose that for sufficiently large n, there are two sequence of positive numbers $a_n \leq b_n$ such that the support of π_n is contained in $[-b_n, -a_n] \cup [a_n, b_n]$, and that

$$\frac{b_n}{a_n} \cdot \max_{i \neq j} |C_{ij}| \leq c_1 / \log(p), \qquad c_1 > 0 \text{ is a constant.}$$

For $k \geq 1$, denote the k-th moment of π_n by

$$\mu_n^{(k)} = \mu_n^{(k)}(\pi_n). \tag{21}$$

Introduce $m_n = m_n(X)$ and $v_n^2 = v_n^2(X)$ by

$$m_n(X) = p\epsilon_n \cdot \max_{1 \leq i \leq p} \Big\{ \Big| \frac{1}{p} \sum_{j \neq i} C_{ij} \Big| \Big\}, \qquad v_n^2(X) = p\epsilon_n \cdot \max_{1 \leq i \leq p} \Big\{ \frac{1}{p} \sum_{j \neq i} C_{ij}^2 \Big\}.$$

Corollary 3.1 *Consider a sequence of regression models as in (5), where the noise component $z^{(n)} = 0$ and $\beta^{(n)}$ is generated as in (16). If there are constants $c_1 > 0$ and $c_2 \in (0, 1/2)$ such that*

$$\frac{b_n}{a_n} \cdot \max_{i \neq j} \{|C_{ij}|\} \leq c_1 / \log(p^{(n)}),$$

and

$$\overline{\lim_{n \to \infty}} \Big(\frac{\mu_n^{(1)}(\pi_n)}{a_n} m_n(X^{(n)}) \Big) \leq c_2, \qquad \overline{\lim_{n \to \infty}} \Big(\frac{\mu_n^{(2)}(\pi_n)}{a_n^2} v_n^2(X^{(n)}) \log(p^{(n)}) \Big) = 0, \tag{22}$$

then

$$\lim_{n \to \infty} A_n(a_n/2, \epsilon_n, \bar{g}^{(n)}; X^{(n)}, \pi_n) = 0,$$

and Condition F" holds.

Corollary 3.1 is proved in the appendix. For interpretation, we consider the special case where there is a generic constant $c > 0$ such that $b_n \leq c a_n$. As a result, $\mu_n^{(1)}/a_n \leq c$, $\mu_n^{(2)}/a_n^2 \leq c^2$. The conditions reduce to that, for sufficiently large n and all $1 \leq i \leq p$,

$$|\frac{1}{p} \sum_{j \neq i}^{p} C_{ij}| \leq O(\frac{1}{p\epsilon_n}), \qquad \frac{1}{p} \sum_{j \neq i}^{p} C_{ij}^2 = o(1/p\epsilon_n).$$

Note that by (16), $s = s^{(n)} \sim \text{Binomial}(p, \epsilon_n)$, so $s \approx p\epsilon_n$. Recall that the Incoherence Condition is

$$\max_{i \neq j} |C_{ij}| \leq O(1/s).$$

In comparison, the Incoherence Condition requires that each coordinate of $(C - I)$ is no greater than $O(1/s)$, while Condition F" only requires that the average of each row of $(C - I)$ is no greater than $O(1/s)$. The latter is much less conservative.

3.1.2 The sparse case

Let $N_n^*(C)$ be the maximum number of nonzero off-diagonal coordinates of C:

$$N_n^*(C) = \max_{1\le i\le p}\{N_n(i)\}, \qquad N_n(i) = N_n(i;C) = \#\{j:\ j\neq i, C_{ij}\neq 0\}.$$

Suppose there is a constant $c_3 > 0$ such that

$$\underline{\lim}_{n\to\infty}\left(\frac{-\log(\epsilon_n N_n^*(C))}{\log(p^{(n)})}\right) \ge c_3. \tag{23}$$

Also, suppose there is a constant $c_4 > 0$ such that for sufficiently large n,

$$\text{the support of } \pi_n \text{ is contained in } [-c_4 a_n, a_n]\cup[a_n, c_4 b_n]. \tag{24}$$

The following corollary is proved in the appendix.

Corollary 3.2 *Consider a sequence of noise-free regression models as in (5), where the noise component $z^{(n)} = 0$ and $\beta^{(n)}$ is randomly generated as in (16). Suppose (23)-(24) hold. If there is a constant $\delta > 0$ such that*

$$\max_{i\neq j}|C_{ij}| \le \delta, \qquad \textit{and} \qquad \delta < \frac{c_3}{2c_4}, \tag{25}$$

then

$$\lim_{n\to\infty} A_n(a_n/2, \epsilon_n, \bar{g}^{(n)}; X^{(n)}, \pi_n) = 0,$$

and Condition F" holds.

For interpretation, consider a special case where $\epsilon_n = p^{-\vartheta}$. In this case, the condition reduces to $N_n^*(C) \ll p^{\vartheta - 2c_4\delta}$. As a result, Condition F" is satisfied if each row of $(C - I)$ contains no more than $p^{\vartheta-2c_4\delta}$ nonzero coordinates each of which $\le \delta$. Compared to the Incoherence Condition $\max_{i\neq j}|C_{ij}| \le O(1/s) = O(p^{-\vartheta})$, our condition is much weaker.

In conclusion, if we alter our attention from the worst-case scenario to the average scenario, and alter our aim from exact variable selection to exact variable selection with probability ≈ 1, then the condition required for success—Condition F"—is much more relaxed than the Incoherence Condition.

4. Hamming Distance for the Gaussian Design and the Phase Diagram

So far, we have focused on exact variable selection. In many applications, exact variable selection is not possible. Therefore, it is of interest to study the Type I and Type II errors of variable selection (a Type I error is a misclassified 0 coordinate of β, and a Type II error is a misclassified nonzero coordinate).

In this section, we use the Hamming distance to measure the variable selection errors. Back to Model (1),

$$Y = X\beta + z, \qquad z\sim N(0, I_n), \tag{26}$$

where without loss of generality, we assume $\sigma_n = 1$. As in the preceding section (i.e. (16)), we suppose

$$\beta_i \overset{iid}{\sim} (1-\epsilon)\nu_0 + \epsilon\pi. \tag{27}$$

For any variable selection procedure $\widehat{\beta} = \widehat{\beta}(Y; X)$, the Hamming distance between $\widehat{\beta}$ and the true β is

$$d(\widehat{\beta}|X) = d(\widehat{\beta}; \epsilon, \pi|X) = \sum_{j=1}^{p} E_{\epsilon,\pi}(E_z[1(\mathrm{sgn}(\widehat{\beta}_j) \neq \mathrm{sgn}(\beta_j))|X]).$$

Note that by Chebyshev's inequality,

$$P(\text{non-exact variable selection by } \widehat{\beta}(Y; X)) \leq d(\widehat{\beta}|X).$$

So a small Hamming distance guarantees exact variable selection with high probability.

How to characterize precisely the Hamming distance is a challenging problem. We approach this by modeling X as random. Assume that the coordinates of X are iid samples from $N(0, 1/n)$:

$$X_{ij} \overset{iid}{\sim} N(0, 1/n). \tag{28}$$

The choice of the variance ensures that most diagonal coordinates of the Gram matrix $C = X^T X$ are approximately 1. Let $P_X(x)$ denote the joint density of the coordinates of X. The expected Hamming distance is then

$$d^*(\widehat{\beta}) = d^*(\widehat{\beta}; \epsilon, \pi) = \int d(\widehat{\beta}; \epsilon, \pi|X = x) P_X(x) dx.$$

We adopt an asymptotic framework where we calibrate p and ϵ with

$$p = n^{1/\theta}, \qquad p\epsilon_n = n^{(1-\vartheta)/\theta} \equiv p^{1-\vartheta}, \qquad 0 < \theta, \vartheta < 1. \tag{29}$$

This models a situation where $p \gg n$ and the vector β gets increasingly sparse as n grows. Note that the parameter ϑ calibrates the sparsity level of the signals. We assume π_n in (16) is a point mass

$$\pi_n = \nu_{\tau_n}. \tag{30}$$

In the literature (e.g. [10, 21]), this model was found to be subtle and rich in theory. In addition, compare two experiments, in one of them $\pi_n = \nu_{\tau_n}$, and in the other the support of π_n is contained in $[\tau_n, \infty)$. Since the second model is easier for inference than the first one, the optimal Hamming distance for the first one gives an upper bound for that for the second one.

With ϵ_n calibrated as above, the most interesting range for τ_n is $O(\sqrt{2\log p})$: when $\tau_n \gg \sqrt{2\log p}$, exact variable selection can be easily achieved by either the lasso or marginal regression. When $\tau_n \ll \sqrt{2\log p}$, no variable selection procedure can achieve exact variable selection. See for example [10]. In light of this, we calibrate

$$\tau_n = \sqrt{2(r/\theta)\log n} \equiv \sqrt{2r\log p}, \qquad r > 0. \tag{31}$$

Note that the parameter r calibrates the signal strength. With these calibrations, we can rewrite

$$d_n^*(\widehat{\beta}; \epsilon, \pi) = d_n^*(\widehat{\beta}; \epsilon_n, \tau_n).$$

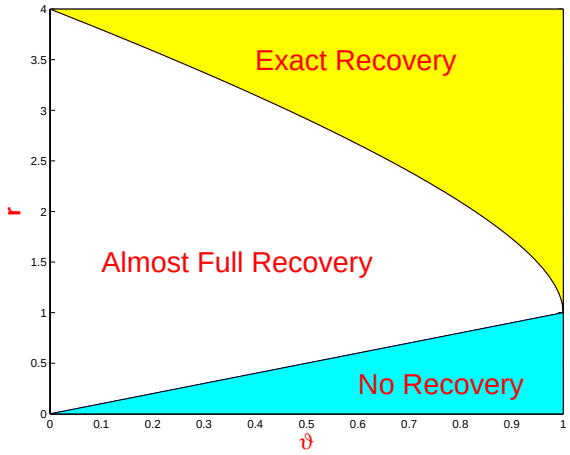

Figure 4: The regions as described in Section 4. In the region of Exact Recovery, both the lasso and marginal regression yield exact recovery with high probability. In the region of Almost Full Recovery, it is impossible to have large probability for exact variable selection, but the Hamming distance of both the lasso and marginal regression $\ll p\epsilon_n$. In the region of No Recovery, optimal Hamming distance $\sim p\epsilon_n$ and all variable selection procedures fail completely. Displayed is the part of the plane corresponding to $0 < r < 4$ only.

Definition 8 *Denote $L(n)$ by a multi-log term which satisfies that $\lim_{n\to\infty}(L(n)\cdot n^{\delta}) = \infty$ and that $\lim_{n\to\infty}(L(n)\cdot n^{-\delta}) = 0$ for any $\delta > 0$.*

We are now ready to spell out the main results. Define

$$\rho(\vartheta) = (1+\sqrt{1-\vartheta})^2, \qquad 0 < \vartheta < 1.$$

The following theorem is proved in the appendix, which gives the lower bound for the Hamming distance.

Theorem 9 *Fix $\vartheta \in (0,1)$, $\theta > 0$, and $r > 0$ such that $\theta > 2(1-\vartheta)$. Consider a sequence of regression models as in (26)-(31). As $n \to \infty$, for any variable selection procedure $\widehat{\beta}^{(n)}$,*

$$d_n^*(\widehat{\beta}^{(n)}; \epsilon_n, \tau_n) \geq \begin{cases} L(n)p^{1-\frac{(\vartheta+r)^2}{4r}}, & r \geq \vartheta, \\ (1+o(1))\cdot p^{1-\vartheta}, & 0 < r < \vartheta. \end{cases}$$

Let $\widehat{\beta}_{mr}$ be the estimate of using marginal regression with threshold

$$t_n = \begin{cases} \frac{\vartheta+r}{2\sqrt{r}}\sqrt{2\log p}, & \text{if } r > \vartheta, \\ t_n = \sqrt{2\widetilde{r}\log p}, & \text{if } r < \vartheta, \end{cases} \tag{32}$$

where $\widetilde{r}$ is some constant $\in (\vartheta, 1)$ (note that in the case of $r < \vartheta$, the choice of t_n is not necessarily unique). We have the following theorem.

Theorem 10 *Fix $\vartheta \in (0,1)$, $r > 0$, and $\theta > (1-\vartheta)$. Consider a sequence of regression models as in (26)-(31). As $p \to \infty$, the Hamming distance of marginal regression with the threshold t_n given in (32) satisfies*

$$d_n^*(\widehat{\beta}_{mr}^{(n)}; \epsilon_n, \tau_n) \leq \begin{cases} L(n)p^{1-\frac{(\vartheta+r)^2}{4r}}, & r \geq \vartheta, \\ (1+o(1)) \cdot p^{1-\vartheta}, & 0 < r < \vartheta. \end{cases}$$

In practice, the parameters (ϑ, r) are usually unknown, and it is desirable to set t_n in a data-driven fashion. Towards this end, we note that our primary interest is in the case of $r > \vartheta$ (as when $r < \vartheta$, successful variable selection is impossible). In this case, the optimal choice of t_n is $(\vartheta + r)/(2r)\tau_p$, which is the Bayes threshold in the literature. The Bayes threshold can be set by the approach of controlling the local False Discovery Rate (Lfdr), where we set the FDR-control parameter as $1/2$; see Efron *et al.* [13] for details.

Similarly, choosing the tuning parameter $\lambda_n = 2(\frac{\vartheta+r}{2\sqrt{r}} \wedge \sqrt{r})\sqrt{2\log p}$ in the lasso, we have the following theorem.

Theorem 11 *Fix $\vartheta \in (0,1)$, $r > 0$, and $\theta > (1-\vartheta)$. Consider a sequence of regression models as in (26)-(31). As $p \to \infty$, the Hamming distance of the lasso with the tuning parameter $\lambda_n = 2t_n$ where t_n is given in (32), satisfies*

$$d_n^*(\widehat{\beta}_{lasso}^{(n)}; \epsilon_n, \tau_n) \leq \begin{cases} L(n)p^{1-\frac{(\vartheta+r)^2}{4r}}, & r \geq \vartheta, \\ (1+o(1)) \cdot p^{1-\vartheta}, & 0 < r < \vartheta. \end{cases}$$

The proofs of Theorems 10-11 are routine and we omit them.

Theorems 9-11 say that in the ϑ-r plane, we have three different regions, as displayed in Figure 4.

- Region I (*Exact Reovery*): $0 < \vartheta < 1$ and $r > \rho(\vartheta)$.
- Region II (*Almost Full Recovery*): $0 < \vartheta < 1$ and $\vartheta < r < \rho(\vartheta)$.
- Region III (*No Recovery*): $0 < \vartheta < 1$ and $0 < r < \vartheta$.

In the Region of Exact Recovery, the Hamming distance for both marginal regression and the lasso are algebraically small. Therefore, except for a probability that is algebraically small, both marginal regression and the lasso give exact recovery.

In the Region of Almost Full Recovery, both the Hamming distance of marginal regression and the lasso are much smaller than the number of relevant variables (which $\approx p\epsilon_n$). Therefore, almost all relevant variables have been recovered. Note also that the number of misclassified irrelevant variables is comparably much smaller than $p\epsilon_n$. In this region, the optimal Hamming distance is algebraically large, so for any variable selection procedure, the probability of exact recovery is algebraically small.

In the Region of No Recovery, the Hamming distance $\sim p\epsilon_n$. In this region, asymptotically, it is impossible to distinguish relevant variables from irrelevant variables, and any variable selection procedure fails completely.

In practice, given a data set, one wishes to know that which of these three regions the true parameters belong to. Towards this end, we note that in the current model, the coordinates of X^TY are approximately iid samples from the following two-component Gaussian mixture

$$(1-\epsilon_p)\phi(x) + \epsilon_n\phi(x-\tau_n),$$

where $\phi(x)$ denotes the density of $N(0,1)$. In principle, the parameters (ϵ_n, τ_n) can be estimated (see the comments we made in Section 3.1 on estimating (ϵ, π)). The estimation can then be used to determine which regions the true parameters belong to.

	$k=4$		$k=10$	
$(a_1,\, a_2)$	lasso	MR	lasso	MR
(0, 0)	0	0	0.8	3.8
(-0.85, 0.85)	0	4	0.6	10.4
(0.85, -0.85)	0	4	0.6	11.2
(-0.4, 0.8)	4	0	10	3.6
(0.4, -0.8)	4	0	10	4.8

Table 1: Comparison of the lasso and marginal regression for different choices of (a_1, a_2) and k. The setting is described in Experiment 1a. Each cell displays the corresponding Hamming error.

The results improve on those by Wainwright [28]. It was shown in [28] that there are constants $c_2 > c_1 > 0$ such that in the region of $\{0 < \vartheta < 1, r > c_2\}$, the lasso yields exact variable selection with overwhelming probability, and that in the region of $\{0 < \vartheta < 1, r < c_2\}$, no procedure could yield exact variable selection. Our results not only provide the exact rate of the Hamming distance, but also tighten the constants c_1 and c_2 so that $c_1 = c_2 = (1+\sqrt{1-\vartheta})^2$. The lower bound argument in Theorem 9 is based on computing the L^1-distance. This gives better results than in [28] which uses Fano's inequality in deriving the lower bounds.

To conclude this section, we briefly comment on the phase diagram in two closely related settings. In the first setting, we replace the identity matrix I_p in (28) by some general correlation matrix Ω, but keep all other assumptions unchanged. In the second setting, we assume that as $n \to \infty$, both ratios $p\epsilon_p/n$ and n/p tend to a constant in $(0,1)$, while all other assumptions remain the same. For the first setting, it was shown in [19] that the phase diagram remains the same as in the case of $\Omega = I_p$, provided that Ω is sparse; see [19] for details. For the second setting, the study is more more delicate, so we leave it to the future work.

5. Simulations and Examples

We conducted a small-scale simulation study to compare the performance of the lasso and marginal regression. The study includes three different experiments (some have more than one sub-experiments). In the first experiment, the rows of X are generated from $N(0, \frac{1}{n}C)$ where C is a diagonal block-wise matrix. In the second one, we take the Gram matrix

$C = X'X$ to be a tridiagonal matrix. In the third one, the Gram matrix has the form of $C = \Lambda + a\xi\xi'$ where Λ is a diagonal matrix, $a > 0$, and ξ is a $p \times 1$ unit-norm vector. Intrinsically, the first two are covered in the theoretic discussion in Section 2.3, but the last one goes beyond that. Below, we describe each of these experiments in detail.

		$k = 2$			$k = 7$		
Method	(a_2, a_3)	$c = 0.5$	$c = 0.7$	$c = 0.85$	$c = 0.5$	$c = 0.7$	$c = 0.85$
MR	(0,0)	0	0	0	3	3.8	4.6
Lasso	(0, 0)	0	0	2	0	0	7
MR	(-0.4, -0.1)	1	1	1	5.4	5.8	5.4
Lasso	(-0.4, -0.1)	0	0	2	0.4	2	7
MR	(0.4, 0.1)	1	1.2	1.2	5.4	5.8	6
Lasso	(0.4, 0.1)	0	0	2	1.2	1.4	7.6
MR	(-0.5, -0.4)	2	2	2	9.6	7.8	7.6
Lasso	(-0.5, -0.4)	1	0	2	3.6	0.2	7
MR	(0.5, 0.4)	2	2	2	9.4	7.4	7.8
Lasso	(0.5, 0.4)	1	0	2	3.4	0	7

Table 2: Comparison of the lasso and marginal regression for different choices of (c, a_2, a_3). The setting is described in Experiment 1b. Each cell displays the corresponding Hamming error.

Experiment 1. In this experiment, we compare the performance of the lasso and marginal regression with the noiseless linear model $Y = X\beta$. We generate the rows of X as iid samples from $N(0, (1/n)C)$, where C is a diagonal block-wise correlation matrix having the form

$$C = \begin{pmatrix} C_{\text{sub}} & 0 & 0 & \dots 0 \\ 0 & C_{\text{sub}} & 0 & \dots 0 \\ & \cdots & \cdots & \\ 0 & 0 & 0 & \dots C_{\text{sub}} \end{pmatrix}.$$

Fixing a small integer m, we take C_{sub} to be the $m \times m$ matrix as follows:

$$C_{\text{sub}} = \begin{pmatrix} D & a^T \\ a & 1 \end{pmatrix},$$

where a is an $(m-1) \times 1$ vector and D is an $(m-1) \times (m-1)$ matrix to be introduced below. Also, fixing another integer $k \geq 1$, according to the block-wise structure of C, we let β be the vector (without loss of generality, we assume p is divisible by m)

$$\beta = (\delta_1 u^T, \delta_2 u^T, \dots, \delta_{p/m} u^T)^T,$$

where $u = (v^T, 0)$ for some $(m-1) \times 1$ vector v and $\delta_i = 0$ for all but k different i, where $\delta_i = 1$.

The goal of this experiment is to investigate how the theoretic results in Section 2.3 shed light on models with more practical interests. To see the point, note that when $k \ll n$, the

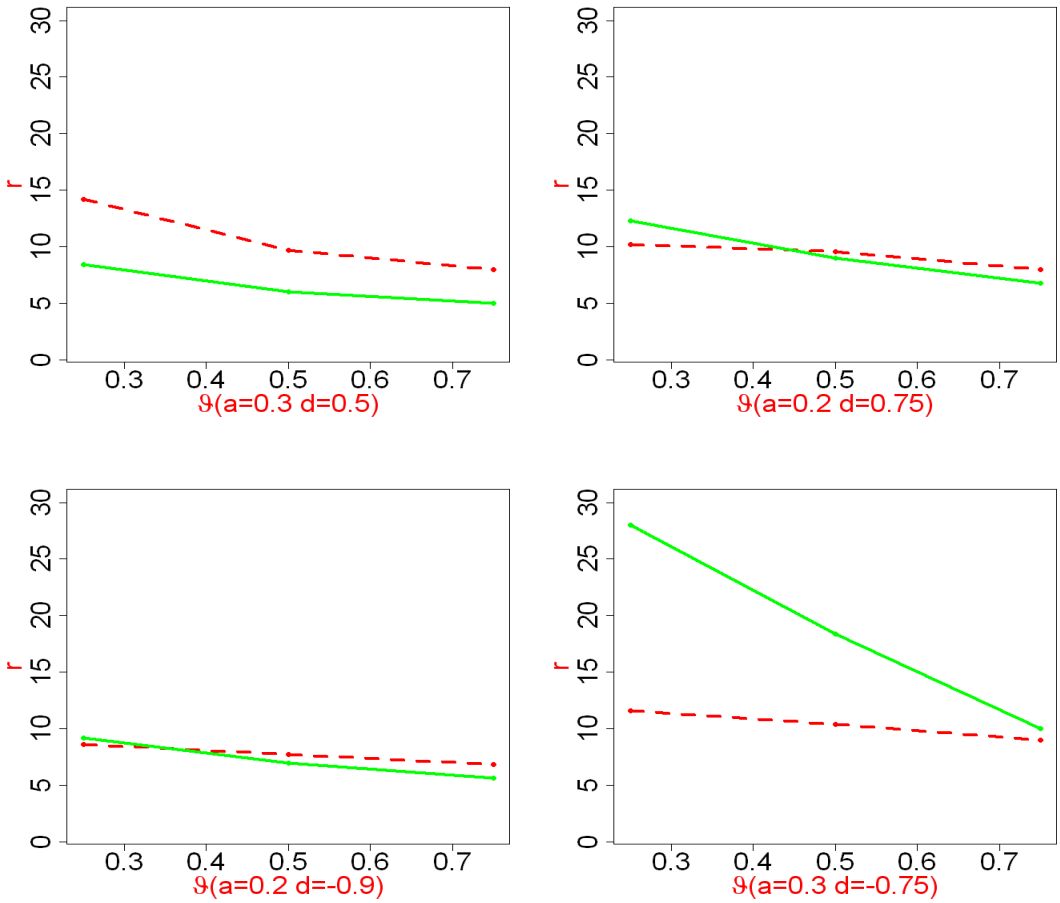

Figure 5: Critical values of exact recovery for the lasso (red dashed) and marginal regression (green solid). See Experiment 2 for the setting and the definition of critical value. For any given set of parameters (ϑ, a, d), the method with a smaller critical value has the better performance in terms of Hamming errors.

signal vector β is sparse, and we expect to see that

$$X'X\beta \approx C\beta, \tag{33}$$

where the right hand side corresponds to the idealized model where $X'X = C$. In this idealized model, if we restrict our attention to any block where the corresponding δ_i is 1, then we have exactly the same model as in Example 1 of Section 2.3, with $C_{SS} = D$ and $\beta_S = v$. As a result, the theoretic results discussed in Section 2.3 apply, at least when the approximation error in equation (33) is negligible. Experiment 1 contains two sub-experiments, Experiment 1a and 1b.

In Experiment 1a, we take $(p, n, m) = (999, 900, 3)$. At the same time, for some numbers a_1 and a_2, we set a, v, and D by

$$a = (a_1, a_2)^T, \qquad v = (2, 1)^T, \qquad D = \begin{pmatrix} 1 & -.75 \\ -.75 & 1 \end{pmatrix}.$$

We investigate the experiment with two different values of k ($k = 4$ and $k = 10$) and five different choices of (a_1, a_2): $(0, 0), \pm(-0.85, 0.85)$, and $\pm(-0.4, 0.8)$. When $k = 4$, we let $\delta_i = 1$ if and only if $i \in \{40, 208, 224, 302\}$, and when $k = 10$, we let $\delta_i = 1$ if and only if $i \in \{20, 47, 83, 86, 119, 123, 141, 250, 252, 281\}$ (such indices are generated randomly; also, note that i are the indices for the blocks, not the indices for the signals).

Consider for a second the idealized case where $X'X = C$ (i.e., n is very large). If we restrict our attention to any block of β where the corresponding δ_i is 1, the setting reduces to that of Example 1 of Section 2.3. In fact, in Figure 1, our first choice of (a_1, a_2) falls inside both the red box and green box, our next two choices fall inside the green box but outside the red box, and our last two choices fall outside the green box but inside the red box. Therefore, at least when k is sufficiently small (so that the setting can be well-approximated by that in the idealized case), we expect to see that the lasso outperforms marginal regression with the second and the third choices, and expect to see the other way around with the last two choices of (a_1, a_2). In the first choice, both methods are expected to perform well.

We now investigate how well these expectations are met. For each combination of these parameters, we generate data and compare the Hamming errors of the lasso and marginal regression, where for each method, the tuning parameters are set ideally. The 'ideal' tuning parameter is obtained through rigorous search from a range. The error rates over 10 repetitions are tabulated in Table 1. More repetitions is unnecessary, partially because the standard deviations of the simulation results are small, and partially because the program is slow (for that we need to choose the 'ideal' tuning parameter through rigorous search. Take the lasso for example. For rigorous search of the 'ideal' tuning parameter, we need to run the glmnet R package many times).

The results suggest that the performances of each method are reasonably close to what are expected for the idealized model, especially in the case of $k = 4$. Take the cases $(a_1, a_2) = \pm(0.85, -0.85)$ for example. The lasso yields exact recovery, while marginal regression, in each of the four blocks where the corresponding δ_i is 1, recovers correctly the stronger signal and mistakenly kills the weaker one. The situation is reversed in the cases where $(a_1, a_2) = \pm(0.4, -0.8)$. The discussion for the case of $k = 10$ is similar, but the approximation error in equation (33) starts to kick in.

In Experiment 1b, we take $(p, n, m) = (900, 1000, 4)$. Also, for some numbers c, a_2, and a_3, we set a, v, and D as

$$a^T = (0, a_2, a_3)^T, \qquad v = (1, 1, 1)^T, \qquad D = \begin{pmatrix} 1 & -1/2 & c \\ -1/2 & 1 & 0 \\ c & 0 & 1 \end{pmatrix}.$$

The primary goal of this experiment is to investigate how different choices of c affect the performance of the lasso and marginal regression. To see the point, note that in the idealized

situation where $X'X = C$, the model reduces to the one discussed in Figure 3, if we restrict our attention to any block of β where $\delta_i = 1$. The theoretic results in Example 4 of Section 2.3 predict that, the performance of the lasso gets increasingly unsatisfactory as c increases, while that of marginal regression stay more or less the same. At the same time, which of this method performs better depends on (a_2, a_3, c), see Figure 3 for details.

We select two different k for experiment: $k = 2$ and $k = 7$. When $k = 2$, we let $\delta_i = 1$ if and only if $i \in \{60, 139\}$, and when $k = 7$, we let $\delta_i = 1$ if and only if $i \in \{34, 44, 58, 91, 100, 183, 229\}$. Also, we investigate five different choices of (a_2, a_3): $(0, 0)$, $(0, 0), \mp(0.4, 0.1)$, and $\mp(0.5, 0.4)$, and three different c: $c = 0.5, 0.7$, and 0.85. For each combination of these parameters, we apply both the lasso and marginal regression and obtain the Hamming errors of both methods, where similarly, the tuning parameters for each method are set ideally. The error rates over 10 repetitions are tabulated in Table 2. The results suggest that different choices of c have a major role over the lasso, but does not have a big influence over marginal regression. The results fit well with the theory illustrated in Section 2.3; see Figure 3 for comparisons.

Experiment 2. In this experiment, we use the linear regression model $Y = X\beta + z$ where $z \sim N(0, I_n)$. We use a different criterion rather than the Hamming errors to compare two methods: with the same parameter settings, the method that yields exact recovery in a larger range of parameters is better. Towards this end, we take $p = n = 500$, and $X = \Omega^{1/2}$, where Ω is the $p \times p$ tridiagonal matrix satisfying

$$\Omega(i, j) = 1\{i = j\} + a \cdot 1\{|i - j| = 1\},$$

and the parameter $a \in (-1/2, 1/2)$ so the matrix is positive definite. At the same time, we generate β as follows. Let ϑ range between 0.25 and 0.75 with an increment of 0.25. For each ϑ, let s be the smallest even number $\geq p^{1-\vartheta}$. We then randomly pick $s/2$ indices $i_1 < i_2 < \ldots < i_{s/2}$. For parameters $r > 0$ and $d \in (-1, 1)$ to be determined, we let $\tau = \sqrt{2r \log p}$ and let $\beta_j = \tau$ if $j \in \{i_1, i_2, \ldots, i_{s/2}\}$, $\beta_j = d\tau$ if $j - 1 \in \{i_1, i_2, \ldots, i_{s/2}\}$, and $\beta_j = 0$ otherwise.

To gain insight on how two procedures perform in this setting, we consider the noiseless counterpart for just a second. Without loss of generality, we assume that the minimum inter-distance of indice $i_1, i_2, \ldots, i_k \geq 4$. Let $\widetilde{Y} = X'Y$. For any i_k, $1 \leq k \leq s/2$, if we restrict $\widetilde{Y}$ to $\{i_k - 1, i_k, i_k + 1, i_k + 2\}$ and call the resulting vector y, then

$$y = A\alpha,$$

where A is the 4 matrix satisfying $A(i, j) = 1\{i = j\} + a \cdot 1\{|i - j| = 1\}$, $1 \leq i, j \leq 4$, and α is the 4×1 vector such that $\alpha_1 = \alpha_4 = 0$, $\alpha_2 = \tau$, and $\alpha_3 = d\tau$. In this simple model, the performance of the lasso and marginal regression can be similarly analyzed as in Section 2.3.

Now, for each of the combination of (d, ϑ), we use the method of exhausting search to determine the smallest r such that the lasso or marginal regression yields exact recovery with 50 repetition of simulations, respectively (similarly, the tuning parameters of each method are set ideally). For each method, we call the resultant value of r the *critical value for exact recovery*. For each ϑ and choices of (a, d), we find the critical values for both methods. The results are summarized in Figure 5. For a given triplet (ϑ, a, d), the method

that gives a larger critical value is inferior to the one with a smaller critical value (as the region of parameters where it yields exact recovery is smaller). Figure 5 suggests that the parameters (a, d) play an important role in determining the performance of the lasso and marginal regression. For example, the performance of both procedures worsen when a get larger. This is because that as a increases, the Gram matrix moves away from that the identity matrix, and the problem of variable selection becomes increasingly harder. Also, the sign of $a \cdot d$ plays an interesting role. For example, when $a \cdot d < 0$, it is known that the marginal regression faces a so-called challenge of signal cancellation (see for example [30]). It seems that the lasso handles signal cancellation better than does marginal regression. However, when (a, d) range, there is no clear winner between two methods.

Experiment 3. So far, we have focused on settings where the regression problem can be decomposed into many parallel small-size regression problems. While how to decompose remains unknown, such insight is valuable, as we can always compare the performance of two methods over each of these small-size regression problems using the theory developed in Section 2.3; the overall performance of each method is then the sum of that on these small-size problems.

With that being said, in this experiment, we investigate an example where such a "decomposition" does not exist or at least is non-obvious. Consider an experiment where $Y = X\beta + z$, and $z \sim N(0, I_n)$. We take $p = n$ and $X = \Omega^{1/2}$, where Ω is a correlation matrix having the form

$$\Omega = \Lambda + a\xi\xi'.$$

which is a rank one perturbation of the diagonal matrix Λ. Here, ξ is the $p \times 1$ vector where its $p/2$ even coordinates are 1, and the remaining coordinates are b, where $a > 0$ and b are parameters calibrating the norm and direction of the rank one perturbation, respectively. Experiment 3 contains two sub-experiments, 3a and 3b.

In Experiment 3a, we investigate how the choices of parameters (a, b) and the signal strength affect the performance of the lasso and marginal regression. Let $p = n = 3000$, and let $\beta_i = \tau$ when $i \in \{k : k = 8 \times (\ell - 1) + 1, 1 \le \ell \le 150\}$ and $\beta_i = 0$ otherwise, where τ calibrates the signal strength. For each of the four choices $(a, b) = (0.01, 0.3), (0.01, 0.5), (0, 0.5), (0.5, -0.1)$, we compare the lasso and marginal regression for $\tau = 2, 3, \ldots, 8$. The Hamming errors are shown in Figure 6. The results suggest that the parameters (a, b) play a key role in the performance of both the lasso and marginal regression. For example, when a increases, the performance of both methods worsen, due to that the Gram matrix moves away from the identity matrix. Also, for relatively small a, it seems that marginal regression outperforms the lasso (see Panel 1 and 2 of Figure 6).

In Experiment 3b, we take a different angle and investigate how the levels of the signal sparsity affect the performance of the lasso and marginal regression. Consider a special case where where $b = 1$. In this case, ξ reduces to the vector of ones, and the Gram matrix is an equi-correlation matrix. This setting can be found in many literature on variable selection. Take $n = p = 500$. We generate the coordinates of β from the mixing distribution of point mass at 0 and the point mass at τ:

$$\beta_i \overset{iid}{\sim} (1 - \epsilon)\nu_0 + \epsilon\nu_\tau,$$

where ϵ calibrates the sparsity level and τ calibrates the signal strength (in this experiment, we take $\tau = 5$). In Figure 7, we plot the Hamming errors of 10 repetition versus the number

of variables retained (which can be thought of different choices of tuning parameters). Interestingly, it seems that the performance of two methods are strikingly similar, with relatively small differences (one way or the other) when the parameters (a, ϵ) are moderate (neither too close to 0 nor to 1). This is interesting as when ρ is moderate, the design matrix X is significantly non-orthogonal. Additionally, the results suggest that the sparsity parameter ϵ has a major influence over the relative performance of two methods. When ϵ get larger (so the signals get denser), marginal regression tends to outperform the lasso. The underlying reason is that when both the correlation and signals are positive, the strength of individual signals are amplified due to correlation, and so have a positive effect on marginal regression.

At the same time, it seems that the correlation parameter a also have a major effect over the performance of two methods, and the error rate of both methods increase as a increases. However, somewhat surprisingly, the parameter a does not seem to have a major effect on the relative performance of two methods.

We conclude this section by mentioning that from time to time, one would like to know for the data at hand, which method is preferable. Generally, this is a hard problem, and generally, there is no clear winner between the lasso and marginal regression. However, there are something can be learned from these simulation examples.

First, the study in this section suggest an interesting perspective, which can be explained as follows. Suppose that the Gram matrix is sparse in the sense that each row has relatively few large coordinates, and that the signal is also sparse. It turns out two types of sparsity interact with each other, and the large-scale regression problem reduces to many small-size regression models, each of which is obtained by restricting the rows of $X'Y$ to a small set of indices. In general, each of such of small-size regression models can be discussed in a similar fashion as those in Section 2.3. The results of these small-size regression problems then decide which of these two methods outperform the other. Take Experiment 1 for example. The performance of each method is determined by that of applying the method block-wise to the regression problem. This echos our previous argument in Section 2.3, where the relative performance of two methods for small-size problems are discussed in detail. Second, it seems that the lasso is comparably more vulnerable to extreme correlation, as discussed in Section 2.3 as well as in Example 1b. Last, it seems that in at least some examples, marginal regression is more vulnerable to the so-called "signal cancellation", which is illustrated in Proposition 3 as well as Example 2 in this section.

6. Proofs

6.1 Proof of Theorem 3

First, let k_i denote the number of non-zero diagonal entries in row i of D. Because D is symmetric but not diagonal, at least two rows must have non-zero k_i. Assume without loss of generality that the rows and columns of D are arranged so that the rows with non-zero k_i form the initial minor. It follows that the initial minor is itself a positive definite symmetric matrix. And because any such matrix A satisfies $|A_{ij}| < \max_k D_{kk}$ for $j \neq i$, there exists a row i of D with $k_i > 0$ and $|D_{ij}| < D_{ii}$ for any $j \neq i$.

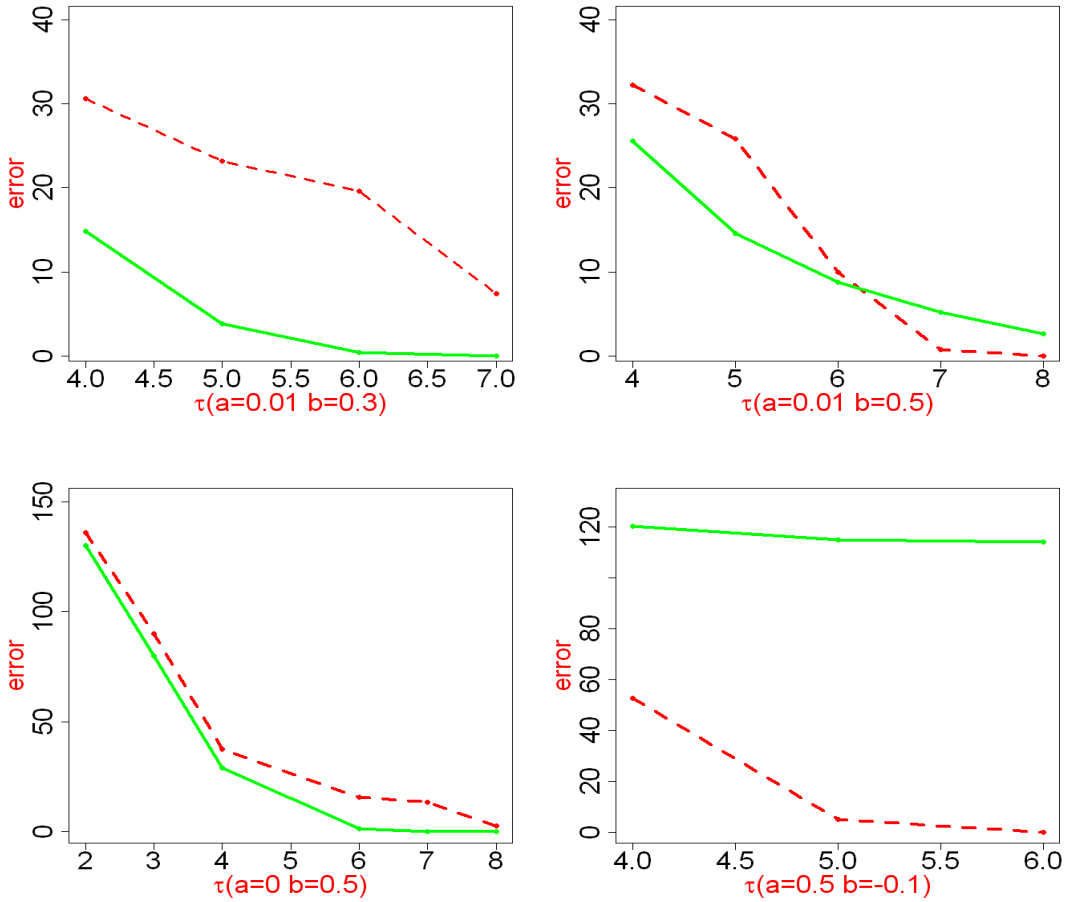

Figure 6: Comparison of Hamming errors by the lasso (red dashed) and marginal regression (green solid). The setting is described in Experiment 3a.

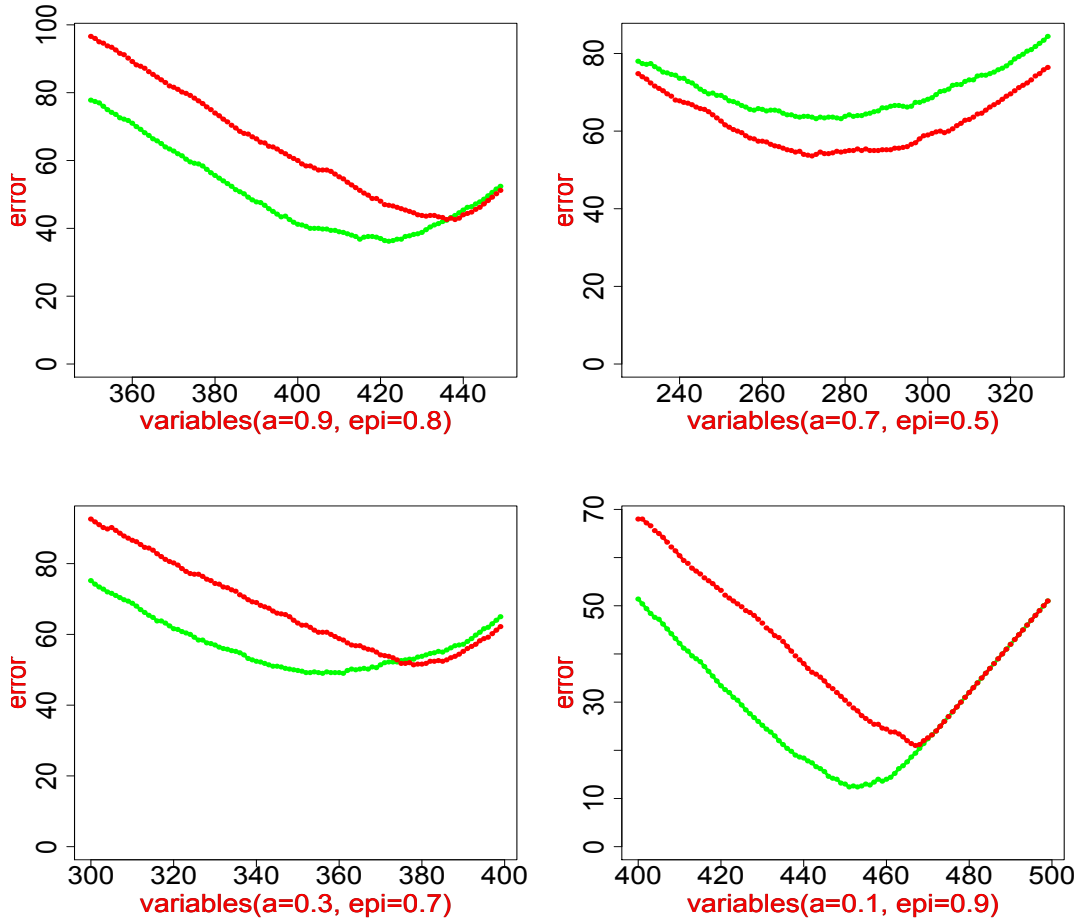

Figure 7: Comparison of Hamming errors by the lasso (red dashed) and marginal regression (green solid). The x-axis shows the number of retained variables. The setting is described in Experiment 3b.

Define β as follows:

$$\beta_j = \begin{cases} \frac{\rho D_{ii}}{D_{ij}} & \text{if } j \neq i \text{ and } D_{ij} \neq 0 \\ \rho & \text{if } j \neq i \text{ and } D_{ij} = 0 \\ -k_i \rho & \text{if } j = i.. \end{cases} \tag{34}$$

Because $|D_{ij}| \leq D_{ii}$, this satisfies $|\beta_j| \geq \rho$, so $\beta \in \mathcal{M}_\rho^s$. Moreover,

$$(D\beta)_i = \sum_j D_{ij}\beta_j = -k_i D_{ii}\rho + \sum_{\substack{j \neq i \\ D_{ij} \neq 0}} \frac{\rho D_{ii}}{D_{ij}} D_{ij} = 0. \tag{35}$$

This proves the theorem.

6.2 Proof of Lemma 4

By the definition of $\widehat{S}_n(s)$, it is sufficient to show that except for a probability that tends to 0,

$$\max |X_N^T Y| < \min |X_S^T Y|.$$

Since $Y = X\beta + z = X_S\beta_S + z$, we have $X_N^T Y = X_N^T(X_S\beta_S + z) = C_{NS}\beta_S + X_N^T z$. Note that $x_i^T z \sim N(0, \sigma_n^2)$. By Boolean algebra and elementary statistics,

$$P(\max |X_N^T z| > \sigma_n \sqrt{2\log p}) \leq \sum_{i \in N} P(|(x_i, z)| \geq \sigma_n \sqrt{2\log p}) \leq \frac{C}{\sqrt{\log p}} \frac{p-s}{p}.$$

It follows that except for a probability of $o(1)$,

$$\max |X_N^T Y| \leq \max |C_{NS}\beta_S| + \max |X_N^T z| \leq \max |C_{NS}\beta_S| + \sigma_n \sqrt{2\log p}.$$

Similarly, except for a probability of $o(1)$,

$$\min |X_S^T Y| \geq \min |C_{SS}\beta_S| - \max |X_S^T z| \geq \min |C_{SS}\beta_S| - \sigma_n \sqrt{2\log p}.$$

Combining these gives the claim. □

6.3 Proof of Theorem 5

Once the first claim is proved, the second claim follows from Lemma 4. So we only show the first claim. Write for short $\widehat{S}_n(s) = \widehat{S}_n(s^{(n)}; X^{(n)}, Y^{(n)}, p^{(n)})$, $s = s^{(n)}$, and $S = S(\beta^{(n)})$. All we need to show is

$$\lim_{n \to \infty} P(\widehat{s}_n \neq s) = 0.$$

Introduce the event

$$D_n = \{\widehat{S}_n(s) = S\}.$$

It follows from Lemma 4 that

$$P(D_n^c) \to 0.$$

Write

$$P(\widehat{s}_n \neq s) \leq P(D_n) P(\widehat{s}_n \neq s | D_n) + P(D_n^c).$$

It is sufficient to show $\lim_{n\to\infty} P(\widehat{s}_n \neq s|D_n) = 0$, or equivalently,

$$\lim_{n\to\infty} P(\widehat{s}_n > s|D_n) = 0 \qquad \text{and} \qquad \lim_{n\to\infty} P(\widehat{s}_n < s|D_n) = 0. \tag{36}$$

Consider the first claim of (36). Write for short $t_n = \sigma_n\sqrt{2\log n}$. Note that the event $\{\widehat{s}_n > s|D_n\}$ is contained in the event of $\cup_{k=s}^{p-1}\{\widehat{\delta}_n(k) \geq t_n|D_n\}$. Recalling $P(D_n^c) = o(1)$,

$$P(\widehat{s}_n > s) \leq \sum_{k=s}^{p-1}(\widehat{\delta}(k) \geq t_n|D_n) \lesssim \sum_{k=s}^{p-1} P(\widehat{\delta}_n(k) \geq t_n), \tag{37}$$

where we say two positive sequences $a_n \lesssim b_n$ if $\overline{\lim}_{n\to\infty}(a_n/b_n) \leq 1$.

Fix $s \leq k \leq p-1$. By definitions, $\widehat{H}(k+1) - \widehat{H}(k)$ is the projection matrix from R^n to $\widehat{V}_n(k+1) \cap \widehat{V}_n(k)^\perp$. So conditional on the event $\{\widehat{V}_n(k+1) = \widehat{V}_n(k)\}$, $\delta_n(k) = 0$, and conditional on the event $\{\widehat{V}_n(k+1) \subsetneq \widehat{V}_n(k)\}$, $\delta_n^2(k) \sim \sigma_n^2\chi^2(1)$. Note that $P(\chi^2(1) \geq 2\log n) = o(1/n)$. It follows that

$$\begin{aligned}\sum_{k=s}^{p-1} P(\widehat{\delta}_n(k) \geq t_n) &= \sum_{k=s}^{p-1} P(\widehat{\delta}_n(k) \geq t_n|\widehat{V}_n(k) \subsetneq \widehat{V}_n(k+1))P(\widehat{V}_n(k) \subsetneq \widehat{V}_n(k+1)) \\ &= o(\frac{1}{n})\sum_{k=s}^{p-1} P(\widehat{V}_n(k) \subsetneq \widehat{V}_n(k+1)).\end{aligned} \tag{38}$$

Moreover,

$$\begin{aligned}\sum_{k=s}^{p-1} P(\widehat{V}_n(k) \subsetneq \widehat{V}_n(k+1)) &= \sum_{k=s}^{p-1} E[1(\dim(\widehat{V}_n(k+1)) > \dim(\widehat{V}_n(k)))] \\ &= E[\sum_{k=s}^{p-1} 1(\dim(\widehat{V}_n(k+1)) > \dim(\widehat{V}_n(k)))].\end{aligned}$$

Note that for any realization of the sequences $\widehat{V}_n(1), \ldots, \widehat{V}_n(p)$, $\sum_{k=s}^{p-1} 1(\dim(\widehat{V}_n(k+1)) > \dim(\widehat{V}_n(k))) \leq n$. It follows that

$$\sum_{k=s}^{p-1} P(\widehat{V}_n(k) \subsetneq \widehat{V}_n(k+1)) \leq n. \tag{39}$$

Combining (37)-(39) gives the claim.

Consider the second claim of (36). By the definition of $\widehat{s}_n$, the event $\{\widehat{s}_n < s|D_n)\}$ is contained in the event $\{\widehat{\delta}_n(s-1) < t_n|D_n\}$. By definitions, $\widehat{\delta}_n(s-1) = \|(\widehat{H}(s) - \widehat{H}(s-1))Y\|$, where $\|\cdot\| = \|\cdot\|_2$ denotes the ℓ^2 norm. So all we need to show is

$$\lim_{n\to\infty} P(\|(\widehat{H}(s) - \widehat{H}(s-1))Y\| < t_n|D_n) = 0. \tag{40}$$

Fix $1 \leq k \leq p$. Recall that i_k denotes the index at which the rank of $|(Y, x_{i_k})|$ among all $|(Y, x_j)|$ is k. Denote $\widetilde{X}(k)$ by the n by k matrix $[x_{i_1}, x_{i_2}, \ldots, x_{i_k}]$, and denote $\widetilde{\beta}(k)$ by the

k-vector $(\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_k})^T$. Conditional on the event D_n, $\widehat{S}_n(s) = S$, and $\beta_{i_1}, \beta_{i_2}, \ldots, \beta_{i_s}$ are all the nonzero coordinates of β. So according to our notations,

$$X\beta = \widetilde{X}(s)\widetilde{\beta}(s) = \widetilde{X}(s-1)\widetilde{\beta}(s-1) + \beta_{i_s}x_{i_s} \tag{41}$$

Now, first, note that $\widehat{H}(s)\widetilde{X}(s) = \widetilde{X}(s)$ and $\widehat{H}(s-1)\widetilde{X}(s-1) = \widetilde{X}(s-1)$. Combine this with (41). It follows from direct calculations that

$$(\widehat{H}(s) - \widehat{H}(s-1))X\beta = (I - \widehat{H}(s-1))x_{i_s}. \tag{42}$$

Second, since $x_{i_s} \in \widehat{V}_n(s)$, $(I - \widehat{H}(s))x_{i_s} = 0$. So

$$(I - \widehat{H}_{s-1})x_{i_s} = (I - \widehat{H}(s))x_{i_s} + (\widehat{H}(s) - \widehat{H}(s-1))x_{i_s} = (\widehat{H}_s - \widehat{H}_{s-1})x_{i_s}. \tag{43}$$

Last, split x_{i_s} into two terms, $x_{i_s} = x_{i_s}^{(1)} + x_{i_s}^{(2)}$ such that $x_{i_s}^{(1)} \in \widehat{V}_n(s-1)$ and $x_{i_s}^{(2)} \in \widehat{V}_n(s) \cap (\widehat{V}_n(s-1))^{\perp}$. It follows that $(\widehat{H}(s) - \widehat{H}(s-1))x_{i_s}^{(1)} = 0$, and so

$$(\widehat{H}(s) - \widehat{H}(s-1))x_{i_s} = (\widehat{H}(s) - \widehat{H}(s-1))x_{i_s}^{(2)}. \tag{44}$$

Combining (42)-(44) gives

$$(\widehat{H}(s) - \widehat{H}(s-1))X\beta = (\widehat{H}(s) - \widehat{H}(s-1))x_{i_s}^{(2)}. \tag{45}$$

Recall that $Y = X\beta + z$, it follows that

$$(\widehat{H}_s - \widehat{H}_{s-1})Y = (\widehat{H}(s) - \widehat{H}(s-1))(\beta_{i_s}x_{i_s}^{(2)} + z). \tag{46}$$

Now, take an orthonormal basis of R^n, say $\widehat{q}_1, \widehat{q}_2, \ldots, \widehat{q}_n$, such that $\widehat{q}_1 \in \widehat{V}_n(s) \cap \widehat{V}_n(s-1)^{\perp}$, $\widehat{q}_2, \ldots, \widehat{q}_s \in \widehat{V}_n(s-1)$, and $\widehat{q}_{s+1}, \ldots, \widehat{q}_n \in \widehat{V}_n(s)^{\perp}$. Recall that $x_{i_s}^{(2)}$ is contained in the one dimensional linear space $\widehat{V}_n(s) \cap \widehat{V}_n(s-1)^{\perp}$, so without loss of generality, assume $(x_{i_s}^{(2)}, \widehat{q}_1) = \|x_{i_s}^{(2)}\|$. Denote the square matrix $[\widehat{q}_1, \ldots, \widehat{q}_n]$ by $\widehat{Q}$. Let $\widetilde{z} = \widehat{Q}z$ and let $\widetilde{z}_1$ be the first coordinate of $\widetilde{z}$. Note that marginally $\widetilde{z}_1 \sim N(0, \sigma_n^2)$. Over the event D_n, it follows from the construction of $\widehat{Q}$ and basic algebra that

$$\|(\widehat{H}(s) - \widehat{H}(s-1))(\beta_{i_s}x_{i_s}^{(2)} + z)\|^2 = (\|\beta_{i_s}x_{i_s}^{(2)}\| + \widetilde{z}_1)^2. \tag{47}$$

Combine (46) and (47),

$$\|(\widehat{H}(s) - \widehat{H}(s-1))Y\|^2 = (\|\beta_{i_s}x_{i_s}^{(2)}\| + \widetilde{z}_1)^2, \qquad \text{over the event } D_n.$$

As a result,

$$P(\|(\widehat{H}(s) - \widehat{H}(s-1))Y\| < t_n | D_n) = P((\|\beta_{i_s}x_{i_s}^{(2)}\| + \widetilde{z}_1)^2 < t_n | D_n). \tag{48}$$

Recall that conditional on the event D_n, $\widehat{S}_n(s) = S$. So by the definition of $\Delta_n^* = \Delta_n(\beta, X, p)$,

$$\|\beta_{i_s}x_{i_s}^{(2)}\| \geq \Delta_n^*,$$

and

$$P((\|\beta_{i_s}x_{i_s}^{(2)}\| + \widetilde{z}_1)^2 < t_n|D_n) \le P(\|\beta_{i_s}x_{i_s}^{(2)}\| + \widetilde{z}_1 < t_n|D_n) \le P(\Delta_n^* + \widetilde{z}_1 < t_n|D_n). \quad (49)$$

Recalling that $\widetilde{z}_1 \sim N(0, \sigma_n^2)$ and that $P(D_n^c) = o(1)$,

$$P(\Delta_n^* + \widetilde{z}_1 < t_n|D_n) \le P(\Delta_n^* + \widetilde{z}_1 < t_n) + o(1). \quad (50)$$

Note that by the assumption of $(\frac{\Delta_n^*}{\sigma_n} - t_n) \to \infty$, $P(\Delta_n^* + \widetilde{z}_1 < t_n) = o(1)$. Combining this with (49)-(50) gives

$$P((\|\beta_{i_s}x_{i_s}^{(2)}\| + \widetilde{z}_1)^2 < t_n^2|D_n) = o(1). \quad (51)$$

Inserting (51) into (48) gives (40). □

6.4 Proof of Lemma 6

For $1 \le i \le p$, introduce the random variable

$$Z_i = \sum_{j \ne i}^{p} \beta_j(x_i, x_j).$$

When $B_i = 0$, $\beta_i = 0$, and so $Z_i = \sum_{j=1}^{p} \beta_j(x_i, x_j)$. By the definition of C_{NS},

$$\max|C_{NS}\beta_S| = \max_{1\le i\le p}\{(1 - B_i) \cdot |\sum_{j=1}^{p} \beta_j(x_i, x_j)|\} = \max_{1\le i\le p}\{(1 - B_i)|Z_i|\}.$$

Also, recalling that the columns of matrix X are normalized such that $(x_i, x_i) = 1$, the diagonal coordinates of $(C_{SS} - I)$ are 0. Therefore,

$$\max|(C_{SS} - I)\beta_S| = \max_{1\le i\le p}\{B_i \cdot |\sum_{j\ne i} \beta_j(x_i, x_j)|\} = \max_{1\le i\le p}\{B_i \cdot |Z_i|\}.$$

Note that Z_i and B_i are independent and that $P(B_i = 0) = (1 - \epsilon)$. It follows that

$$P(\max|C_{NS}\beta_S| \ge \delta) \le \sum_{i=1}^{p} P(B_i = 0)P(|Z_i| \ge \delta|B_i = 0) = (1 - \epsilon)\sum_{i=1}^{p} P(|Z_i| \ge \delta),$$

and

$$P(\max|(C_{SS} - I)\beta_S| \ge \delta) \le \sum_{i=1}^{p} P(B_i = 1)P(|Z_i| \ge \delta|B_i = 1) = \epsilon\sum_{i=1}^{p} P(|Z_i| \ge \delta).$$

Compare these with the lemma. It is sufficient to show

$$P(|Z_i| \ge \delta) \le e^{-\delta t}[e^{\epsilon\bar{g}_i(t)} + e^{\epsilon\bar{g}_i(-t)}]. \quad (52)$$

Now, by the definition of $g_{ij}(t)$, the moment generating function of Z_i satisfies that

$$E[e^{tZ_i}] = E[e^{t\sum_{j\ne i}\beta_j(x_i,x_j)}] = \Pi_{j\ne i}[1 + \epsilon g_{ij}(t)]. \quad (53)$$

Since $1+x \le e^x$ for all x, $1+\epsilon g_{ij}(t) \le e^{\epsilon g_{ij}(t)}$, so by the definition of $\bar{g}_i(t)$,

$$E[e^{tZ_i}] \le \Pi_{j\neq i} e^{\epsilon g_{ij}(t)} = e^{\epsilon \bar{g}_i(t)}. \tag{54}$$

It follows from Chebyshev's inequality that

$$P(Z_i \ge \delta) \le e^{-\delta t} E[e^{tZ_i}] \le e^{-\delta t} e^{\epsilon \bar{g}_i(t)}. \tag{55}$$

Similarly,

$$P(Z_i < -\delta) \le e^{-\delta t} e^{\epsilon \bar{g}_i(-t)} \tag{56}$$

Inserting (55)-(56) into (52) gives the claim. □

6.5 Proof of Corollary 3.1

Choose a constant q such that $q/2 - c_2 q > 1$ and let $t_n = q\log(p)/a_n$. By the definition of $A_n(a_n/2, \epsilon_n, \bar{g})$, it is sufficient to show that for all $1 \le i \le p$,

$$e^{-a_n t_n/2} e^{\epsilon_n \bar{g}_i(t_n)} = o(1/p), \qquad e^{-a_n t_n/2} e^{\epsilon_n \bar{g}_i(-t_n)} = o(1/p).$$

The proofs are similar, so we only show the first one. Let u be a random variable such that $u \sim \pi_n$. Recall that the support of $|u|$ is contained in $[a_n, b_n]$. By the assumptions and the choice of t_n, for all fixed i and $j \neq i$, $|t_n u(x_i, x_j)| \le q\log(p)(b_n/a_n)|(x_i, x_j)| \le c_1 q$. Since $e^x - 1 \le x + e^x x^2/2$, it follows from Taylor expansion that

$$\epsilon_n \bar{g}_i(t_n) = \epsilon_n [e^{t_n u(x_i,x_j)} - 1] \le \epsilon_n \sum_{j\neq i} E_{\pi_n}[t_n u(x_i, x_j) + \frac{e^{c_1 q}}{2} t_n^2 u^2 (x_i, x_j)^2].$$

By definitions of $m_n(X)$ and $v_n^2(X)$, $\epsilon_n \sum_{j\neq i} E_\pi[t_n u(x_i, x_j)] = t_n \mu_n^{(1)} m_n(X)$, and $\epsilon_n \sum_{j\neq i} E_{\pi_n}[t_n^2 u^2(x_i, x_j)^2] = t_n^2 \mu_n^{(2)} v_n^2(X)$. It follows from (22) that

$$\epsilon_n \bar{g}_i(t_n) \le q\log(p) \cdot [\frac{\mu_n^{(1)}}{a_n} m_n(X) + \frac{e^{c_1 q}}{2} \frac{\mu_n^{(2)}}{a_n^2} v_n^2(X) q\log(p)] \lesssim q c_2 \log(p).$$

Therefore,

$$e^{-a_n t_n/2} e^{\epsilon_n \bar{g}_i(t_n)} \le e^{-[q/2 - c_2 q + o(1)]\log(p)},$$

and claim follows by the choice of q. □

6.6 Proof of Corollary 3.2

Choose a constant q such that $2 < q < \frac{c_3}{c_4 \delta}$. Let $t_n = a_n q \log(p)$, and u be a random variable such that $u \sim \Pi_n$. Similar to the proof of Lemma 3.1, we only show that

$$e^{-a_n t_n/2} e^{\epsilon_n \bar{g}_i(t_n)} = o(1/p), \qquad \text{for all } 1 \le i \le p.$$

Fix $i \neq j$. When $(x_i, x_j) = 0$, $e^{tu(x_i,x_j)} - 1 = 0$. When $(x_i, x_j) \neq 0$, $e^{t_n u(x_i,x_j)} - 1 \le e^{t_n (b_n/a_n)\delta} \le e^{c_4 q \delta \log p}$. Also, $\epsilon_n N_n^* \le e^{-[c_3 + o(1)]\log(p)}$. Therefore,

$$\epsilon_n \bar{g}_i(t) \le \epsilon_n N_n^* e^{c_4 q \delta \log(p)} \le e^{-[c_3 - c_4 q \delta + o(1)]\log p}.$$

By the choice of q, $c_3 - c_4 q\delta > 0$, so $\epsilon_n \bar{g}_i(t) = o(1)$. It follows that

$$e^{-a_n t_n/2} e^{\epsilon_n \bar{g}_i(t_n)} \le o(e^{-a_n t_n/2}) = o(e^{-q\log(p)/2}),$$

which gives the claim by $q > 2$. □

6.7 Proof of Theorem 9

Write

$$X = [x_1, \widetilde{X}], \qquad \beta = (\beta_1, \widetilde{\beta})^T.$$

Fix a constant $c_0 > 3$. Introduce the event

$$D_n(c_0) = \{1_S^T \widetilde{X}_S^T \widetilde{X}_S 1_S \leq |S|[1 + \sqrt{\frac{|S|}{n}}(1 + \sqrt{2c_0 \log p})]^2, \text{for all } S\}. \tag{57}$$

The following lemma is proved in Section 6.7.1.

Lemma 12 *Fix $c_0 > 3$. As $p \to \infty$,*

$$P(D_n^c(c_0)) = o(1/p^2).$$

Since $d_n(\widehat{\beta}|X) \leq p$ for any variable selection procedure $\widehat{\beta}$, Lemma 12 implies that the overall contribution of D_n^c to the Hamming distance $d_n^*(\widehat{\beta})$ is $o(1/p)$. In addition, write

$$d_n(\widehat{\beta}|X) = \sum_{j=1}^{p} E[1(\widehat{\beta}_j \neq \beta_j)].$$

By symmetry, it is sufficient to show that for any realization of $(X, \beta) \in D_n(c_0)$,

$$E[1(\widehat{\beta}_j \neq \beta_j)] \geq \begin{cases} L(n)p^{-\frac{(\vartheta+r)^2}{4r}}, & r \geq \vartheta, \\ p^{-\vartheta}, & 0 < r < \vartheta, \end{cases} \tag{58}$$

where $L(n)$ is a multi-log term that does not depend on (X, β).

We now show (58). Toward this end, we relate the estimation problem to the problem of testing the null hypothesis of $\beta_1 = 0$ versus the alternative hypothesis of $\beta_1 \neq 0$. Denote ϕ by the density of $N(0, 1)$. Recall that $X = [x_1, \widetilde{X}]$ and $\beta = (\beta_1, \widetilde{\beta})^T$. The joint density associated with the null hypothesis is

$$f_0(y) = f_0(y; \epsilon_n, \tau_n, n|X)\phi(y - \widetilde{X}\widetilde{\beta})d\widetilde{\beta} = \phi(y) \int e^{y^T \widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2} d\widetilde{\beta},$$

and the joint density associated with the alternative hypothesis is

$$\begin{aligned} f_1(y) = f_1(y; \epsilon_n, \tau_n, n|X) &= \int \phi(y - \tau_n x_1 - \widetilde{X}\widetilde{\beta})d\widetilde{\beta} \\ &= \phi(y - \tau_n x_1) \int e^{y^T \widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2} e^{-\tau_n x_1^T \widetilde{X}\widetilde{\beta}} d\widetilde{\beta}. \end{aligned} \tag{59}$$

Since the prior probability that the null hypothesis is true is $(1 - \epsilon_n)$, the optimal test is the Neyman-Pearson test that rejects the null if and only if

$$\frac{f_1(y)}{f_0(y)} \geq \frac{(1 - \epsilon_n)}{\epsilon_n}.$$

The optimal testing error is equal to

$$1 - \|(1 - \epsilon_n)f_0 - \epsilon_n f_1\|_1.$$

Compared to (2), $\|\cdot\|_1$ stands for the L^1-distance between two functions, not the ℓ^1 norm of a vector.

We need to modify f_1 into a more tractable form, but with negligible difference in L^1-distance. Toward this end, let $N_n(\widetilde{\beta})$ be the number of nonzeros coordinates of $\widetilde{\beta}$. Introduce the event

$$B_n = \{|N_n(\widetilde{\beta}) - p\epsilon_n| \le \frac{1}{2}p\epsilon_n\}.$$

Let

$$a_n(y) = a_n(y; \epsilon_n, \tau_n|X) = \frac{\int (e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2})(e^{-\tau_n x_1^T\widetilde{X}\widetilde{\beta}}) \cdot 1_{\{B\}} d\widetilde{\beta}}{\int (e^{-y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2}) \cdot 1_{\{B\}} d\widetilde{\beta}}. \tag{60}$$

Note that the only difference between the numerator and the denominator is the term $e^{-\tau_n x_1^T\widetilde{X}\widetilde{\beta}}$ which ≈ 1 with high probability. Introduce

$$\widetilde{f}_1(y) = a_n(y)\phi(y - \tau_n x_1) \int e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2} d\widetilde{\beta}. \tag{61}$$

The following lemma is proved in Section 6.7.2.

Lemma 13 *As $p \to \infty$, there is a generic constant $c > 0$ that does not depend on y such that $|a_n(y) - 1| \le c\log(p)p^{(1-\vartheta)-\theta/2}$ and $\|f_1 - \widetilde{f}_1\|_1 = o(1/p)$.*

We now ready to show the claim. Define $\Omega_n = \{y : a_n(y)\phi(y - \tau_n x_1) \ge \phi(y)\}$. Note that by the definitions of $f_0(y)$ and $\widetilde{f}_1(y)$, $y \in \Omega_n$ if and only if

$$\frac{\epsilon_n\widetilde{f}_1(y)}{(1-\epsilon_n)f_0(y)} \ge 1.$$

By Lemma 13,

$$|\int \widetilde{f}_1(y)dy - 1| \le \|\widetilde{f}_1 - f_1\|_1 \le o(1/p).$$

It follows from elementary calculus that

$$1 - \|(1-\epsilon_n)f_0 - \epsilon_n\widetilde{f}_1\|_1 = \int_{\Omega_n}(1-\epsilon_n)f_0(y)dy + \int_{\Omega_n^c}\epsilon_n\widetilde{f}_1(y)dy + o(1/p).$$

Using Lemma 13 again, we can replace $\widetilde{f}_1$ by f_1 on the right hand side, so

$$1 - \|(1-\epsilon_n)f_0 - \epsilon_n\widetilde{f}_1\|_1 = \int_{\Omega_n}(1-\epsilon_n)f_0(y)dy + \int_{\Omega_n^c}\epsilon_n f_1(y)dy + o(1/p).$$

At the same time, let $\delta_p = c\log(p)p^{(1-\vartheta)-\theta/2}$ be as in Lemma 13, and let

$$t_0 = t_0(\vartheta, r) = \frac{\vartheta + r}{2\sqrt{r}}\sqrt{2\log p}.$$

be the unique solution of the equation $\phi(t) = \epsilon_n\phi(t - \tau_n)$. It follows from Lemma 13 that,

$$\{\tau_n x^T y \ge t_0(1+\delta_p)\} \subset \Omega_n \subset \{\tau_n x_1^T y \ge t_0(1-\delta_p)\}.$$

As a result,

$$\int_{\Omega_n} f_0(y)dy \geq \int_{\tau_n x_1^T y \geq t_0(1+\delta_p)} f_0(y) \equiv P_0(\tau_n x_1^T Y \geq t_0(1+\delta_p)),$$

and

$$\int_{\Omega_n^c} f_1(y)dy \geq \int_{\tau_n x_1^T y \leq t_0(1-\delta_p)} f_1(y) \equiv P_1(\tau_n x_1^T Y \leq t_0(1-\delta_p)).$$

Note that under the null, $x_1^T Y = x_1^T \widetilde{X}\widetilde{\beta} + x_1^T z$. It is seen that given x_1, $x_1^T z \sim N(0, |x_1|^2)$, and $|x_1|^2 = 1 + O(1/\sqrt{n})$. Also, it is seen that except for a probability of $o(1/p)$, $x_1^T \widetilde{X}\widetilde{\beta}$ is algebraically small. It follows that

$$P_0(\tau_n x_1^T Y \geq t_0(1+\delta_p)) \lesssim \bar{\Phi}(t_0) = L(n)p^{-\frac{(\vartheta+r)^2}{4r}},$$

where $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0,1)$. Similarly, under the alternative,

$$x_1^T y = \tau_n(x_1, x_1) + x_1^T \widetilde{X}\widetilde{\beta} + x_1^T z,$$

where $(x_1, x_1) = 1 + O(1/\sqrt{n})$. So

$$\epsilon_n P_1(\tau_n x_1^T y \leq t_0(1-\delta_p)) \lesssim \Phi(t_0 - \tau_n) = \begin{cases} L(n)p^{-\frac{(\vartheta+r)^2}{4r}}, & r \geq \vartheta, \\ L(n)p^{-\vartheta}, & 0 < r < \vartheta, \end{cases}$$

Combine these gives the theorem. □

6.7.1 Proof of Lemma 12

It is seen that

$$P(D_n^c(c_0)) \leq \sum_{k=1}^{p} P\bigg(1_S^T X^T X 1_S \geq k[1 + \sqrt{\frac{k}{n}}(1 + \sqrt{2c_0 \log p})]^2, \text{for all } S \text{ with } |S| = k\bigg).$$

Fix $k \geq 1$. There are $\binom{p}{k}$ different S with $|S| = k$. It follows from [27, Lecture 9] that except a probability of $2\exp(-c_0 \log(p) \cdot k)$ that the largest eigenvalue of $X_S^T X_S$ is no greater than $[1 + \sqrt{\frac{k}{n}}(1 + \sqrt{2c_0 \log p})]^2$. So for any S with $|S| = k$, it follows from basic algebra that

$$P(1_S^T X^T X 1_S \geq k[1 + \sqrt{\frac{k}{n}}(1 + \sqrt{2c_0 \log p})]^2) \leq 2\exp(-c_0 \log(p) \cdot k).$$

Combining these with $\binom{p}{k} \leq p^k$ gives

$$P(D_n^c(c_0)) \leq 2\sum_{k=1}^{p} \binom{p}{k} \exp(-c_0(\log p)k) \leq 2\sum_{k=1}^{p} \exp(-(c_0 - 1)\log(p)k).$$

The claim follows by $c_0 > 3$. □

6.7.2 Proof of Lemma 13

First, we claim that for any X in event $D_n(c_0)$,

$$|x_1^T \widetilde{X}\widetilde{\beta}| \leq c\log(p)(N(\widetilde{\beta})/\sqrt{n}), \tag{62}$$

where $c > 0$ is a generic constant. Suppose $N_n(\widetilde{\beta}) = k$ and the nonzero coordinates of $\widetilde{\beta}$ are $i_1, i_2, \ldots, i_k$. Denote the $(k+1)\times(k+1)$ submatrix of X^TX containing the 1^{st}, $(1+i_1)$-th, ..., and $(1+i_k)$-th rows and columns by U_{k+1}. Let ξ_1 be the $(k+1)$-vector with 1 on the first coordinate and 0 elsewhere, let ξ_2 be the $(k+1)$-vector with 0 on the first coordinate and 1 elsewhere. Then

$$x_1^T \widetilde{X}\widetilde{\beta} = \tau_n \xi_1^T U_{k+1}\xi_2 \equiv \tau_n \xi_1^T (U_{k+1} - I_{k+1})\xi_2.$$

Let $(U_{k+1} - I_{k+1}) = Q_{k+1}\Lambda_{k+1}Q_{k+1}^T$ be the orthogonal decomposition. By the definition of $D_n(c_0)$, all eigenvalues of $(U_{k+1} - I_{k+1})$ are no greater than $(1 + \sqrt{c\log(p)k/n})^2 - 1 \leq \sqrt{c\log p}\sqrt{k/n}$ in absolute value. As a result, all diagonal coordinates of Λ_{k+1} are no greater than

$$\sqrt{c\log p}\sqrt{k/n}$$

in absolute value, and

$$\|\xi_1^T(U_{k+1} - I_{k+1})\xi_2\| \leq \|\xi_1^T Q_{k+1}\Lambda_{k+1}\| \cdot \|Q_{k+1}\xi_2\| \leq \sqrt{c\log p}\sqrt{k/n}\|\xi_1^T Q_{k+1}\| \cdot \|Q_{k+1}\xi_2\|.$$

The claim follows from $\|\xi_1^T Q_{k+1}\| = 1$ and $\|Q_{k+1}\xi_2\| = \sqrt{k}$.

We now show the lemma. Consider the first claim. Consider a realization of X in the event $D_n(c_0)$ and a realization of $\widetilde{\beta}$ in the event B_n. By the definitions of B_n, $N_n(\widetilde{\beta}) \leq p\epsilon_n + \frac{1}{2}p\epsilon_n$. Recall that $p\epsilon_n = p^{1-\vartheta}$, $n = p^\theta$. It follows that $\log(p)N(\widetilde{\beta})/\sqrt{n} \leq c\log(p)p\epsilon_n/\sqrt{n} = c\log(p)p^{1-\vartheta-\theta/2}$. Note that by the assumption of $(1-\vartheta) < \theta/2$, the exponent is negative. Combine this with (62),

$$|e^{-\tau_n x_1^T \widetilde{X}\widetilde{\beta}} - 1| \leq c\log(p)(N(\widetilde{\beta})/\sqrt{n}), \tag{63}$$

Now, note that in the definition of $a_n(y)$ (i.e. (60)), the only difference between the integrand on the top and that on the bottom is the term $e^{-\tau_n x_1^T \widetilde{X}\widetilde{\beta}}$. Combine this with (63) gives the claim.

Consider the second claim. By the definitions of $\widetilde{f}_1(y)$ and $a_n(y)$,

$$\begin{aligned}\widetilde{f}_1(y) &= a_n(y)\phi(y-\tau_n x_1)\cdot\left[\int [e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2}1_{B_n}]d\widetilde{\beta} + \int [e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2}1_{B_n^c}]d\widetilde{\beta}\right] \\ &= \phi(y-\tau_n x_1)\cdot\left[\int [e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2}e^{-\tau_n x_1^T\widetilde{X}\widetilde{\beta}}1_{B_n^c}]d\widetilde{\beta} + a_n(y)\int [e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2}1_{B_n^c}]d\widetilde{\beta}\right].\end{aligned}$$

By the definition of $f_1(y)$,

$$f_1(y) = \phi(y-\tau_n x_1)\cdot\left[\int [e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2}e^{-\tau_n x_1^T\widetilde{X}\widetilde{\beta}}1_{B_n}]d\widetilde{\beta} + \int [e^{y^T\widetilde{X}\widetilde{\beta} - |\widetilde{X}\widetilde{\beta}|^2/2}e^{-\tau_n x_1^T\widetilde{X}\widetilde{\beta}}1_{B_n^c}]d\widetilde{\beta}\right].$$

Compare two equalities and recall that $a_n(y) \sim 1$ (Lemma 12),

$$\begin{aligned}\|f_1 - \widetilde{f}_1\|_1 &\lesssim \int \phi(y-\tau_n x_1)[\int (e^{y^T\widetilde{X}\widetilde{\beta}-|\widetilde{X}\widetilde{\beta}|^2/2} + e^{y^T\widetilde{X}\widetilde{\beta}-|\widetilde{X}\widetilde{\beta}|^2/2}e^{-\tau_n x_1^T\widetilde{X}\widetilde{\beta}})1_{B_n^c}d\widetilde{\beta}]dy \\ &= \int\int \phi(y-\tau_n x_1-\widetilde{X}\widetilde{\beta})[e^{\tau_n x_1^T\widetilde{X}\widetilde{\beta}}+1]1_{B_n^c}d\widetilde{\beta}dy. \qquad (64)\end{aligned}$$

Integrating over y, the last term is equal to $\int[1+e^{\tau_n x_1^T\widetilde{X}\widetilde{\beta}}]\cdot 1_{B_n^c}d\widetilde{\beta}$.

At the same time, by (62) and the definition of B_n^c,

$$\int[1+e^{\tau_n x_1^T\widetilde{X}\widetilde{\beta}}]\cdot 1_{B_n^c}d\widetilde{\beta} \le \sum_{\{k:|k-p\epsilon_n|\ge\frac{1}{2}p\epsilon_n\}}[1+e^{c\log(p)k/\sqrt{n}}]P(N(\widetilde{\beta})=k). \qquad (65)$$

Recall that $p\epsilon_n = p^{1-\vartheta}$, $n=p^\theta$, and $(1-\vartheta)<\theta/2$. Using Bennett's inequality for $P(N(\widetilde{\beta})=k)$ (e.g. [31, Page 440]), it follows from elementary calculus that

$$\sum_{\{k:|k-p\epsilon_n|\ge\frac{1}{2}p\epsilon_n\}}[1+e^{c\log(p)k/\sqrt{n}}]P(N(\widetilde{\beta})=k) = o(1/p). \qquad (66)$$

Combining (64)–(66) gives the claim. □

Acknowledgement: We would like to thank David Donoho, Robert Tibshirani, and anonymous referees for helpful discussions. CG was supported in part by NSF grant DMS-0806009 and NIH grant R01NS047493, JJ was supported in part by NSF CAREER award DMS-0908613, and LW was supported in part by NSF grant DMS-0806009.

References

[1] Bühlmann, P., Kalisch, M. and Maathuis, M. H. (2009). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika*, **97** 261–278.

[2] Cai, T., Wang, L. and Xu, G. (2010). Shifting inequality and recovery of sparse signals. *IEEE Transactions on Signal Processing*, **59**(3) 1300–1308.

[3] Candés, E. and Plan, Y. Near-ideal model selection by ℓ^1 minimization. *Ann. Statist.*, **37** 2145–2177.

[4] Candés, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35** 2313–2351.

[5] Chen. S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, **20**(1) 33–61.

[6] Donoho, D. (2006a). For most large underdetermined systems of linear equations the minimal ℓ^1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, **59**(7) 907–934.

[7] Donoho, D. (2006b). High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension. *Disc. Comput. Geometry*, **35**(4) 617–652.

[8] Donoho, D. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proc. Natl. Acad. Sci.*, **100**(5) 2197–2202.

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

[10] Donoho, D. and Jin, J. (2004). Higher Criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32** 962–994.

[11] Donoho, D. and Tanner, J. (2005). Neighborliness of randomly-projected simplices in high dimensions. *Proc. Natl. Acad. Sci.*, **102**(27) 9452–9457.

[12] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32**(2) 407–499.

[13] Efron, B., Tibshirani, R., Storey, J. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, **96** 1151–1160.

[14] Fan, J. and Li, R. (1999). Variable Selection via Penalized Likelihood. *J. Amer. Statist. Assoc.*, **96** 1348–1360.

[15] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Roy. Statist. Soc. B*, **70** 849–911.

[16] Fuchs, J.J. (2005). Recovery of exact sparse representations in the presence of noise. *IEEE Trans. Info. Theory*, **51**(10) 3601–3608.

[17] Jin, J. (2007). Proportion of nonzero normal means: oracle equivalence and uniformly consistent estimators. *J. R. Roy. Soc. B*, **70**(3) 461–493.

[18] Knight, K. and Fu, W.J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.*, **28** 1356–1378.

[19] Ji, P. and Jin, J. (2010). UPS delivers optimal phase diagram in high dimensional variable selection. *To appear in Ann. Statist.*

[20] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34**(3) 1436–1462.

[21] Meinshausen, M. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, **34** 373–393.

[22] Ravikumar, P. (2007). *Personal communication.*

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

[24] Spirtes, P., Glymour, C. and Scheines. R. (1993). *Causation, Prediction, and Search*. Springer-Verlag Lecture Notes in Statistics **81**, NY.

[25] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B*, **58**(1) 267–288.

[26] Tropp, J. (2004). Greed is good: algorithic results for sparse approximation. *IEEE Trans. Info. Theory*, **50**(10) 2231–2242.

[27] Vershynin, R. (2007). *Nonasymptotic theory of random matrices*. Lecture notes, Department of Mathematics, University of Michigan. *www-personal.umich.edu/~romanv/teaching/2006-07/280/course.html*.

[28] Wainwright, M. (2006). Sharp threshold for high-dimensional and noisy recovery of sparsity. *Technical report*, Department of Statistics, University of Berkeley.

[29] Wasserman, L. (2006). *All of nonparametric statistics*. Springer, NY.

[30] Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist*. **37**(5), 2178–2201.

[31] Shorack, G.R. and Wellner, J.A. (1986). *Empirical Process with Application to Statistics*. John Wiley & Sons, NY.

[32] Sun, T. and Zhang, C.-H. (2011). Scaled sparse linear regression. *arXiv:1104.4595*.

[33] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learning Research*, **7** 2541–2563.

[34] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**(476) 1418–1429.