The Lasso under Heteroscedasticity

Jinzhu Jia JZJIA@STAT.BERKELEY.EDU
Karl Rohe KARLROHE@STAT.BERKELEY.EDU
Department of Statistics
University of California
Berkeley, CA 94720, USA

Bin Yu BINYU@STAT.BERKELEY.EDU
Department of Statistics,
and Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720, USA

Editor:

Abstract

Lasso is a popular method for variable selection in regression. Much theoretical understanding has been obtained recently on its model selection or sparsity recovery properties under sparse and homoscedastic linear regression models. Since these standard model assumptions are often not met in practice, it is important to understand how Lasso behaves under nonstandard model assumptions.

In this paper, we study the sign consistency of the Lasso under one such model where the variance of the noise scales linearly with the expectation of the observation. This sparse Poisson-like model is motivated by medical imaging. In addition to studying the sign consistency, we also give sufficient conditions for ℓ_∞ consistency. With theoretical and simulation studies, we provide conditions for when the Lasso should not be expected to be sign consistent. One interesting finding is that β^* can not be spread out. Precisely, for both deterministic design and random Gaussian design, the sufficient conditions for the Lasso to be sign consistent require $\|\beta^*\|_2/[M(\beta^*)]^2$ to be not too big, where $M(\beta^*)$ is the smallest nonzero element of $|\beta^*|$. By special designs of $\mathbf{X}$, we show that $\|\beta^*\|_2/[M(\beta^*)]^2 = o(n)$ is almost necessary. For Positron Emission Tomography (PET), this suggests that when there are dense areas of the positron emitting substance, less dense areas are not well detected by the Lasso; this is of particular concern when imaging tumors; the periphery of the tumor will produce a much weaker signal than the center, leading to a big $\|\beta^*\|_2/[M(\beta^*)]^2$.

We compare the sign consistency of the Lasso under the Poisson-like model to its sign consistency on the standard model which assumes the noise is homoscedastic. The comparison shows that when β^* is spread out, the Lasso performs worse for data from the Poisson-like model than those from the standard model, confirming our theoretical findings.

Keywords: Lasso, Poisson-like Model, Sign Consistency, Heteroscedasticity

1. Introduction

The Lasso (Tibshirani, 1996) is now widely used in high dimensional regression for variable selection. Its model selection performance has been well studied under sparse and homoskedastic regression models. Several researchers have shown that under sparsity and

regularity conditions, the Lasso can select the true model asymptotically even when $p \gg n$ (Meinshausen and Buhlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009; Tropp, 2006; Donoho et al., 2006).

To define the Lasso estimate, suppose we observe independent pairs $\{(Y_i, x_i)\} \in R \times R^p$ for $i = 1, 2, \ldots, n$ following linear regression model

$$Y_i = x_i^T \beta^* + \epsilon_i, \tag{1}$$

where x_i^T is a row vector representing the predictors for the ith observation, Y_i is the corresponding ith response variable, ϵ_i are i.i.d. mean zero noise terms and independent of the predictors, and $\beta^* \in R^p$. Let us use $\mathbf{X} \in R^{n \times p}$ to denote the $n \times p$ design matrix with $x_k^T = (\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kp})$ as its kth row and with $X_j = (\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn})^T$ as its jth column, then

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = (X_1, X_2, \ldots, X_p).$$

Let $Y = (Y_1, \ldots, Y_n)^T$ and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T \in R^n$. The Lasso estimate is then defined as the solution to a penalized least squares problem (with regularization parameter λ):

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{2n} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1, \tag{2}$$

where for some vector $x \in R^k$, $\|x\|_r = (\sum_{i=1}^k |x_i|^r)^{1/r}$. In previous research with the Lasso (Knight and Fu, 2000; Meinshausen and Buhlmann, 2006; Zhao and Yu, 2006; Wainwright, 2009; Tropp, 2006; Donoho et al., 2006), the above model has been assumed where the noise terms are i.i.d. and independent of the predictors (hence homoskedastic). We call this the standard model.

Lustig et al. (2008) applied compressed sensing, a sparse method similar to the Lasso, to Magnetic Resonance Imaging (MRI). Candes and Tao (2007) suggests that the standard model could be useful for medical imaging technology like MRI. In this scenario, one hopes to collect far fewer measurements than usually required. However, the standard model is not the only model used for medical imaging. For imaging methods PET and SPECT the Poisson model is more appropriate (Fessler, 2000).

In PET, a subject is injected with a biochemical metabolite which is attractive to the tissue being studied. The biochemical metabolite is labeled with a positron emitting radioactive material. As the metabolite gathers around the tissue, so does the positron emitting radioactive material. The positron emissions are modeled by a Poisson process with an intensity rate which varies over the subject in direct relationship to the varying levels of biochemical metabolite. Therefore, an estimate of the intensity rate is an estimate of the level of biochemcial metabolite. Unfortunately, we do not observe the positron emission Poisson process directly. When a positron is emitted, it annihilates a nearby electron, sending two X-ray photons in nearly opposite directions (at the speed of light) Vardi et al. (1985). We observe these X-rays with several sensors in a ring around the subject. In our model, the sample size n represents the number of sensors; β_j^* represents the Poisson intensity rate for

a small cubic volume (a voxel) inside the subject; the design matrix $\mathbf{X}$ specifies the physics of the tomography and emissions process; finally, p is the number of voxels wanted, the more voxels, the finer the resolution of the final image.

With Poissonian noise, the variance of the noise is equal to the mean of the measurement. Motivated by the Poissonian model, we study Lasso under the following Poisson-like model:

$$\begin{aligned} Y &= \mathbf{X}\beta^* + \epsilon, \\ E(\epsilon \mid \mathbf{X}) &= 0, \\ Cov(\epsilon \mid \mathbf{X}) &= \sigma^2 \times diag(|\mathbf{X}\beta^*|), \\ \epsilon &\perp\!\!\!\perp X(S^c) \mid X(S), \end{aligned} \tag{3}$$

where $\sigma^2 > 0$ and the sparsity index set is defined as

$$S = \{1 \leq j \leq p : \beta_j \neq 0\}.$$

In the definition of the Poisson-like model, ϵ conditioned on $\mathbf{X}$ consists of independent Gaussian variables; $Cov(\epsilon \mid \mathbf{X})$, the variance-covariance matrix of ϵ conditioned on X, is $diag(|\mathbf{X}\beta^*|)$, an $n \times n$ diagonal matrix with the vector $|\mathbf{X}\beta^*|$ down the diagonal; and $X(S)$ and $X(S^c)$ denote two matrices consisting of the relevant column vectors (nonzero coefficients) and irrelevant column vectors (zero coefficients) respectively. Define $\beta^*(S) = \{\beta_j^* : \beta_j^* \neq 0\}$, $\beta^*(S^c) = \{\beta_j^* : \beta_j^* = 0\}$. This is a heteroscedastic model.

Since the Lasso provides a computationally feasible way to select a model (Osborne et al., 2000; Efron et al., 2004; Rosset, 2004; Zhao and Yu, 2007), it can be applied in the non-standard settings to give sparse solutions. It is possible that an altered version of the Lasso could better suit different non-standard settings. In the classical setup when n is growing and p is fixed, with heteroskedastic data, ordinary least squares is not desirable (Freedman, 2005). It is sometimes recommended to use weighted least squares instead of ordinary least squares. If we know the Poisson model holds, one could propose a similar fix for the Lasso for heteroscedastic data. However, we believe it is important to understand how the Lasso behaves before we propose fixes. Moreover, we would like to use this realistic Poisson model to start understanding how sensitive the Lasso is to nonstandard model assumptions.

With the Poisson-like model, for general scalings of p, q, n, and β^*, where $q = \#\ S$ is the number of true predictors in the linear model, we investigate when the Lasso is sign consistent and when it is not with theoretical and simulation studies. We also give sufficient conditions for the Lasso to be ℓ_∞ consistent. As far as we know, this is the first study of sign consistency and ℓ_∞ consistency using the Lasso in a non-homoscadestic setting, for general scalings of p, q, n and β^*.

1.1 Overview of Previous Work

The Lasso (Tibshirani, 1996) has been a popular technique to simultaneously select a model and provide regularized estimated coefficients. There is a substantial literature on the use of the Lasso for sparsity recovery and subset selection under the standard model. We provide only a very brief overview here.

In noiseless setting (when $\epsilon = 0$), with contributions from a broad range of researchers (Chen et al., 1998; Donoho and Huo., 2001; E. Candes and Tao., 2004; Elad and Bruckstein.,

2002, 2003; Tropp., 2004), there is now much understanding of sufficient conditions on deterministic predictors $\{X_i, i = 1, \ldots, n\}$ and sparsity index $S = \{j : \beta_j^* \neq 0\}$ for which the true β^* can be recovered exactly. Results by Donoho (2004), as well as Candes and Tao (2005) provide high probability results for random ensembles of $\mathbf{X}$. More specifically, as independently established by both sets of authors using different methods, if entries of $\mathbf{X}$ are i.i.d. from standard normal distribution $N(0, 1)$, with the number of predictors p scaling linearly in terms of the number of observations (i.e., $p = \gamma n$, for some $\gamma > 1$), there exists a constant $\alpha > 0$ such that all sparsity patterns with $q \leq \alpha p$ can be recovered with high probability.

There is also a substantial body of work focusing on the noisy setting (where ϵ is random noise). Knight and Fu (2000) analyze the asymptotic behavior of the optimal solution for fixed dimension (p); not only for L_1 regularization, but for L_r regularization with $r \in (0, 2]$. Both Tropp (2006) and Donoho et al. (2006) provide sufficient conditions for the support of the optimal solution to the Lasso problem (2) to be contained within the support of β^*. Recent work on the use of the Lasso for model selection by Meinshausen and Buhlmann (2006), focuses on Gaussian graphical models. Zhao and Yu (2006) considers linear regression and more general noise distributions. For the case of Gaussian noise and Gaussian predictors, both papers established that under particular mutual incoherence conditions and the appropriate choice of the regularization parameter λ, the Lasso can recover the sparsity pattern with probability converging to one for particular regimes of n, p and q. Zhao and Yu (2006) termed the mutual incoherence condition Irrepresentable Condition which they show is almost necessary when p is fixed. The Irrepresentable Condition was found in Fuchs (2005) and Zou (2006) as well. For i.i.d. Gaussian noise, Wainwright (2009) established a sharp relation between the problem dimension p, the number q of nonzero elements in β^*, and the number of observations n that are required for sign consistency.

1.2 Our Contributions

Before giving our contributions, we first give some definitions used throughout this paper. Define

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

We say that $\hat{\beta}(\lambda) =_s \beta^*$ if and only if $\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)$ elementwise.

Definition 1 *The Lasso is* ***sign consistent*** *if there exists a sequence λ_n such that,*

$$P\left(\hat{\beta}(\lambda_n) =_s \beta^*\right) \to 1, \text{ as } n \to \infty.$$

We study the sign consistency of the Lasso applied to data from the Poisson-like model. We give non-asymptotic results for both the deterministic design and the Gaussian random design. The non-asymptotic results give the probability that $\hat{\beta}(\lambda) =_s \beta^*$, for any λ, p, q, and n. A class of models which describes the relationship between n, p, q, and design matrix $\mathbf{X}$ are specified for the case of a deterministic design such that within this class the Lasso is sign consistent and ℓ_∞ consistent. A similar class is given for the Gaussian random design case. We also give necessary conditions for the Lasso to be sign consistent

under the Poisson-like model. We show that the Irrepresentable Condition is necessary for the Lasso's sign consistency under the Poisson-like Model. This condition is also necessary under the standard model (Wainwright, 2009; Zhao and Yu, 2006; Zou, 2006). The sufficient conditions for both the deterministic design and random Gaussian design requires that $\|\beta^*\|_2$ is not too large and $M(\beta^*)$ is not too small. Specifically, for deterministic design, assume that

$$\Lambda_{\min}\left(\frac{1}{n}X(S)^T X(S)\right) \geq C_{\min} > 0,$$

where $\Lambda_{\min}(\cdot)$ denotes the minimal eigenvalue of a matrix and $C_{\min}$ is some positive constant; for random Gaussian design, assume that

$$\Lambda_{\min}(\Sigma_{11}) \geq \tilde{C}_{\min} > 0 \quad \text{and} \quad \Lambda_{\max}(\Sigma) \leq \tilde{C}_{\max} < \infty,$$

where $\Sigma_{11} \in R^{q\times q}$ is the variance-covariance matrix of the true predictors, $\Sigma \in R^{p\times p}$ is the variance-covariance matrix of all predictors, $\Lambda_{\max}(\cdot)$ denotes the maximal eigenvalue of a matrix, and $\tilde{C}_{\min}$ and $\tilde{C}_{\max}$ are some positive constants. Then, the sufficient condition for deterministic design requires that, for some arbitrary $0 < \alpha < 1$,

$$\frac{\|\beta^*\|_2}{[M(\beta^*)]^{\frac{2}{\alpha}}} \leq \frac{n\eta^2}{2\sigma^2 \max_i \|x_i\|_2 (\, C_{\min}^{-1/2} + \sqrt{q}\, C_{\min}^{-1})^2 \log(p+1)}.$$

The sufficient condition for random Gaussian design requires that

$$\frac{\|\beta^*\|_2}{[M(\beta^*)]^2} \leq \frac{n\tilde{C}_{\min}}{4\sigma^2 \log n \sqrt{2\max(16q, 4\log n)}},$$

and

$$\frac{\|\beta^*\|_2}{[M(\beta^*)]^{2/\alpha}} \leq \frac{n\tilde{C}_{\min}^{2/\alpha}}{3\sigma^2 2^{2/\alpha} q \log(p-q+1)\sqrt{\tilde{C}_{\max}}}.$$

With a special design, we show that

$$\frac{\|\beta^*\|_2}{[M(\beta^*)]^2} = o\,(n)$$

is almost necessary for the Lasso to be sign consistent. These findings show $\|\beta^*\|_2$ is an important factor to both the sufficient conditions and necessary conditions. These results are different from the results for the standard model and they give insight into when the Lasso will not be sign consistent for the Poisson-like model. For Positron Emission Tomography (PET), the necessary condition means that if there are regions of very dense emissions (corresponding to very big β_j^*) and there are regions of very low (but still positive) emissions, then it is hard to reconstruct the positron emission Poisson intensity rate β^* with the Lasso. If the tissue of interest is a tumor, it might be difficult to estimate β^* with the Lasso, because there might be some area in the tumor with very dense emissions than other parts and the normal tissues.

With Gaussian design, we also find that it is necessary that the sample size n must grow faster than a lower bound defined as $c\, q \log(p-q)$. Where c is a constant which depends on

the variance-covariance matrix of the predictors. This condition is also needed under the standard model (Wainwright, 2009).

We use several techniques from Wainwright (2009), but the proofs in this paper are more difficult because of the heteroscedasticity. To control the variances of the the noise, we apply random matrix results regarding the Gaussian distribution and large deviation results regarding the χ^2 distribution.

The remainder of the paper is organized as follows. Section 2 analyzes the Lasso estimator under deterministic designs. Section 3 considers the case where the rows of $\mathbf{X}$ are i.i.d. Gaussian vectors. For both the deterministic and random designs we give sufficient conditions for the Lasso to be sign consistent and ℓ_∞ consistent. We also give necessary conditions for the Lasso's sign consistency. We then give simulations in Section 4 which demonstrate our theoretical findings. We conclude in Section 5.

2. Deterministic Design

In this section we consider $\mathbf{X}$ to be nonrandom. We study when the Lasso is sign consistent and when it is not sign consistent under the Poisson-like model. First, some notation,

$$x_i(S) = e_i^T X(S),$$

where e_i is the unit vector with ith element one and the rest zero. Because $S = \{j : \beta_j^* \neq 0\}$ is the sparsity index set, $x_i(S)$ is a row vector of dimension $\#\ S = q$. Define

$$M(\beta^*) = \min_{j \in S} |\beta_j^*| \quad \text{and} \quad \overrightarrow{b} = \text{sign}(\beta^*(S)).$$

Suppose the Irrepresentable Condition holds. That is, for some constant $\eta \in (0, 1]$,

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \leq 1 - \eta. \tag{4}$$

The ℓ_∞ norm of a vector, $\|\cdot\|_\infty$, is defined as the vector's largest element in absolute value. In addition, assume that

$$\Lambda_{\min} \left(\frac{1}{n} X(S)^T X(S) \right) \geq C_{\min} > 0, \tag{5}$$

where $\Lambda_{\min}$ denotes the minimal eigenvalue and $C_{\min}$ is some positive constant. Condition (5) guarantees that matrix $X(S)^T X(S)$ is invertible. These conditions are also needed in Wainwright (2009) for sign consistency of the Lasso under the standard model. Now define

$$\Psi(\mathbf{X}, \beta^*, \lambda) = \lambda \left[\eta \ (C_{\min})^{-1/2} + \left\| \left(\frac{1}{n} X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \right].$$

With this, we have:

Theorem 2 *Suppose that data $(\mathbf{X}, Y)$ follows Poisson-like model described by equations (3) and each column of $\mathbf{X}$ is normalized to l_2-norm $\sqrt{n}$. Assume that (4) and (5) hold. Then for any λ such that*

$$M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda),$$

each of the following properties holds with probability greater than

$$1 - 2\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2\|\beta^*\|_2 \max_{1\le i\le n}\|x_i(S)\|_2} + \log(p)\right\},$$

(a) The Lasso has a unique solution $\hat{\beta}(\lambda)$ *with* $\hat{\beta}(\lambda) =_s \beta^*$,

(b) $\left\|\hat{\beta}(\lambda) - \beta^*\right\|_\infty \le \Psi(\mathbf{X}, \beta^*, \lambda)$.

A proof of Theorem 2 can be found in Appendix A.1.

Theorem 2 gives a non-asymptotic result on the Lasso's sparsity pattern recovery property. The next corollary gives a well behaved class of models such that for a given choice of λ, when sample size n goes to infinity, the Lasso estimate recovers the sparsity pattern and $\|\hat{\beta}(\lambda) - \beta^*\|_\infty \to 0$ in probability. This class of models restricts the relationship between the data ($\mathbf{X}$), the coefficients (β^*), and the distribution of the noise (ϵ).

For any constant $0 < \alpha < 1$, define,

$$\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha) = \left(\frac{2\sigma^2\|\beta^*\|_2 \max_i \|x_i\|_2 (\, C_{\min}^{-1/2} + \sqrt{q}\; C_{\min}^{-1})^2 \log(p+1)}{n\eta^2}\right)^{\alpha/2}.$$

Corollary 3 *As in Theorem 2, suppose that data* $(\mathbf{X}, Y)$ *follows Poisson-like model described by equations (3) and each column of* $\mathbf{X}$ *is normalized to* l_2*-norm* $\sqrt{n}$. *Assume that (4) and (5) hold. If in addition,*

$$M(\beta^*) \ge \Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha),$$

for some arbitrary $0 < \alpha < 1$, *then by taking* λ *such that*

$$\lambda = \frac{\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)}{\eta\; C_{\min}^{-1/2} + \sqrt{q}\; C_{\min}^{-1}}, \tag{6}$$

we have that each of the following holds with probability greater than

$$1 - 2\exp\left\{-\left(\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)^{2-2/\alpha} - 1\right)\log(p+1)\right\},$$

(a) $\hat{\beta}(\lambda) =_s \beta^*$,

(b) $\|\hat{\beta}(\lambda) - \beta^*\|_\infty < \Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)$.

If $\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha) \to 0$, *then the probability of these events converges to one.*

The proof of this corollary can be found in Appendix A.2.

This corollary gives a class of heteroscedastic models for which the Lasso gives a sign consistent and ℓ_∞ consistent estimate of β^*. Sign consistency is important because it suggests which predictors should be included in the model and whether they have a positive or negative influence on the response. ℓ_∞ consistency is important because it says that each element of $\hat{\beta}(\lambda)$ cannot be far away from the corresponding element of β^*. From Corollary

3, this class requires that $M(\beta^*)$ not be too small; it should be greater than $\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)$. To be precise, condition $M(\beta^*) \geq \Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)$ suggests that there exists some $0 < \alpha < 1$,

$$\frac{\|\beta^*\|_2}{[M(\beta^*)]^{\frac{2}{\alpha}}} \leq \frac{n\eta^2}{2\sigma^2 \max_i \|x_i\|_2 (\, C_{\min}^{-1/2} + \sqrt{q}\, C_{\min}^{-1})^2 \log(p+1)}.$$

In analyzing the sign consistency of the Lasso under the standard model, Wainwright (2009) also requires that $M(\beta^*)$ not be too small. However, one big difference is that for the Poisson-like model in this paper, the results also depend on $\|\beta^*\|_2$.

The next corollary addresses the classical setting, where p, q, and β^* are all fixed and n goes to infinity. This is a straightforward result from Corollary 3 and yields better understanding of the Lasso applied to data from the Poisson-like model. Since $M(\beta^*)$ and $\|\beta^*\|_2$ do not change with n, $\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha) \to 0$ in Corollary 3 when $\frac{1}{n} \max_{1 \leq i \leq n} \|x_i(S)\|_2 \to 0$. Then we have:

Corollary 4 *As in Theorem 2, suppose that data $(\mathbf{X}, Y)$ follows Poisson-like model described by equations (3) and each column of $\mathbf{X}$ is normalized to l_2-norm $\sqrt{n}$. Assume that (4) and (5) hold. In the classical case when p, q and β^* are fixed, if*

$$\frac{1}{n} \max_{1 \leq i \leq n} \|x_i(S)\|_2 \to 0, \tag{7}$$

then by choosing λ as in equation (6),

$$P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \to 1, \text{ and } \|\hat{\beta}(\lambda) - \beta^*\|_\infty \to 0 \text{ in probability,}$$

as $n \to \infty$.

Condition (7) is not strong and it is easy to be satisfied. Suppose

$$0 < \Lambda_{\max}\left(\frac{1}{n} X(S)^T X(S)\right) \leq C_{\max},$$

where $\Lambda_{\max}(\cdot)$ is the maximum eigenvalue of a matrix and $C_{\max}$ is a positive constant, then

$$\left\| \frac{1}{\sqrt{n}} x_i(S) \right\|_2^2 = \left\| \frac{1}{\sqrt{n}} e_i^T X(S) \right\|_2^2 \leq \Lambda_{\max}\left(\frac{1}{n} X(S)^T X(S)\right) \leq C_{\max}.$$

Consequently,

$$\frac{1}{n} \max_{1 \leq i \leq n} \|x_i(S)\|_2 = \frac{1}{\sqrt{n}} \max_{1 \leq i \leq n} \left\| \frac{1}{\sqrt{n}} x_i(S) \right\|_2 \leq \frac{1}{\sqrt{n}} C_{\max} \to 0.$$

Corollary 4 states that in the classical settings, the Lasso can consistently select the true model under the Poisson-like model.

So far the results have given sufficient conditions for sign consistency of the Lasso. To understand how the Lasso might be sensitive to the heteroscedastic model, the next theorem gives necessary conditions which show that $\|\beta^*\|_2$ is in fact an important quantity for the Poisson-like model, due to the presence of β^* in $Cov(\epsilon|\mathbf{X})$.

Theorem 5 (Necessary Conditions) *Suppose that data* $(\mathbf{X}, Y)$ *follows Poisson-like model described by equations* (3) *and each column of* $\mathbf{X}$ *is normalized to* l_2*-norm* $\sqrt{n}$*. Assume that* (5) *holds.*

(a) Consider $\frac{1}{n}X(S)^T X(S) = I_{q\times q}$*. For any* j*, define*

$$c_{n,j}^2 = \frac{n^2\beta_j^{*2}}{\sigma^2 e_j^T \left[X(S)^T diag(|X\beta^*|)X(S)\right] e_j}. \tag{8}$$

Define $c_n = \min_j c_{n,j}$*. Then, for sign consistency, it is necessary that* $c_n \to \infty$*. Specifically,*

$$P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \le 1 - \frac{\exp\left\{-c_n^2/2\right\}}{\sqrt{2\pi}(1+c_n)}.$$

(b) If the Irrepresentable Condition (4) *does not hold, specifically,*

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S)\right)^{-1} \overrightarrow{b} \right\|_\infty = 1 + \zeta \text{ for some } \zeta \ge 0, \tag{9}$$

then, the Lasso estimate is not sign consistent: $P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \le \frac{1}{2}$;

A proof of Theorem 5 can be found in Appendix A.3.

Statement (a) says that under the Poisson-like model, the Lasso is sensitive to the scale of $\beta_j^*, i = 1, \dots q$. This is quite different from the results on the standard model which state that the Lasso is only sensitive to the smallest $|\beta_j^*|$ and not sensitive to the largest $|\beta_j^*|$. This makes sense, because when β_j^* becomes big in the standard model, the signal will increase while the noise keeps at the same level. But for Poisson-like data, when β_j^* grows, both the signal and the noise will grow. Statement (b) says that the Irrepresentable Condition (4) is necessary for the Lasso's sign consistency. This necessary condition can also be found in both Zhao and Yu (2006) and Wainwright (2009). Zhao and Yu (2006) points out that the Irrepresentable Condition is almost necessary and sufficient for the Lasso to be sign consistent under the standard model when p and q are fixed. Wainwright (2009) says that it is necessary for the Lasso's sign consistency under the standard model for any p and q. To understand the quantity $c_{n,j}$, we can consider $(X_j)_{j\in S}$ to follow a joint normal distribution $N(0, I_{q\times q})$, under which we have the following result.

Proposition 1 *Suppose that* $(X_j)_{j\in S}$ *follow a joint normal distribution* $N(0, I_{q\times q})$*. For* n *observations of the random vector* $(X_j)_{j\in S}$ *,* $c_{n,j}$ *as defined in* (8) *has the following property,*

$$P\left[c_{n,j}^2 \asymp \frac{n[\beta_j^*]^2}{\sigma^2\|\beta^*\|_2}\right] \to 1,$$

where the notation $A \asymp B$ *means that there exists two positive constants* c_1 *and* c_2*, such that* $c_2 B \le A \le c_1 B$*.*

The proof of Proposition 1 can be found in Appendix A.4.

Theorem 5 says that $c_{n,j} \to \infty$ is a necessary condition for sign consistency when $X(S)^T X(S)/n = I$. To understand the quantity $c_{n,j}$, this proposition gives an approximation by assuming a probabilistic model on $X(S)$. It shows that it is reasonable to suspect $c_{n,j} \to \infty$ if and only if $\|\beta^*\|_2/[M(\beta^*)]^2 = o(n)$. Therefore, $\|\beta^*\|_2/[M(\beta^*)]^2 = o(n)$ is almost necessary for sign consistency. This is not an exact necessary condition because $X(S)$ is approximated by a random matrix in Proposition 1. We will study how the Lasso is sensitive to $M(\beta^*)$ and $\|\beta^*\|_2$ with simulation studies in Section 4.

3. Gaussian Random Design

We now turn to the Gaussian random design where rows of $\mathbf{X}$ are i.i.d. from a p-dimensional multivariate Gaussian distribution with mean 0 and variance-covariance matrix Σ, which has unit diagonal entries. Define the variance-covariance matrix of the relevant predictors to be Σ_{11} and the covariance between the irrelevant predictors and the relevant predictors to be Σ_{21}. Specifically,

$$\begin{aligned} \Sigma_{11} &= E\left(\frac{1}{n}X(S)^T X(S)\right) \quad \text{and} \\ \Sigma_{21} &= E\left(\frac{1}{n}X(S^c)^T X(S)\right). \end{aligned}$$

Let $\Lambda_{\min}(\cdot)$ denote the minimum eigenvalue of a matrix and $\Lambda_{\max}(\cdot)$ denote the maximum eigenvalue of a matrix. To get the main results which allow p to grow with n, we need the following regularity conditions on the $p \times p$ matrix Σ. First, for some positive constants $C_{\min}$ and $C_{\max}$ which do not depend on n,

$$\Lambda_{\min}(\Sigma_{11}) \geq \tilde{C}_{\min} \quad \text{and} \quad \Lambda_{\max}(\Sigma) \leq \tilde{C}_{\max}, \tag{10}$$

and second, the Irrepresentable Condition,

$$\|\Sigma_{21}(\Sigma_{11})^{-1} sign(\beta^*(S))\|_\infty \leq 1 - \eta, \tag{11}$$

for some constant $\eta \in (0, 1]$. These are standard assumptions in previous work under standard models. Define,

$$\begin{aligned} V^*(n, \beta^*, \lambda, \sigma^2) &= \frac{2\lambda^2 q}{n\tilde{C}_{\min}} + \frac{3\sigma^2\sqrt{\tilde{C}_{\max}}\|\beta^*\|_2}{n}, \\ A(n, \beta^*, \sigma^2) &= \sqrt{\frac{4\sigma^2\|\beta^*\|_2 \log n \sqrt{2\max(16q, 4\log n)}}{n\tilde{C}_{\min}}} \quad \text{and} \\ \tilde{\Psi}(n, \beta^*, \lambda, \sigma^2) &= A(n, \beta^*, \sigma^2) + \frac{2\lambda\sqrt{q}}{\tilde{C}_{\min}}. \end{aligned}$$

With these quantities defined above, we have

Theorem 6 *Consider the Poisson-like model described by (3), under Gaussian random design. Suppose that the variance-covariance matrix Σ satisfies condition (10) and condition (11) with unit diagonal. Further, suppose that $q/n \to 0$. Then for any λ such that*

$$M(\beta^*) > \tilde{\Psi}(n, \beta^*, \lambda, \sigma^2),$$

each of the following properties holds with probability greater than

$$1 - 2\exp\left\{-\frac{\lambda^2\eta^2}{2V^*(n,\beta^*,\lambda,\sigma^2)\tilde{C}_{\max}} + \log(p-q)\right\} - (2q+3)\exp\{-0.03n\} - \frac{1+4q}{n},$$

(a) $\hat{\beta}(\lambda) =_s \beta^*$,

(b) $\left\|\hat{\beta}(\lambda) - \beta^*\right\|_\infty \le \tilde{\Psi}(n, \beta^*, \lambda, \sigma^2)$.

A proof of Theorem 6 can be found in Appendix A.5.

Theorem 6 gives a non-asymptotic result of the Lasso's sparsity pattern recovery property when the predictors are from Gaussian random ensemble. The next corollary gives a well behaved class of models such that for a given choice of λ, when sample size n goes to infinity, the Lasso estimate recovers the sparsity pattern and is ℓ_∞ consistent. This class of models restricts the relationship between the data ($\mathbf{X}$), the coefficients (β^*), and the distribution of the noise (ϵ). For any $\alpha \in (0,1)$, define

$$\tilde{\Gamma}(n, \beta^*, \sigma^2, \alpha) = \left(\frac{3\sigma^2 q\|\beta^*\|_2 \log(p-q+1)\sqrt{\tilde{C}_{\max}}}{n}\right)^{\alpha/2}.$$

Corollary 7 *As in Theorem 6, consider the Poisson-like model described by (3), under Gaussian random design. Suppose the variance-covariance matrix Σ satisfies condition (10) and condition (11) with unit diagnal. Further, suppose that $q/n \to 0$. If in addition,*

$$M(\beta^*) \ge A(n, \beta^*, \sigma^2) + \frac{2\tilde{\Gamma}(n, \beta^*, \sigma^2, \alpha)}{\tilde{C}_{\min}},$$

then by taking λ such that

$$\lambda = \frac{\tilde{\Gamma}(n, \beta^*, \sigma^2, \alpha)}{\sqrt{q}},$$

we have that each of the following holds with probability greater than

$$1 - 2\exp\left\{-\frac{\log(p-q+1)\eta^2}{2[\frac{2q\log(p-q+1)}{n\tilde{C}_{\min}} + \tilde{\Gamma}(n,\beta^*,\sigma^2,\alpha)^{\alpha/2-2}]\tilde{C}_{\max}} + \log(p-q)\right\}$$
$$-(2q+3)\exp\{-0.03n\} - \frac{1+4q}{n}$$

(a) $\hat{\beta}(\lambda) =_s \beta^*$,

(b) $\|\hat{\beta}(\lambda) - \beta^*\|_\infty < 2\tilde{\Gamma}(n, \beta^*, \sigma^2, \alpha)/\tilde{C}_{\min} + A(n, \beta^*, \sigma^2)$.

If

$$(\sigma^2\|\beta^*\|_2 + 1)q\log(p-q+1)/n \to 0, \tag{12}$$

then the probability of these events converges to one. If in addition $A(n, \beta^*, \sigma^2) \to 0$ *and* $\tilde{\Gamma}(n, \beta^*, \sigma^2, \alpha) \to 0$*, then* $\hat{\beta}(\lambda)$ *is* ℓ_∞ *consistent.*

A proof of Corollary 7 can be found in Appendix A.6.

This corollary gives a class of heteroscedastic models for which the Lasso gives a sign consistent and ℓ_∞ consistent estimate of β^*, when the predictors are from a Gaussian random ensemble. This class requires that $M(\beta^*)$ is not too small and $\|\beta^*\|_2$ not be too large. The sufficient conditions require that $M(\beta^*) \geq A(n, \beta^*, \sigma^2)$, which suggests that

$$\frac{\|\beta^*\|_2}{[M(\beta^*)]^2} \leq \frac{n\tilde{C}_{\min}}{4\sigma^2\log n\sqrt{2\max(16q, 4\log n)}};$$

and the sufficient conditions also require that $M(\beta^*) \geq 2\tilde{\Gamma}(n\beta^*, \sigma^2, \alpha)/\tilde{C}_{\min}$, which suggests that

$$\frac{\|\beta^*\|_2}{[M(\beta^*)]^{2/\alpha}} \leq \frac{n\tilde{C}_{\min}^{2/\alpha}}{3\sigma^2 2^{2/\alpha} q\log(p-q+1)\sqrt{\tilde{C}_{\max}}}.$$

To make the Lasso sign consistent, the sufficient condition (12) also needs n to grow faster than $q\log(p-q+1)$. The next theorem gives necessary conditions for the Lasso to be sign consistent. It says that if n is less than $2(\theta_l - v)q\log(p-q)$, where θ_l and v are two positive constants defined in next theorem, then the Lasso cannot be sign consistent for any $\lambda > 0$.

Theorem 8 (Necessary Conditions) *Consider the Poisson-like model described by (3), under Gaussian random design. Suppose the variance-covariance matrix* Σ *satisfies condition (10).*

(a) Suppose the Irrepresentable Condition (11) holds, and $q/n \to 0$*. Define*

$$\theta_l = \frac{(\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}})^2}{(2-\eta)^2\tilde{C}_{\max}}.$$

If for any $v > 0$*,* $n < 2(\theta_l - v)q\log(p-q)$*, then* $P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \to 0$*, for any* λ*.*

(b) If the Irrepresentable Condition (11) does not hold, specifically,

$$\|\Sigma_{21}(\Sigma_{11})^{-1}sign(\beta^*(S))\|_\infty = 1 + \zeta \text{ for some } \zeta \geq 0, \tag{13}$$

then, the Lasso estimate is not sign consistent: $P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \leq \frac{1}{2}$;

A proof of Corollary 8 can be found in Appendix A.7.

Claim (a) gives a necessary relationship between p, q, and n for the Lasso can be sign consistent. Claim (b) says that the Irrepresentable Condition (11) is a necessary condition for the Lasso's sign consistency.

In both deterministic design and Gaussian random design, we find that the sufficient conditions for the Lasso to be sign consistent require that $\|\beta^*\|_2/[M(\beta^*)]^2$ grow slow enough. By a special design, we showed that $\|\beta^*\|_2/[M(\beta^*)]^2 = o(n)$ is almost necessary. When using the Lasso on data which follows the Poisson-like model, one should be careful when β^* is spread out. If $\|\beta^*\|_2/[M(\beta^*)]^2$ is large, then the Lasso might not be able to select the true model. We will show the effect of $\|\beta^*\|_2/[M(\beta^*)]^2$ with simulations in the next section.

4. Simulation Studies

We present two simulations which investigate how the spread of β^* affects sign consistency and compares the standard model to the Poisson-like model. Our findings show that the β^* can not be spread out, that is $\|\beta^*\|_2/[M(\beta^*)]^2$ can not be too big. In the first example, we investigate how changing the values of $M(\beta^*)$ and $\|\beta^*\|_2$ result in changing the probability of sign consistency. The second experiment compares how the standard model and the Poisson-like model react to changing the value of $\|\beta^*\|_2$. All simulations were done in R with the LARS package (Efron et al., 2004).

Example 1 (Changing $c_{n,j}$) In this example, we study how the Lasso is sensitive to the spread of β^*, which is measured as Sp (spread) defined as

$$Sp = \frac{\|\beta^*\|_2}{[M(\beta^*)]^2}. \tag{14}$$

It can be argued that the bigger Sp, the β^* is more spread out. Our theoretical results in the previous sections suggest that when Sp is big, the probability that the Lasso is sign consistent might be small.

Consider an initial model with the parameters such that $n = 400$, $p = 1000$, $q = 20$, $\sigma^2 = 1$, and each element of the design matrix $\mathbf{X}$ is drawn independently from $N(0, 1)$. Once $\mathbf{X}$ is drawn, it will be fixed. β^* is designed this way,

$$\beta_j^* = \begin{cases} \beta_{\max} & \text{if } j \le 10 \\ \beta_{\min} & \text{if } 11 \le j \le 20 \\ 0 & \text{otherwise} \end{cases}$$

for $\beta_{\max} = 40$ and $\beta_{\min} = 5$. In this initial model, $\|\beta\|_2 = 127$, $Sp = 127/5^2 = 5.10$.

By changing $\beta_{\min}$ or $\|\beta^*\|_2$, we change the value of Sp. First, we fix $M(\beta^*) = \beta_{\min} = 5$, and change the value of $\beta_{\max}$. In this example, we take

$$\beta_{\max} \in \{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300\}.$$

Later we fix $\|\beta\|_2 = 127$, and then choose $\beta_{\min}$, such that Sp does not change from the first design. Part values of the parameters for the two designs are described in the following two tables.

For each design, we draw a plot of "success" versus Sp in Figure 1. The horizontal axis in Figure 1 is Sp. Each point along the solid line in Figure 1 corresponds to Design 1, and each point along the dashed line corresponds to Design 2. Success is defined as the existence of a λ which makes $\hat{\beta}(\lambda) =_s \beta^*$. The probability of success for each point is estimated with 1000 trials.

Table 1: This describes the relationship between $\|\beta^*\|_2, M(\beta^*)$, and Sp for the first simulation design. $M(\beta^*) = 5$ is fixed. With $\beta_{\max}$ changing, $\|\beta^*\|_2$ changes, so does $Sp = \|\beta^*\|_2/[M(\beta^*)]^2$.

$\beta_{\max}$	5	20	40	60	80	100	150	200	250	300
$\|\beta^*\|_2$	22	65	127	190	253	316	475	633	791	949
Sp	0.89	2.61	5.10	7.62	10.14	12.66	18.98	25.31	31.63	37.95

Table 2: This describes the relationship between $\|\beta^*\|_2, M(\beta^*)$, and Sp for the second simulation design. $\|\beta^*\|_2 = 127$ is fixed. Sp are keeping at the same values as in Design 1. Then $\beta_{\min}$ and $\beta_{\max}$ are decided by Sp and $\|\beta^*\|_2$. $\beta_{\min} = \sqrt{\|\beta^*\|_2/Sp}$ and $\beta_{\max} = \sqrt{\|\beta^*\|_2/10 - \beta_{\min}^2}$.

$\beta_{\min}$	11.94	6.99	5.00	4.09	3.55	3.17	2.59	2.24	2.01	1.83
$\beta_{\max}$	38.50	39.70	40.00	40.10	40.16	40.19	40.23	40.25	40.26	40.27
Sp	0.89	2.61	5.10	7.62	10.14	12.66	18.98	25.31	31.63	37.95

Figure 1 shows as Sp increases, the probability of success decreases. What is especially remarkable is that the dashed and solid lines are nearly identical.

Example 2 (Comparison to Standard Model) As seen in the theorems and the previous example, $\|\beta^*\|_2$ is an important quantity for sign consistency under the Poisson-like model. In this example, we compare the sign consistency of the Lasso on the Poisson-like model to the sign consistency of the Lasso on the standard model when the value of $\|\beta^*\|_2$ changes.

The following parameters do not change throughout this example: $n = 400$, $p = 1000$, $q = 20$, $\sigma^2 = 1$, and the design matrix has independent $N(0, 1)$ entries. We only want to change $\|\beta^*\|_2$, we do this by changing $\beta_{\max}$:

$$\beta_j^* = \begin{cases} \beta_{\max} & \text{if } j \leq 10 \\ 5 & \text{if } 11 \leq j \leq 20 \\ 0 & \text{otherwise.} \end{cases}$$

Under the initial model, $\beta_{\max} = 10$. This is the point furthest to the left of Figure 2. For this initial value of $\beta_{\max}$, $\|\beta^*\|_2 = \sqrt{1250} \approx 35.4$. For each subsequent point, $\beta_{\max}$ is chosen so that $\|\beta^*\|_2$ is a multiple of $\sqrt{1250}$. On the horizontal axis is how many times larger it should be. So, for the kth point, $\beta_{\max}$ is set so that $\|\beta^*\|_2 = k\sqrt{1250}$.

The final parameter to choose is the variance of the noise term in the standard model. It is chosen so that the expected standard deviation of the noise terms are equal under the initial model. If Z is a $N(0, 1)$ variable,

$$E\sqrt{|X_i\beta^*|} = E\sqrt{\|\beta^*\|_2|Z|} = \sqrt{\|\beta^*\|_2}E\sqrt{|Z|} \approx 4.9.$$

Success is defined as the existence of a λ which makes $\hat{\beta}(\lambda) =_s \beta^*$. The probability of success for each point is estimated with 1000 trials.

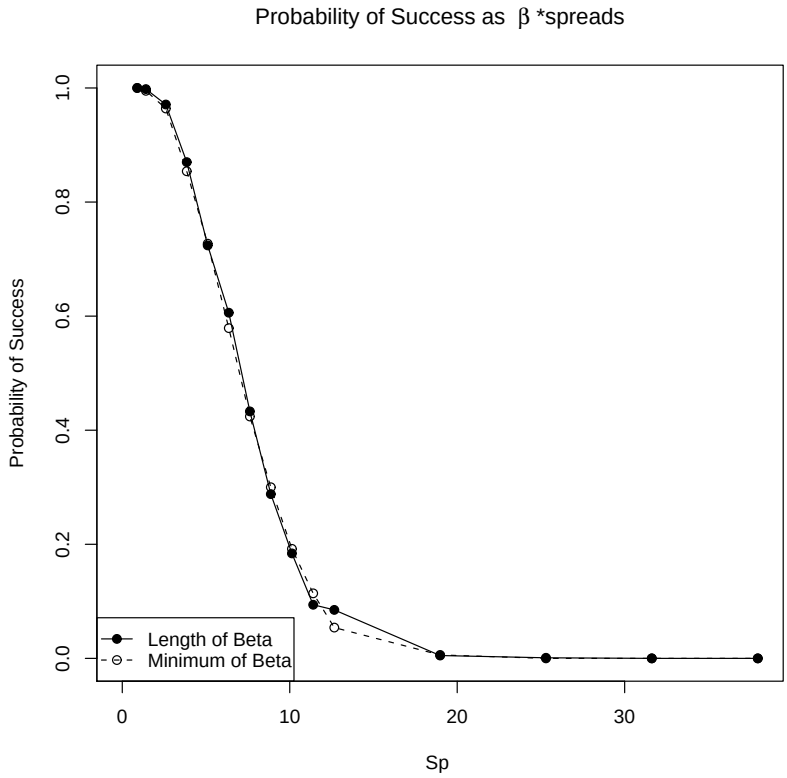

Figure 1: Probability of Success vs. Sp. For the solid line, $\|\beta^*\|_2$ decreases while $M(\beta^*)$ is kept constant. For the dashed line, $M(\beta^*)$ increases while $\|\beta^*\|_2$ is kept constant. The values of $M(\beta^*)$ and $\|\beta^*\|_2$ are chosen so that Sp as defined in (14) takes the values specified on the horizontal axis. Each probability is estimated with 1000 simulations.

What we see in Figure 2 is that increasing $\|\beta^*\|_2$ can have adverse effects on the sign consistency performance of the Lasso under the Poisson-like model. However, the performance on the standard model is almost constant.

The reason that sign consistency fails for the Poisson-like model when $\|\beta^*\|_2$ grows is that the variance of the noise also grows. As the variance of the noise grows, it becomes more and more difficult to detect the smallest elements of β^*. This does not occur in the standard model because $\|\beta^*\|_2$ does not affect the variance of the noise.

5. Conclusion

In this paper, we studied the sign consistency of the Lasso when the data is from a non-standard Poisson-like linear model which has heteroscedastic errors. This setup is different from the standard homoscedastic error model in preivous research. The Poisson-like model is

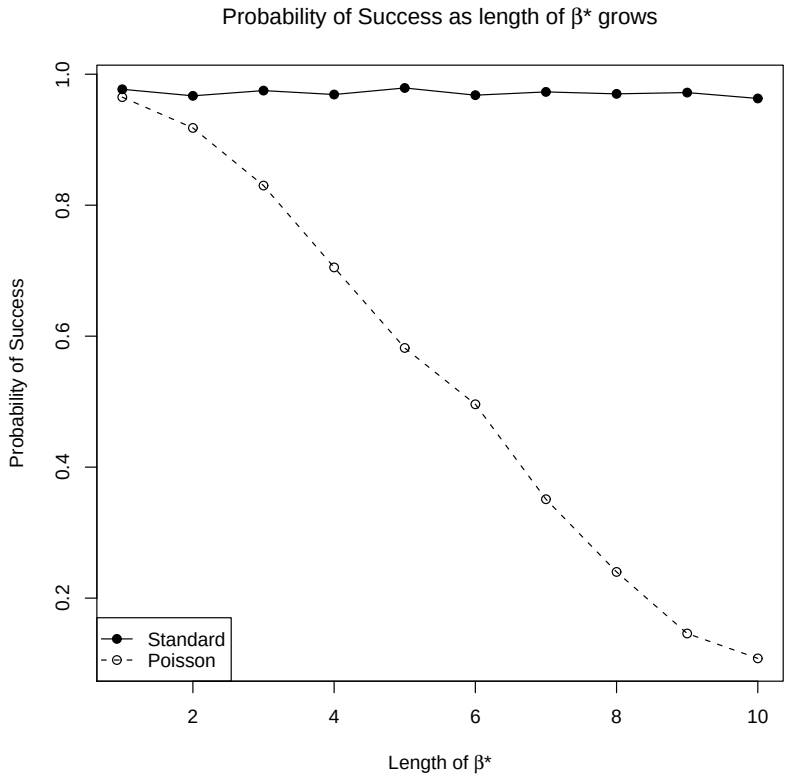

Figure 2: Probability of Success vs. the length of β^*. The solid line is the probability of success for the standard model and the dashed line is the probability of success for the Poisson-like model. The length of β^*, $\|\beta^*\|_2$ for the first point is $\sqrt{1250}$. For the kth point, $\|\beta^*\|_2 = k\sqrt{1250}$.

motivated by attempts of practitioners using the Lasso in high-dimensional medical imaging problems.

We gave non-asymptotic results for the Lasso's sign consistency property for both deterministic design and random Gaussian ensemble, under the Poisson-like model. Followed by these non-asymptotic results, a class of models dependent on the parameters $\beta^*, p, q, n, \sigma^2$, and the design matrix are specified. Under the specified models, a suitable λ was chosen such that the Lasso is sign consistent and ℓ_∞ consistent. We also studied how sensitive the Lasso is to the heteroscedastic model by finding necessary conditions. Both the deterministic design and the Gaussian random design show that $\|\beta^*\|_2/[M(\beta^*)]^2$ is an important factor. When it is too large, then the Lasso might not be sign consistent. With one special design, we show that $\|\beta^*\|_2/[M(\beta^*)]^2 = o(n)$ is almost necessary for the Lasso to be sign consistent. We call this condition spread-control condition. In PET, a patient is injected with a positron emitting substance which is attracted to a tissue of interest. Our results show that if there are regions of very dense emissions, then it is difficult for the Lasso to

detect regions of lower (but still positive) emissions. This is of particular interest when imaging tumors because the periphery of the tumor will have lower emissions than the center. By simulation studies, we demonstrated our theoretical findings. We showed that when β^* is spread out, the probability that the Lasso is sign consistent is very small. We also showed that when minimum of β_j^* is fixed, the growing $\|\beta^*\|_2$ has an adverse effect on the sign consistency performance of the Lasso under the Poisson-like model, while it almost has no any effect on the performance of the Lasso under the standard model.

We have not addressed possible improvements to the Lasso for heteroskedastic data. However, here are two possibilities. First, you might find a transformation of your variables which makes the errors more homoskedastic. Second, the Lasso estimator minimizes the residual sum of squares plus an ℓ_1 penalty. The residual sum of squares could be replaced with the weighted residual sum of squares, where the weights are inversely proportional to the variance of the noise. These suggestions are similar to those for least squares regression in classical case where p is fixed. They lead to a future study of the Poisson-like model.

Acknowledgments

This work is inspired by a personal communication between Bin Yu and Professor Peng Zhang from Capital Normal University in Beijing. We would like to thank Professor Ming Jiang and Vincent Vu for their helpful comments and suggestions on this paper. Jinzhu Jia is partially supported by NSF grants DMS-0605165 and SES-0835531. Karl Rohe is partially supported by a NSF VIGRE Graduate Fellowship. Bin Yu is partially supported by NSF grants DMS-0605165 and SES-0835531, ARO grant W911NF-05-1-0104, NSFC grant 60628102, and a grant from MSRA.

Appendix A. Proofs

A.1 Proof of Theorem 2

To prove the theorem, we need the next Lemma which gives necessary and sufficient conditions for the property $\mathcal{R}(\mathbf{X}, \beta^*, \epsilon, \lambda)$. They are important to the asymptotic analysis. Wainwright (2009) gives this condition which follows from KKT conditions.

Lemma 9 *For linear model $Y = \mathbf{X}\beta^* + \epsilon$, assume that the matrix $X(S)^T X(S)$ is invertible. Then for any given $\lambda > 0$ and any noise term $\epsilon \in R^n$, there exists a Lasso estimate $\hat{\beta}(\lambda)$ which satisfies $\hat{\beta}(\lambda) =_s \beta^*$, if and only if the following two conditions hold*

$$\left| X(S^c)^T X(S)(X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] - \frac{1}{n} X(S^c)^T \epsilon \right| \leq \lambda, \tag{15}$$

$$sign\left(\beta^*(S) + (\frac{1}{n} X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] \right) = sign(\beta^*(S)), \tag{16}$$

where the vector inequality and equality are taken elementwise. Moreover, if (15) holds strictly, then

$$\hat{\beta} = (\hat{\beta}^{(1)}, 0)$$

is the unique optimal solution to the Lasso problem (2), where

$$\hat{\beta}^{(1)} = \beta^*(S) + (\frac{1}{n} X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\hat{\beta}^{(1)}) \right]. \tag{17}$$

As in Wainwright (2009), we state sufficient conditions for (15) and (16). Define

$$\overrightarrow{b} = \text{sign}(\beta^*(S)),$$

and denote by e_i the vector with 1 in the ith position and zeroes elsewhere. Define

$$U_i = e_i^T (\frac{1}{n} X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda \overrightarrow{b} \right],$$

$$V_j = X_j^T \left\{ X(S)(X(S)^T X(S))^{-1} \lambda \overrightarrow{b} - \left[X(S)(X(S)^T X(S))^{-1} X(S)^T - I \right] \frac{\epsilon}{n} \right\}.$$

By rearranging terms, it is easy to see that (15) holds strictly if and only if

$$\mathcal{M}(V) = \left\{ \max_{j \in S^c} |V_j| < \lambda \right\} \tag{18}$$

holds. If we define $M(\beta^*) = \min_{j \in S} |\beta_j^*|$ (recall that $S = \{j : \beta_j^* \neq 0\}$ is the sparsity index), then the event

$$\mathcal{M}(U) = \left\{ \max_{i \in S} |U_i| < M(\beta^*) \right\}, \tag{19}$$

is sufficient to guarantee that condition (16) holds. So, under condition (18) and (19), by (17) the unique solution $\hat{\beta}(\lambda)$ satisfies

$$\|\hat{\beta}(\lambda) - \beta\|_\infty = \max_i |U_i|.$$

Finally, a proof of theorem 2.

Proof This proof is divided into two parts. First we analysis the asymptotic probability of event $\mathcal{M}(V)$, and then we analysis the event of $\mathcal{M}(U)$.

Analysis of $\mathcal{M}(V)$: Note from (18) that $\mathcal{M}(V)$ holds if and only if $\frac{\max_{j\in S^c}|V_j|}{\lambda} < 1$. Each random variable V_j is Gaussian with mean

$$\mu_j = \lambda X_j^T X(S)(X(S)^T X(S))^{-1}\overrightarrow{b}.$$

Define $\tilde{V}_j = X_j^T\left[I - X(S)(X(S)^T X(S))^{-1}X(S)^T)\right]\frac{\epsilon}{n}$, then $V_j = \mu_j + \tilde{V}_j$. Using condition (4), we have $|\mu_j| \leq (1-\eta)\lambda$ for all $j \in S^c$, from which we obtain that

$$\frac{1}{\lambda}\max_{j\in S^c}|\tilde{V}_j| < \eta \Rightarrow \frac{\max_{j\in S^c}|V_j|}{\lambda} < 1.$$

By the Gaussian comparison result (34) stated in Lemma 16, we have

$$P\left[\frac{1}{\lambda}\max_{j\in S^c}|\tilde{V}_j| \geq \eta\right] \leq 2(p-q)\exp\{-\frac{\lambda^2\eta^2}{2\max_{j\in S^c}E(\tilde{V}_j^2)}\}.$$

Since

$$E(\tilde{V}_j^2) = \frac{1}{n^2}X_j^T H[VAR(\epsilon)]HX_j,$$

where $H = I - X(S)(X(S)^T X(S))^{-1}X(S)^T$ which has maximum eigenvalue equal to 1, and $VAR(\epsilon)$ is the variance-covariance matrix of ϵ, which is a diagonal matrix with the ith diagonal element equal to $\sigma^2 \times |x_i^T\beta^*|$.

Since $|x_i^T\beta^*| \leq \sqrt{\|x_i(S)\|_2^2\|\beta^*\|_2^2} \leq \max_i\|x_i(S)\|_2\|\beta^*\|_2$, an operator bound yields

$$E(\tilde{V}_j^2) \leq \frac{\sigma^2}{n^2}\max_i\|x_i(S)\|_2\|\beta^*\|_2\|X_j\|_2^2 = \frac{\sigma^2}{n}\max_i\|x_i(S)\|_2\|\beta^*\|_2.$$

Therefore,

$$P\left[\frac{1}{\lambda}\max_j|\tilde{V}_j| \geq \eta\right] \leq 2(p-q)\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2\max_i\|x_i(S)\|_2\|\beta^*\|_2}\right\}.$$

So, we have

$$\begin{aligned} P\left[\frac{1}{\lambda}\max_j|V_j| < 1\right] &\geq 1 - P\left[\frac{1}{\lambda}\max_j|\tilde{V}_j| \geq \eta\right] \\ &\geq 1 - 2(p-q)\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}\right\}. \end{aligned}$$

Analysis of $\mathcal{M}(U)$:

$$\max_i|U_i| \leq \|(\frac{1}{n}X(S)^T X(S))^{-1}\frac{1}{n}X(S)^T\epsilon\|_\infty + \lambda\|(\frac{1}{n}X(S)^T X(S))^{-1}\overrightarrow{b}\|_\infty$$

Define $Z_i := e_i^T(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T\epsilon$. Each Z_i is a normal Gaussian with mean 0 and variance

$$\begin{aligned} var(Z_i) &= e_i^T(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T[VAR(\epsilon)]\frac{1}{n}X(S)(\frac{1}{n}X(S)^TX(S))^{-1}e_i \\ &\leq \frac{\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}{nC_{\min}}. \end{aligned}$$

So, for any $t>0$, by (34)

$$P(\max_{i\in S}|Z_i|\geq t)\leq 2q\exp\{-\frac{t^2nC_{\min}}{2\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}\},$$

by taking $t=\frac{\lambda\eta}{\sqrt{C_{\min}}}$, we have

$$P(\max_{i\in S}|Z_i|\geq\frac{\lambda\eta}{\sqrt{C_{\min}}})\leq 2q\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}\right\}.$$

Recall the definition of $\Psi(\mathbf{X},\beta^*,\lambda)=\lambda\left[\eta\ (C_{\min})^{-1/2}+\left\|\left(\frac{1}{n}X(S)^TX(S)\right)^{-1}\overrightarrow{b}\right\|_\infty\right]$, we have

$$P(\max_i|U_i|\geq\Psi(\mathbf{X},\beta^*,\lambda))\leq 2q\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}\right\}.$$

By condition $M(\beta^*)>\Psi(\mathbf{X},\beta^*,\lambda)$, we have

$$P(\max_i|U_i|<M(\beta^*))\geq 1-2q\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}\right\}.$$

At last, we have

$$P\left[\mathcal{M}(V)\&\ \mathcal{M}(U)\right]\geq 1-2p\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}\right\}$$

■

A.2 Proof of Corollary 3

Proof Recall the definition of $\Gamma(\mathbf{X},\beta^*,\sigma^2,\alpha)$:

$$\Gamma(\mathbf{X},\beta^*,\sigma^2,\alpha)=\left(\frac{2\sigma^2\|\beta^*\|_2\max_i\|x_i\|_2(\eta\ C_{\min}^{-1/2}+\sqrt{q}\ C_{\min}^{-1})^2\log(p+1)}{n\eta^2}\right)^{\alpha/2}.$$

So,

$$\frac{n\eta^2}{2\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}=\Gamma(\mathbf{X},\beta^*,\sigma^2,\alpha)^{-2/\alpha}(\eta\ C_{\min}^{-1/2}+\sqrt{q}\ C_{\min}^{-1})^2\log(p+1)$$

By taking

$$\lambda = \frac{\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)}{\eta\ C_{\min}^{-1/2} + \sqrt{q}\ C_{\min}^{-1}},$$

we have

$$\Psi(\mathbf{X}, \beta^*, \lambda) \leq \lambda \left[\eta\ C_{\min}^{-1/2} + \sqrt{q} C_{\min}^{-1}\right] = \Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha),$$

and

$$\frac{n\lambda^2\eta^2}{2\sigma^2\|\beta^*\|_2 \max_i \|x_i(S)\|_2} = \Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)^{2-2/\alpha} \log(p+1).$$

So, the probability bound in Theorem 3 greater than

$$1 - 2\exp\left\{-\left(\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha)^{2-2/\alpha} - 1\right)\log(p+1)\right\},$$

which goes to one when $\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha) \to 0$.

■

A.3 Proof of Theorem 5

Proof First prove (b). Without loss of generality, assume for some $j \in S^c$, $X_j^T X(S)\left(X(S)^T X(S)\right)^{-1}\overrightarrow{b} = 1+\zeta$, then $V_j = \lambda(1+\zeta) + \tilde{V}_j$, where $\tilde{V}_j = -[X(S)\left(X(S)^T X(S)\right)^{-1} X(S)^T - I]\frac{\epsilon}{n}$ is a Gaussian random variable with mean 0, so $P(\tilde{V}_j > 0) = \frac{1}{2}$. So, $P(V_j > \lambda) \geq \frac{1}{2}$, which implies that for any λ, Condition (15) (a necessary condition) is violated with probability greater than 1/2.

For claim (a). Condition (16),

$$sign\left(\beta^*(S) + (\frac{1}{n}X(S)^T X(S))^{-1}\left[\frac{1}{n}X(S)^T\epsilon - \lambda sign(\beta^*(S))\right]\right) = sign(\beta^*(S))$$

is also a necessary condition for sign consistency. Since $\frac{1}{n}X(S)^T X(S) = I_{q\times q}$, (16) becomes

$$sign\left(\beta^*(S) + \left[\frac{1}{n}X(S)^T\epsilon - \lambda sign(\beta^*(S))\right]\right) = sign(\beta^*(S)),$$

which implies that

$$sign\left(\beta^*(S) + \frac{1}{n}X(S)^T\epsilon\right) = sign(\beta^*(S)). \tag{20}$$

Without loss of generality, assume for some $j \in S$, $\beta_j^* > 0$. Then (20) implies $\beta_j^* + Z_j > 0$, where $Z_j = e_j^T \frac{1}{n}X(S)^T\epsilon$ is a Gaussian random variable with mean 0, and variance

$$\begin{aligned} var(Z_j) &= e_j^T \frac{1}{n}X(S)^T VAR(\epsilon)\frac{1}{n}X(S)e_j \\ &= \frac{\sigma^2 e_j^T\left[X(S)^T diag(|X\beta^*|)X(S)\right]e_j}{n^2} \\ &= \frac{\beta_j^{*2}}{c_{n,j}^2}, \end{aligned}$$

where the last equality uses the definition of $c_{n,j}^2$ in Theorem 5. To Summarize,

$$\begin{aligned}
P[\hat{\beta}(\lambda) =_s \beta^*] &\leq P[\beta_j^* + Z_j > 0] \\
&= P[Z_j > -\beta_j^*] \\
&= P[Z_j < \beta_j^*] \\
&= 1 - \int_{\beta_j^*}^{\infty} \frac{1}{\sqrt{2\pi var(Z_j)}} \exp\{-\frac{x^2}{2var(Z_j)}\}dx \\
&= 1 - \int_{\beta_j^*/\sqrt{var(Z_j)}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}dx \\
&\leq 1 - \frac{1}{\sqrt{2\pi}} \int_{\beta_j^*/\sqrt{var(Z_j)}}^{\infty} (\frac{x}{1+x} + \frac{1}{(1+x)^2}) \exp\{-\frac{x^2}{2}\}dx \\
&= 1 - \frac{\exp\left\{-\frac{\beta_j^{*2}}{var(Z_j)}\right\}}{\sqrt{2\pi}(1 + \frac{\beta_j^*}{\sqrt{var(Z_j)}})} \\
&= 1 - \frac{\exp\left\{-\frac{c_{n,j}^2}{2}\right\}}{\sqrt{2\pi}(1 + c_{n,j})}.
\end{aligned}$$

■

A.4 Proof of Proposition 1

Proof Let $Z_j = e_j^T X(S)^T diag(|X\beta^*|)X(S)e_j$. Since

$$e_j^T X(S)^T diag(|X\beta^*|)X(S)e_j = \sum_{i=1}^{n} |x_i^T\beta^*|X_{ij}^2,$$

$$|x_i^T\beta^*|X_{ij}^2 \perp\!\!\!\perp |x_k^T\beta^*|X_{ki}^2$$

for any $i \neq k$. So, we have

$$E(Z_j) = \sum_{i=1}^{n} E(|x_i^T\beta^*|X_{ij}^2),$$

and

$$var(Z_j) = \sum_{i=1}^{n} var(|x_i^T\beta^*|X_{ij}^2).$$

Since $var(x_i^T\beta^*) = \|\beta^*\|_2^2$, $var(X_{ij}) = 1$, $\rho(x_i^T\beta^*, X_{ij}) = \frac{\beta_j^*}{\|\beta^*\|_2}$, by Lemma 10, we have

$$c_3\|\beta^*\|_2 \leq E(|x_i^T\beta^*|X_{ij}^2) \leq c_1\|\beta^*\|_2,$$

and

$$3\|\beta^*\|_2^2 \leq E([|x_i^T\beta^*|X_{ij}^2]^2) \leq 15\|\beta^*\|_2^2.$$

So,

$$\begin{aligned} P[|Z_j - E(Z_j)| > \frac{1}{2}E(Z_j)] &\leq \frac{4\ var(Z_j)}{E(Z_j)^2} \\ &\leq \frac{4 \times 15n\|\beta^*\|_2^2}{[c_3 n\|\beta^*\|_2]^2} \\ &= \frac{60}{c_3^2 n}. \end{aligned}$$

So, we have

$$P\left[\frac{c_3 n}{2}\|\beta^*\|_2 \leq Z_j \leq \frac{3c_1 n}{2}\|\beta^*\|_2\right] \geq 1 - \frac{60}{c_3^2 n}.$$

From which we have

$$P\left[\frac{2n\beta_j^{*2}}{3c_1\sigma^2\|\beta^*\|_2} \leq \frac{n^2\beta_j^{*2}}{\sigma^2 Z_j} \leq \frac{2n\beta_j^{*2}}{c_3\sigma^2\|\beta^*\|_2}\right] \geq 1 - \frac{60}{c_3^2 n}.$$

Substituting $c_{n,j}^2 = n^2\beta_j^{*2}/(\sigma^2 Z_j)$, yields the results.

Lemma 10 *Suppose that random variables X and Y, follow a joint normal distribution, with means $E(X) = 0$ and $E(Y) = 0$, variances $var(X) = \sigma_X^2$ and $var(Y) = \sigma_Y^2$, and correlation $\rho(X, Y) = \rho$. Then we have*

$$E(|X|Y^2) = c_3\rho^2\sigma_X\sigma_Y^2 + c_1(1-\rho^2)\sigma_X\sigma_Y^2,$$

and

$$E(|X|^2Y^4) = (12\rho^2 + 3)\sigma_X^2\sigma_Y^4.$$

The proof of this lemma can be found in Appendix A.8.

■

A.5 Proofs of Theorem 6

To prove Theorem 6, we need some preliminary results.

Lemma 11 *Conditioned on $X(S)$ and ϵ, the random vector V is Gaussian. Its mean vector is upper bound as*

$$|\ E[V|\epsilon, X(S)]\ | \leq \lambda(1-\eta)\mathbf{1}. \tag{21}$$

Moreover, its conditional covariance takes the form

$$cov[V|\epsilon, X(S)] = M_n\Sigma_{2|1} = M_n[\Sigma_{22} - \Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}], \tag{22}$$

where

$$M_n = \lambda^2\overrightarrow{b}^T(X(S)^TX(S))^{-1}\overrightarrow{b} + \frac{1}{n^2}\epsilon^T[I - X(S)(X(S)^TX(S))^{-1}X(S)^T]\epsilon. \tag{23}$$

Lemma 12 *Let* $M_1 = \lambda^2 \overrightarrow{b}^T (X(S)^T X(S))^{-1} \overrightarrow{b}$ *and* $M_2 = \frac{1}{n^2} \epsilon^T [I - X(S)(X(S)^T X(S))^{-1} X(S)^T] \epsilon$, *then* $M_n = M_1 + M_2$. *We have*

$$P\left[\frac{\lambda^2 q}{2n\tilde{C}_{\max}} \leq M_1 \leq \frac{2\lambda^2 q}{n\tilde{C}_{\min}}\right] \geq 1 - \exp\{-0.03n\}, \tag{24}$$

$$P\left[M_2 \geq \frac{3\sigma^2\sqrt{\tilde{C}_{\max}}\|\beta^*\|_2}{n}\right] \leq \frac{1}{n}. \tag{25}$$

Lemma 13

$$P\left[\max_{i=1,\ldots n} \|x_i(S)\|_2^2 \geq 2\tilde{C}_{\max} \max(16q, 4\log n)\right] \leq \frac{1}{n}. \tag{26}$$

Proofs of these lemmas can be found in Appendix A.8. Now, we prove Theorem 6.

Analysis of $M(V)$**:** Define the event $T = \{M_n \geq v^*\}$, where

$$v^* = \frac{2\lambda^2 q}{n\tilde{C}_{\min}} + \frac{3\sigma^2\sqrt{\tilde{C}_{\max}}\|\beta^*\|_2}{n}.$$

By Lemma 12, we have $P[T] \leq \exp\{-0.03n\} + \frac{1}{n}$.

Let $\mu_j = E[V_j|\epsilon, X(S)]$, $Z_j = V_j - \mu_j$, and $Z = (Z_j)_{j\in S^c}$, then $E[Z|X(S),\epsilon] = 0$ and $cov(Z|X(S),\epsilon) = cov(V|X(S),\epsilon) = M_n\Sigma_{2|1}$.

$$\begin{aligned}
\max_{j\in S^c}|V_j| &= \max_{j\in S^c}|\mu_j + Z_j| \\
&\leq \max_{j\in S^c}[|\mu_j| + |Z_j|] \\
&\leq (1-\eta)\lambda + \max_{j\in S^c}|Z_j|.
\end{aligned}$$

From this inequality, we have

$$\{\max_{j\in S^c}|Z_j| < \eta\lambda\} \subset \{\max_{j\in S^c}|V_j| < \lambda\}.$$

Define $\tilde{Z}$ to be a zero-mean Gaussian with covariance $v^*\Sigma_{2|1}$. Since

$$\begin{aligned}
P\left[\max_{j\in S^c}|Z_j| \geq \eta\lambda \mid T^c\right] &\leq \sum_j P\left[|Z_j| > \eta\lambda \mid T^c\right] \\
&\leq (p-q)P\left[\max_{j\in S^c}|\tilde{Z}_j| > \eta\lambda\right] \\
&\leq 2(p-q)\exp\{-\frac{\eta^2\lambda^2}{2v^*\tilde{C}_{\max}}\},
\end{aligned}$$

we have

$$\begin{aligned}
P[\max_{j\in S^c}|V_j| \geq \lambda] &\leq P\left[\max_{j\in S^c}|Z_j| \geq \lambda \mid T^c\right] + P[T] \\
&\leq 2(p-q)\exp\{-\frac{\eta^2\lambda^2}{2v^*\tilde{C}_{\max}}\} + \exp\{-0.03n\} + \frac{1}{n}.
\end{aligned}$$

This says that

$$P[\mathcal{M}(V)] \geq 1 - 2(p-q)\exp\{-\frac{\eta^2\lambda^2}{2v^*\tilde{C}_{\max}}\} - \exp\{-0.03n\} - \frac{1}{n}.$$

Analysis of $\mathcal{M}(U)$: Now we analyze $\max_{j\in S}|U_j|$.

$$\max_j |U_j| \leq \left\|(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T\epsilon\right\|_\infty + \lambda\left\|(\frac{1}{n}X(S)^TX(S))^{-1}\overrightarrow{b}\right\|_\infty.$$

Define $\Lambda_i(\cdot)$ to be the ith largest eigenvalue of the matrix. Since

$$\lambda\left\|(\frac{1}{n}X(S)^TX(S))^{-1}\overrightarrow{b}\right\|_\infty \leq \frac{\lambda\sqrt{q}}{\Lambda_{\min}(\frac{1}{n}X(S)^TX(S))},$$

by Equation (37) in Lemma 20,

$$P\left[\frac{1}{2}\tilde{C}_{\min} \leq \Lambda_i(\frac{1}{n}X^TX) \leq 2\tilde{C}_{\max}\right] \geq 1 - 2\exp(-0.03n),$$

we have

$$P\left[\lambda\left\|(\frac{1}{n}X(S)^TX(S))^{-1}\overrightarrow{b}\right\|_\infty \leq \frac{2\lambda\sqrt{q}}{\tilde{C}_{\min}}\right] \geq 1 - 2\exp(-0.03n).$$

Let

$$W_i = e_i^T(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T\epsilon,$$

then conditioned on $X(S)$, W_i is a Gaussian random variable with mean 0, and variance

$$\begin{aligned} var(W_i|X(S)) &= e_i^T(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T[VAR(\epsilon)]\frac{1}{n}X(S)(\frac{1}{n}X(S)^TX(S))^{-1}e_i \\ &\leq \frac{\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}{n\Lambda_{\min}(\frac{1}{n}X(S)^TX(S))}. \end{aligned}$$

Using (37)

$$P\left[\frac{1}{2}\tilde{C}_{\min} \leq \Lambda_i(\frac{1}{n}X^TX) \leq 2\tilde{C}_{\max}\right] \geq 1 - 2\exp(-0.03n),$$

and Lemma 13, we have

$$\frac{\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}{n\Lambda_{\min}(\frac{1}{n}X(S)^TX(S))} \leq \frac{2\sigma^2\|\beta^*\|_2\sqrt{2\tilde{C}_{\max}\max(16q, 4\log n)}}{n\tilde{C}_{\max}}$$

with probability no less than $1 - 2\exp\{-0.03n\} - \frac{1}{n}$.

Define event

$$\mathcal{T} = \left\{\frac{\sigma^2\|\beta^*\|_2\max_i\|x_i(S)\|_2}{n\Lambda_{\min}(\frac{1}{n}X(S)^TX(S))} \leq \frac{2\sigma^2\|\beta^*\|_2\sqrt{2\tilde{C}_{\max}\max(16q, 4\log n)}}{n\tilde{C}_{\min}}\right\},$$

then $P(\mathcal{T}) \geq 1 - 2\exp\{-0.03n\} - \frac{1}{n}$. From the proof of Lemma 16, for any $t > 0$,

$$P(|W_i| > t \mid X(S), \mathcal{T}) \leq 2\exp(-\frac{t^2}{2var(W_i \mid X(S), \mathcal{T})}).$$

The above is also true if we replace $var(W_i \mid X(S), \mathcal{T})$ with any upper bound. So, we have

$$P(|W_i| > t \mid X(S), \mathcal{T}) \leq 2\exp\left\{-\frac{t^2}{2\frac{2\sigma^2\|\beta^*\|_2\sqrt{2\tilde{C}_{\max}\max(16q, 4\log n)}}{n\tilde{C}_{\min}}}\right\}.$$

So,

$$\begin{aligned} P(|W_i| > t) &\leq P(|W_i| > t \mid \mathcal{T}) + P(\mathcal{T}) \\ &\leq 2\exp\left\{-\frac{t^2}{2\frac{2\sigma^2\|\beta^*\|_2\sqrt{2\tilde{C}_{\max}\max(16q, 4\log n)}}{n\tilde{C}_{\min}}}\right\} + 2\exp\{-0.03n\} + \frac{1}{n}. \end{aligned}$$

By takeing $t = A(n, \beta^*, \sigma^2) := \sqrt{\frac{4\sigma^2\|\beta^*\|_2\log n\sqrt{2\max(16q, 4\log n)}}{n\tilde{C}_{\min}}}$, we have

$$\begin{aligned} P\left[\max_i |W_i| > A(n, \beta^*, \sigma^2)\right] &\leq \frac{2q}{n} + 2q\exp\{-0.03n\} + \frac{q}{n} \\ &= \frac{3q}{n} + 2q\exp\{-0.03n\}. \end{aligned}$$

Summarize,

$$\begin{aligned} &P\left[\max_i U_i \geq A(n, \beta^*, \sigma^2) + \frac{2\lambda\sqrt{q}}{\tilde{C}_{\min}}\right] \\ \leq &\frac{3q}{n} + 2q\exp\{-0.03n\} + \frac{q}{n} + 2\exp\{-0.03n\} \quad . \end{aligned}$$

At last, we have

$$P\left[\mathcal{M}(V) \;\&\; \mathcal{M}(U)\right] \leq 1 - 2(p-q)\exp\{-\frac{\eta^2\lambda^2}{2v^*\tilde{C}_{\max}}\} - (2q+3)\exp\{-0.03n\} - \frac{1+4q}{n}.$$

A.6 Proofs of Corollary 7

Proof By taking $\lambda = \frac{\tilde{\Gamma}(n, \beta^*, \sigma^2)}{\sqrt{q}}$, where

$$\tilde{\Gamma}(n, \beta^*, \sigma^2) = \left(\frac{3\sigma^2 q\|\beta^*\|_2\log(p-q+1)\sqrt{\tilde{C}_{\max}}}{n}\right)^{\alpha/2},$$

and $\alpha < 1$, we have

$$\begin{aligned}\frac{\lambda^2}{V^*(n,\beta^*,\lambda,\sigma^2)} &= \frac{\lambda^2}{\frac{2\lambda^2 q}{n\tilde{C}_{\max}} + \frac{3\sigma^2\sqrt{\tilde{C}_{\max}}\|\beta^*\|_2}{n}} \\ &= \frac{1}{\frac{2q}{n\tilde{C}_{\max}} + \frac{3\sigma^2\sqrt{\tilde{C}_{\max}}\|\beta^*\|_2}{n\lambda^2}} \\ &= \frac{\log(p-q+1)}{\frac{2q\log(p-q+1)}{n\tilde{C}_{\max}} + \tilde{\Gamma}(n,\beta^*,\sigma^2,\alpha)^{2/\alpha-2}}.\end{aligned}$$

By Condition (12), $q\log(p-q+1)/n \to 0$ and $\tilde{\Gamma}(n,\beta^*,\sigma^2,\alpha) \to 0$,

$$\frac{\lambda^2}{V^*(n,\beta^*,\lambda,\sigma^2)\log(p-q+1)} \to \infty.$$

This guarantees $P[\hat{\beta}(\lambda) =_s \beta^*] \to 1$. In fact, the probability bound in Theorem 6 now becomes,

$$\begin{aligned}&1 - 2\exp\left\{-\frac{\lambda^2\eta^2}{2V^*(n,\beta^*,\lambda,\sigma^2)\tilde{C}_{\max}} + \log(p-q)\right\} - (2q+3)\exp\{-cn\} - \frac{1+4q}{n} \\ &= 1 - 2\exp\left\{-\frac{\log(p-q+1)\eta^2}{2[\frac{2q\log(p-q+1)}{n\tilde{C}_{\max}} + \tilde{\Gamma}(n,\beta^*,\sigma^2,\alpha)^{\alpha/2-2}]\tilde{C}_{\max}} + \log(p-q)\right\} \\ &-(2q+3)\exp\{-cn\} - \frac{1+4q}{n}\end{aligned}$$

By the choice of λ, we have that $\Psi(n,\beta^*,\lambda,\sigma^2) = A(n,\beta^*,\sigma^2) + \frac{2\tilde{\Gamma}(n,\beta^*,\sigma^2)}{\tilde{C}_{\min}}$, which goes to 0, if (12) holds and $A(n,\beta^*,\sigma^2) \to 0$. ■

A.7 Proof of Theorem 8

Proof First, prove part (b). Without loss of generality, assume

$$e_j^T\Sigma_{21}(\Sigma_{11})^{-1}sign(\beta^*(S)) = 1+\zeta,$$

for some $j \in S^c$. Since $E[V|X(S),\epsilon] = \lambda\Sigma_{21}(\Sigma_{11})^{-1}sign(\beta^*(S))$, V_j, conditioned on $X(S)$ and ϵ, is a Gaussian random variable with mean $\lambda(1+\zeta)$. So $P[V_j \geq \lambda(1+\zeta)|X(S),\epsilon] = \frac{1}{2}$, which implies $P[V_j > \lambda|X(S),\epsilon] \geq \frac{1}{2}$. Then we have $P(V_j > \lambda) \geq \frac{1}{2}$. So for any λ,

$$P[\hat{\beta}(\lambda) =_s \beta^*] \leq P[\max_j V_j \leq \lambda] \leq \frac{1}{2}.$$

Now, a proof for claim (a). Let $V_j = E[V_j] + \tilde{V}_j$, then $|E[V|X(S),\epsilon]| = \left|\lambda\Sigma_{21}(\Sigma_{11})^{-1}sign(\beta^*(S))\right| \leq \lambda$ by Condition (11), and $\tilde{V}_j$ is a zero-mean random variable. Since

$$\begin{aligned}\max_{j\in S^c}|V_j| &\geq \max_{j\in S^c}|\tilde{V}_j| - \max_{j\in S^c}|E[V_j]| \\ &\geq \max_{j\in S^c}|\tilde{V}_j| - (1-\eta)\lambda,\end{aligned}$$

So,

$$P\left[\max_{j\in S^c}|V_j|>\lambda\right]\geq P\left[\max_{j\in S^c}|\tilde{V}_j|>(2-\eta)\lambda\right].$$

Conditioned on $X(S)$ and ϵ, the random vector $(V_j)_{j\in S^c}$ is Gaussian with covariance matrix $M_n\Sigma_{2|1}$; thus the zero-mean version $(\tilde{V})_{j\in S^c}$has the same covariance matrix. Defining the event $\mathcal{T}=\{M_1>\frac{\lambda^2 q}{2nC_{\max}}\}$, we have $P[\mathcal{T}]\to 0$ by Lemma 12, and

$$\begin{aligned}P\left[\max_{j\in S^c}|\tilde{V}_j|>(2-\eta)\lambda\right] &\geq (1-P[\mathcal{T}])P\left[\max_{j\in S^c}|\tilde{V}_j|>(2-\eta)\lambda\mid\mathcal{T}^c\right]\\ &\geq (1-P[\mathcal{T}])P\left[\max_{j\in S^c}|Z_j(v^*)|>(2-\eta)\lambda\right],\end{aligned}$$

where each $Z_j(v^*)$ is the conditioned version of $\tilde{V}_j$ with the scaling factor M_n in the variance fixed to $v^*=\frac{\lambda^2 q}{2nC_{\max}}$.

Lemma 14 *Under the stated assumptions in Theorem 8, $\frac{\lambda^2}{v^*}\to+\infty$, and there exists some $\gamma>0$ such that $\frac{1}{\lambda}E[\max_{j\in S^c}Z_j(v^*)]>(2-\eta)[1+\gamma]$ for all sufficiently large n.*

Lemma 15 *For any $\xi>0$, we have*

$$P\left[\max_{j\in S^c}Z_j(v^*)<E[\max_{j\in S^c}Z_j(v^*)]-\xi\right]\leq\exp\left(-\frac{\xi^2}{2v^*}\right). \tag{27}$$

Using these two lemmas, we complete the proof as follows. Set $\xi=\frac{(2-\eta)\gamma\lambda}{2}$ in inequality (27) to obtain that

$$P\left[\max_{j\in S^c}Z_j(v^*)\geq(2-\eta)(1+\frac{\gamma}{2})\lambda\right]\geq 1-\exp\left(-\frac{(2-\eta)^2\gamma^2\lambda^2}{8v^*}\right),$$

which converges to 1, since $\frac{\lambda^2}{v^*}\to+\infty$ from Lemma 14. So,

$$P\left[\max_{j\in S^c}|V_j|>\lambda\right]\geq P\left[\max_{j\in S^c}|\tilde{V}_j|>(2-\eta)\lambda\right]\to 1.$$

■

A.8 Proofs of Lemma 10-Lemma 15

Proof of Lemma 10

Proof Y can be decomposed into two parts $Y=aX+e$, where $a=\rho\frac{\sigma_Y}{\sigma_X}$ is a constant, and e is a normal random variable independent of X, with mean 0, and variance $E(e^2)=$

$(1-\rho^2)\sigma_Y^2$.

$$\begin{aligned} E(|X|Y^2) &= E(|X|(a^2X^2+2aXe+e^2)) \\ &= a^2E(|X|^3)+E(|X|)E(e^2) \\ &= \rho^2(\frac{\sigma_Y}{\sigma_X})^2\sigma_X^3c_3+c_1\sigma_X(1-\rho^2)\sigma_Y^2 \\ &= c_3\rho^2\sigma_X\sigma_Y^2+c_1(1-\rho^2)\sigma_X\sigma_Y^2, \end{aligned}$$

where $c_1=E(|Z|), c_3=E(|Z|^3)$; Z is a variable with standard normal distribution.

The same way, X can be decomposed as $X=bY+e_2$, where $b=\rho\frac{\sigma_X}{\sigma_Y}$, $e_2 \perp\!\!\!\perp Y$, and $E(e_2^2)=(1-\rho^2)\sigma_X^2$. Then we have

$$\begin{aligned} E(|X|^2Y^4) &= E((b^2Y^2+2bYe_2+e_2^2)Y^4) \\ &= b^2E(|Y|^6)+E(Y^4)E(e_2^2) \\ &= \rho^2(\frac{\sigma_X}{\sigma_Y})^2(15\sigma_Y^6)+3\sigma_Y^4(1-\rho^2)\sigma_X^2 \\ &= (12\rho^2+3)\sigma_X^2\sigma_Y^4. \end{aligned}$$

■

Proof of Lemma 11

Proof Conditioned on $X(S)$ and ϵ, the only random component in V_j is the column in the column vector X_j, $j\in S^c$. We know that $(X(S^c)|X(S),\epsilon)\sim(X(S^c)|X(S))$ is Gaussian with mean and covariance

$$E[X(S^c)^T|X(S),\epsilon] = \Sigma_{21}(\Sigma_{11})^{-1}X(S)^T, \tag{28}$$

$$var(X(S^c)|X(S)) = \Sigma_{2|1}=\Sigma_{22}-\Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}. \tag{29}$$

Consequently, we have,

$$\begin{aligned} & |E[V|X(S),\epsilon]| \\ = & \left|\Sigma_{21}(\Sigma_{11})^{-1}X(S)^T\left\{X(S)(X(S)^TX(S))^{-1}\lambda\overrightarrow{b}\right.\right. \\ & \left.\left.-\left[X(S)(X(S)^TX(S))^{-1}X(S)^T-I\right]\frac{\epsilon}{n}\right\}\right| \\ = & |\Sigma_{21}(\Sigma_{11})^{-1}\lambda\overrightarrow{b}| \\ \leq & \lambda(1-\delta)\mathbf{1}, \end{aligned}$$

where the last inequality uses condition (11).

Now, we compute the elements of the conditional covariance

$$cov(V_j,V_k|\epsilon,X(S)).$$

Let $\vec{\alpha}=X(S)(X(S)^TX(S))^{-1}\lambda\overrightarrow{b}-\left[X(S)(X(S)^TX(S))^{-1}X(S)^T-I)\right]\frac{\epsilon}{n}$, then $V_j=X_j^T\vec{\alpha}$. So we have

$$cov(V_j,V_k|\epsilon,X(S))=\vec{\alpha}^Tcov(X_j^T,X_k^T|\epsilon,X(S))\vec{\alpha}=[var(X(S^c)|X(S))]_{jk}\,\vec{\alpha}^T\vec{\alpha}.$$

Consequently,

$$cov(V|\epsilon, X(S)) = \vec{\alpha}^T\vec{\alpha}\ var(X(S^c)|X(S)) = \vec{\alpha}^T\vec{\alpha}\Sigma_{2|1} = \vec{\alpha}^T\vec{\alpha}[\Sigma_{22} - \Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}].$$

By careful calculation, we have $\vec{\alpha}^T\vec{\alpha} = M_n$. ■

Proof of Lemma 12

Proof Recall that $M_1 = \lambda \overrightarrow{b}^T(X(S)^TX(S))^{-1}\overrightarrow{b}$. So,

$$\frac{\lambda q}{\Lambda_{\max}(X(S)^TX(S))} \leq M_1 \leq \frac{\lambda q}{\Lambda_{\min}(X(S)^TX(S))}.$$

From (37)

$$P\left[\frac{1}{2}\tilde{C}_{\min} \leq \Lambda_i(\frac{1}{n}X^TX) \leq 2\tilde{C}_{\max}\right] \geq 1 - 2\exp(-0.03n),$$

then we have,

$$P\left[\frac{\lambda q}{2nC_{\max}} \leq M_1 \leq \frac{2\lambda q}{n\tilde{C}_{\max}}\right] \geq 1 - 2\exp(-0.03n).$$

Define $\varrho = E[|Z|]$, where $Z \sim N(0,1)$, then for any random variable $R \sim N(0,\sigma^2)$, $E[|R|] = \sigma\varrho$. Since $x_i^T\beta \sim N(0, \beta_1^T\Sigma_{11}\beta^*(S))$, we have

$$E[|x_i^T\beta|] = \sqrt{\beta_1^T\Sigma_{11}\beta^*(S)}\varrho.$$

We know that $M_2 \leq \frac{1}{n^2}\epsilon^T\epsilon$. Since $E[\epsilon_i^2] = E[E[\epsilon_i^2|X(S)]] = E[\sigma^2|x_i^T\beta|] = \sigma^2\sqrt{\beta^*(S)^T\Sigma_{11}\beta^*(S)}\varrho$, and $E[\epsilon_i^4] = E[E[\epsilon_i^4|X(S)]] = 3E[\sigma^4|x_i^T\beta|^2] = 3\sigma^4\beta^*(S)^T\Sigma_{11}\beta^*(S)$, we have

$$\begin{aligned}
& P\left[\frac{\sum_i \epsilon_i^2}{n^2} \geq \frac{\sigma^2(\varrho + \sqrt{3-\varrho^2})\sqrt{\beta^*(S)^T\Sigma_{11}\beta^*(S)}}{n}\right] \\
= & P\left[\sum_i \epsilon_i^2 - n\sigma^2\varrho\sqrt{\beta^*(S)^T\Sigma_{11}\beta^*(S)} \geq n\sigma^2\sqrt{3-\varrho^2}\sqrt{\beta^*(S)^T\Sigma_{11}\beta^*(S)}\right] \\
\leq & \frac{nvar(\epsilon_i^2)}{n^2\sigma^4(3-\varrho^2)\beta^*(S)^T\Sigma_{11}\beta^*(S)} \\
= & \frac{3\sigma^4\beta^*(S)^T\Sigma_{11}\beta^*(S) - \sigma^4\beta^*(S)^T\Sigma_{11}\beta^*(S)\varrho^2}{n\sigma^4(3-\varrho^2)\beta^*(S)^T\Sigma_{11}\beta^*(S)} \\
= & \frac{1}{n}
\end{aligned}$$

So,

$$P\left[M_2 \geq \frac{\sigma^2(\varrho + \sqrt{3-\varrho^2})\sqrt{\beta^*(S)^T\Sigma_{11}\beta^*(S)}}{n}\right] \leq \frac{1}{n}.$$

While $\sqrt{\beta_1^T \Sigma_{11} \beta^*(S)} \leq \sqrt{\tilde{C}_{\max}} \|\beta\|_2$ and $\varrho = E(|Z|) \leq \sqrt{E(|Z|^2)} = 1$, where Z is a standard normal random variable, so

$$\frac{\sigma^2(\varrho + \sqrt{3 - \varrho^2})\sqrt{\beta^*(S)^T \Sigma_{11} \beta^*(S)}}{n} \leq \frac{3\sigma^2 \sqrt{\tilde{C}_{\max}} \|\beta\|_2}{n}.$$

Then we have

$$P[M_2 \geq \frac{3\sigma^2 \sqrt{\tilde{C}_{\max}} \|\beta\|_2}{n}] \leq \frac{1}{n}.$$

■

Proof of Lemma 13

Proof

By lemma 17, we have for any $t > q$,

$$P\left[\max_{i=1,\ldots n} \|\Sigma_{11}^{-\frac{1}{2}} x_i(S)\|_2^2 \geq 2t\right] \leq n \exp(-t\left[1 - 2\sqrt{\frac{q}{t}}\right]).$$

Take $t = \max(16q, 4\log n)$, we have

$$\begin{array}{rcl} \exp(-t\left[1 - 2\sqrt{\frac{q}{t}}\right]) & \leq & \exp(-t\left[1 - 2\sqrt{\frac{1}{16}}\right]) \\ & = & \exp(-\frac{t}{2}) \\ & \leq & \frac{1}{n^2}, \end{array}$$

so,

$$P\left[\max_{i=1,\ldots n} \|\Sigma_{11}^{-\frac{1}{2}} x_i(S)\|_2^2 \geq 2\max(16q, 4\log n)\right] \leq \frac{1}{n}.$$

Since $\|\Sigma_{11}^{-\frac{1}{2}} x_i(S)\|_2^2 \geq \frac{1}{\tilde{C}_{\max}} \|x_i(S)\|_2^2$, we have

$$P\left[\max_{i=1,\ldots n} \|x_i(S)\|_2^2 \geq 2\tilde{C}_{\max} \max(16q, 4\log n)\right] \leq \frac{1}{n}. \tag{30}$$

■

Proof of Lemma 14

Proof Recall that the Gaussian random vector Z is zero-mean with covariance $v^* \Sigma_{2|1}$, where $v^* = \frac{\lambda^2}{2n\tilde{C}_{\max}}$. For any index i, let e_i be equal to 1 in position i, and zero otherwise. For any two indices $i \neq j$, we have

$$\begin{array}{rcl} \Delta_Z(i,j) & = & E(Z_i - Z_j)^2 \\ & = & v^*(e_i - e_j)^T \Sigma_{2|1}(e_i - e_j) \\ & \leq & 2v^* \lambda_{\max}(\Sigma_{2|1}) \\ & \leq & 2v^* \tilde{C}_{\max} \end{array}$$

Now let $(X_i)_{i\in S^c}$ be an i.i.d. zero-mean Gaussian vector with $var(X_i) = \tilde{C}_{\max}v^*$, so that $\Delta_X(i,j) = E(X_i - X_j)^2 = 2\tilde{C}_{\max}v^*$. If we set

$$\Delta^* = \max_{i,j\in S^c} |\Delta_X(i,j) - \Delta_Z(i,j)|,$$

then, by applying a known error bound for the Sudakov-Fernique inequality (Chatterjee., 2005), we are guaranteed that

$$E(\max_{j\in S^c} Z_j) \geq E(\max_{j\in S^c} X_j) - \sqrt{\Delta^* \log(p-q)}. \tag{31}$$

We now show that the quantity Δ^* is upper bounded by

$$\Delta^* \leq 2v^*(\tilde{C}_{\max} - \tilde{C}_{\min}). \tag{32}$$

Using the inversion formula for block-partitioned matrices, we have

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \left[[\Sigma^{-1}]_{22}\right]^{-1}.$$

Consequently,

$$\begin{aligned} E(Z_i - Z_j)^2 &= v^*(e_i - e_j)\Sigma_{2|1}(e_i - e_j)^T \\ &\geq 2v^*\Lambda_{\min}(\Sigma_{2|1}) \\ &= 2v^*/\Lambda_{\max}([\Sigma^{-1}]_{22}) \\ &\geq 2v^*/\Lambda_{\max}([\Sigma^{-1}]) \\ &= 2v^*\tilde{C}_{\min}. \end{aligned}$$

So,

$$\begin{aligned} \Delta^* &= \max_{i,j\in S^c} |\Delta_X(i,j) - \Delta_Z(i,j)| \\ &= \max_{i,j\in S^c} |2v^*\tilde{C}_{\max} - \Delta_Z(i,j)| \\ &\leq 2v^*(\tilde{C}_{\max} - \tilde{C}_{\min}). \end{aligned}$$

An argument on page 80 of Ledoux and Talagrand (1991) can be used to show the following result which appears in Wainwright (2009): for any $\delta' > 0$, there exists an $N(\delta')$, for all $N > N(\delta')$,

$$E(\max_{j\in S^c} X_j) \geq (1-\delta')\sqrt{2v^*\tilde{C}_{\max}\log N}.$$

Applying this lower bound to the bound (31), we have

$$\begin{aligned} \frac{1}{\lambda}E(\max_{j\in S^c} Z_j) &\geq \frac{1}{\lambda}\left[(1-\delta')\sqrt{2v^*\tilde{C}_{\max}\log N} - \sqrt{\Delta^*\log N}\right] \\ &\geq \frac{1}{\lambda}\left[(1-\delta')\sqrt{2v^*\tilde{C}_{\max}\log N} - \sqrt{2v^*(\tilde{C}_{\max} - \tilde{C}_{\min})\log N}\right] \\ &= \left[(1-\delta')\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}}\right]\sqrt{\frac{2v^*}{\lambda^2}\log N}. \end{aligned}$$

Since $\frac{v^*}{\lambda^2} = \frac{q}{nC_{\max}}$, we now apply the condition

$$\frac{2q\log N}{n} > \frac{1}{\theta_l - v} = \frac{\tilde{C}_{\max}(2-\eta)^2}{(\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}})^2 - vC_{\max}(2-\eta)^2}$$

to obtain that

$$\begin{aligned}
\frac{1}{\lambda}E(\max_{j\in S^c} Z_j) &\geq \left[(1-\delta')\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}}\right]\sqrt{\frac{2v^*}{\lambda^2}\log N} \\
&= \left[(1-\delta')\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}}\right]\sqrt{\frac{2q\log N}{nC_{\max}}} \\
&\geq \frac{(1-\delta')\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}}}{\sqrt{(\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}})^2 - vC_{\max}(2-\eta)^2}}(2-\eta) \qquad (33)
\end{aligned}$$

Let $F(\delta')$ be the lower bound on the RHS (33). Note that F is a continuous function, and moreover that

$$F(0) = \frac{\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}}}{\sqrt{(\sqrt{\tilde{C}_{\max}} - \sqrt{\tilde{C}_{\max} - \tilde{C}_{\min}})^2 - vC_{\max}(2-\eta)^2}}(2-\eta) > (2-\eta).$$

Therefore, by the continuity of $F(\cdot)$ and the arbitrariness of δ', we can choose $\delta' > 0$ sufficiently small to ensure that for some $\gamma > 0$, we have $\frac{1}{\lambda}E(\max_{j\in S^c} Z_j) > (2-\eta)[1+\gamma]$ for all sufficiently large n. ■

Proof of Lemma 15

Proof Consider the function $f : R^N \to N$ given by

$$f(w) = \max_{1\leq j\leq N}[\sqrt{v^*\Sigma_{2|1}}w]_j,$$

then, for a Gaussian random vector $V \sim N(0, I_{N\times N})$, we have $f(V) = \max_{j\in S^c}\tilde{Z}_j$.

We now bound the Lipschitz constant of f. Let $R = \sqrt{\Sigma_{2|1}}$, then for each $w, v \in R^N$,

$$\begin{aligned}
|[\sqrt{v^*}Rw]_j - [\sqrt{v^*}Rv]_j| &= \sqrt{v^*}[\sum_k R_{jk}(w_k - v_k)] \\
&\leq \sqrt{v^*}\sqrt{\sum_k R_{jk}^2}\|w-v\|_2 \\
&\leq \sqrt{v^*}\|w-v\|_2,
\end{aligned}$$

where the last inequality holds since $\sum_k R_{jk}^2 = [\Sigma_{2|1}]_{jj} \leq 1$.

$$\begin{aligned}
|f(w) - f(v)| &= |\max_j[\sqrt{v^*}Rw]_j - \max_j[\sqrt{v^*}Rv]_j| \\
&\leq \max_j|[\sqrt{v^*}Rw]_j - [\sqrt{v^*}Rv]_j|
\end{aligned}$$

Therefore, by Gaussian concentration of measure for Lipschitz functions [Massart. (2003.)], we conclude that for any $\zeta > 0$, it holds that

$$P\left[f(V) > \zeta + E[f(V)]\right] \leq \exp\left(-\frac{\zeta^2}{2v^*}\right), \text{ and}$$
$$P\left[f(V) > E[f(V)] - \zeta\right] \leq \exp\left(-\frac{\zeta^2}{2v^*}\right)$$

■

Appendix B. Some Gaussian Comparison Results

Lemma 16 *For any mean zero Gaussian random vector $(X_1, \ldots, X_n)$, and $t > 0$, we have*

$$P(\max_{1\leq i\leq n} |X_i| \geq t) \leq 2n \exp\{-\frac{t^2}{2\max_i E(X_i^2)}\} \tag{34}$$

Proof Note that the generate function of X_i is

$$E(e^{tX_i}) = \exp\{\frac{E(X_i^2)t^2}{2}\}.$$

So, for any $t > 0$,

$$P(X_i \geq x) = P(e^{tX_i} \geq e^{tx}) \leq \frac{E(e^{tX_i})}{e^{tx}} = \exp\{\frac{E(X_i^2)t^2}{2} - xt\},$$

by taking $t = \frac{x}{E(X_i^2)}$, we have

$$P(X_i \geq x) \leq \exp\{-\frac{x^2}{2E(X_i^2)}\}.$$

So,

$$P(|X_i| \geq t) = 2P(X_i \geq t) \leq 2\exp\{-\frac{t^2}{2E(X_i^2)}\} \leq 2\exp\{-\frac{t^2}{2\max_i E(X_i^2)}\}.$$

So,

$$P(\max_{1\leq i\leq n} |X_i| \geq t) \leq 2n \exp\{-\frac{t^2}{2\max_i E(X_i^2)}\}.$$

■

Appendix C. Large deviation for χ^2 distribution

Lemma 17 *Let $Z_1, \ldots, Z_n$ be i.i.d. χ^2-variates with q degrees of freedom. Then for all $t > q$, we have*

$$P\left[\max_{i=1,\ldots,n} Z_i > 2t\right] \leq n \exp(-t\left[1 - 2\sqrt{\frac{q}{t}}\right]). \tag{35}$$

The proof of this lemma can be found in Obozinski et al. (2008).

Appendix D. Some useful random matrix results

In this appendix, we use some known concentration inequalities for the extreme eigenvalues of Gaussian random matrices (Davidson and Szarek, 2001) to bound the eigenvalues of a Gaussian random matrix. Although these results hold more generally, our interest here is on scalings (n, q) such that $q/n \to 0$.

Lemma 18 (Davidson and Szarek (2001)) *Let $\Gamma \in R^{n\times q}$ be a random matrix whose entries are i.i.d. from $N(0, 1/n)$, $q \le n$. Let the singular values of Γ be $s_1(\Gamma) \ge \ldots \ge s_q(\Gamma)$. Then*

$$\max\left\{P\left[s_1(\Gamma) \ge 1 + \sqrt{\frac{q}{n}} + t\right], P\left[s_q(\Gamma) \le 1 - \sqrt{\frac{q}{n}} - t\right]\right\} < \exp\{-nt^2/2\}.$$

Using Lemma 18, we now have some useful results.

Lemma 19 *Let $U \in R^{n\times q}$ be a random matrix with elements from the standard normal distribution (i.e., $U_{ij} \sim N(0, 1)$, i.i.d.) Assume that $q/n \to 0$. Let the eigenvalues of $\frac{1}{n}U^TU$ be $\Lambda_1(\frac{1}{n}U^TU) \ge \ldots \ge \Lambda_q(\frac{1}{n}U^TU)$. Then there exist a constant c, when n is big enough,*

$$P\left[\frac{1}{2} \le \Lambda_i(\frac{1}{n}U^TU) \le 2\right] \ge 1 - 2\exp(-0.03n). \tag{36}$$

Proof Let $\Gamma = \frac{1}{\sqrt{n}}U$, then $\Lambda_i(\frac{1}{n}U^TU) = s_i^2(\Gamma)$. By Lemma 18,

$$P\left[s_q(\Gamma) \le 1 - \sqrt{\frac{q}{n}} - t\right] < \exp\{-nt^2/2\},$$

by taking $t = t_0 = 1 - \frac{\sqrt{2}}{2} - 0.1$, we have

$$P\left[s_q(\Gamma) \le \frac{\sqrt{2}}{2} + 0.1 - \sqrt{\frac{q}{n}}\right] < \exp\{-nt_0^2/2\}.$$

Since $q/n \to 0$ by assumption, we have when n is big enough, $q/n < 0.1$, then

$$P\left[s_q(\Gamma) < \frac{\sqrt{2}}{2}\right] < \exp\{-nt_0^2/2\},$$

which implies that, for any $i = 1, \ldots, q$,

$$P\left[\Lambda_i(\frac{1}{n}(U^TU)) < \frac{1}{2}\right] < \exp\{-nt_0^2/2\}.$$

Followed the same procedures,

$$P\left[\Lambda_i(\frac{1}{n}(U^TU)) > 2\right] < \exp\{-nt_1^2/2\},$$

for $t_1 = \sqrt{2} - 1.1$. Then inequality (36) holds immediately. ■

Corollary 20 *Let $X \in R^{n\times q}$ be a random matrix, of which, the rows are i.i.d. from the normal distribution with mean 0 and covariance Σ. Assume that $\tilde{C}_{\min} \leq \Lambda_i(\Sigma) \leq \tilde{C}_{\max}$ and $q/n \to 0$, then there exist a constant c, when n is big enough,*

$$P\left[\frac{1}{2}\tilde{C}_{\min} \leq \Lambda_i(\frac{1}{n}X^TX) \leq 2\tilde{C}_{\max}\right] \geq 1 - 2\exp(-0.03n). \tag{37}$$

Proof Let $U = X\Sigma^{-\frac{1}{2}}$, then $var(U) = I_{q\times q}$ and U satisfies the condition in Lemma 19. Then

$$P\left[\frac{1}{2} \leq \Lambda_i(\frac{1}{n}U^TU) \leq 2\right] \geq 1 - 2\exp(-0.03n),$$

for some constant c. Since

$$\tilde{C}_{\min}\Lambda_i(\frac{1}{n}U^TU) \leq \Lambda_i(\frac{1}{n}X^TX) \leq \tilde{C}_{\max}\Lambda_i(\frac{1}{n}U^TU),$$

result (37) is obtained immediately. ■

References

E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Info Theory,*, 51(12): 4203 – 4215, 2005.

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

S. Chatterjee. An error bound in the sudakov-fernique inequality. *Technical report, UC Berkeley. arXiv:math.PR/0510424*, 2005.

S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *J. Sci. Computing*, 20(1):33–61, 1998.

K. R. Davidson and S. J. Szarek. *Local operator theory, random matrices, and Banach spaces. In Handbook of Banach Spaces, volume 1, pages 317-366.* Elsevier, Amsterdan, NL., 2001.

Donoho. For most large undetermined system of linear equations the minimal l1-norm near-solution is also the sparsest solution. *Technical report, Statistics Department, Stanford University*, 2004.

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

D. Donoho, M. Elad, and V. M. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info Theory*, 52(1):6–18, 2006.

J. Romberg E. Candes and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Technical report, Applied and Computational Mathematics, Caltech*, 2004.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, pages 407–451, 2004.

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

M. Elad and A. M. Bruckstein. On sparse representation in pairs of bases. *IEEE Trans. Info Theory,*, 49(6):1579 – 1581, 2003.

J. Fessler. Statistical image reconstruction methods for transmission tomography. *Handbook of Medical Imaging*, 2:1–70, 2000.

D. Freedman. *Statistical models: Theory and practice*. Cambridge University Press, 2005.

J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

K. Knight and W. J. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics,*, 28: 1356 – 1378, 2000.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* New York, Springer-Verlag, 1991.

M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing MRI. *IEEE Signal Processing Magazine*, 25(2):72–82, 2008.

P. Massart. *Concentration Inequalties and Model Selection.* Ecole d'Eté de Probabilités,Saint- Flour. Springer, New York, 2003.

N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

G. Obozinski, M.J. Wainwright, M.I. Jordan, et al. Union support recovery in high-dimensional multivariate regression. *stat*, 1050:5, 2008.

M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–37, 2000.

S. Rosset. Tracking curved regularized optimization solution paths. *NIPS*, 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

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

J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Info Theory*, 52(3):1030 – 1051, 2006.

Y. Vardi, LA Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985.

M. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *IEEE Transactions on Information Theory*, To appear, 2009.

P. Zhao and B. Yu. Stagewise lasso. *The Journal of Machine Learning Research*, 8:2701–2726, 2007.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.