Statistica Sinica **23** (2013), 99-118
doi:http://dx.doi.org/10.5705/ss.2010.254

THE LASSO UNDER POISSON-LIKE HETEROSCEDASTICITY

Jinzhu Jia, Karl Rohe and Bin Yu

Peking University, University of Wisconsin-Madison and University of California, Berkeley

Abstract: The performance of the Lasso is well understood under the assumptions of the standard sparse linear model with homoscedastic noise. However, in several applications, the standard model does not describe the important features of the data. This paper examines how the Lasso performs on a non-standard model that is motivated by medical imaging applications. In these applications, the variance of the noise scales linearly with the expectation of the observation. Like all heteroscedastic models, the noise terms in this Poisson-like model are *not* independent of the design matrix. Under a sparse Poisson-like model for the high-dimension regime that allows the number of predictors (p) $\gg$ sample size (n), we give necessary and sufficient conditions for the sign consistency of the Lasso estimate. Simulations reveal that the Lasso performs equally well in terms of model selection performance on both Poisson-like data and homoscedastic data (with properly scaled noise variance), across a range of parameterizations.

Key words and phrases: Heteroscedasticity, Lasso, Poisson-like model, sign consistency.

1. Introduction

The Lasso (Tibshirani (1996)) is widely used in high dimensional regression for variable selection. Its model selection performance has been well studied under a standard sparse and homoskedastic regression model. Several researchers have shown that, under sparsity and regularity conditions, the Lasso can select the true model asymptotically even when $p \gg n$ (Donoho, Elad, and Temlyakov (2006); Meinshausen and Buhlmann (2006); Tropp (2006); Zhao and Yu (2006); Wainwright (2009)).

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

$$Y_i = x_i^T \beta^* + \epsilon_i, \tag{1.1}$$

where x_i^T is a row vector representing the predictors for the ith observation, Y_i is the corresponding ith response variable, the ϵ_i's are independent and mean zero

noise terms, and $\beta^* \in \mathbb{R}^p$. If $\mathbf{X} \in \mathbb{R}^{n\times p}$ denotes 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 $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T \in \mathbb{R}^n$. The Lasso estimate (Tibshirani (1996)) is defined as the solution to a penalized least squares problem (with regularization parameter λ):

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{2n}\|Y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1, \tag{1.2}$$

where for some vector $x \in \mathbb{R}^k$, $\|x\|_r = (\sum_{i=1}^k |x_i|^r)^{1/r}$.

In previous research on the Lasso, 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.*

Candes and Tao (2007) suggested that compressed sensing, a sparse method similar to the Lasso, could reduce the number of measurements needed by such medical technology as Magnetic Resonance Imaging (MRI). This methodology was later applied to MRI by Lustig et al. (2008). The standard model was useful to their analyses, but it is not appropriate for such other imaging methods as PET and SPECT (Fessler (2000)).

PET provides an indirect measure for the metabolic activity of a specific tissue. To take an image, a biochemical metabolite must be identified that is attractive to the tissue under investigation. This biochemical metabolite is labeled with a positron emitting radioactive material and is then injected into the subject. The substance circulates through the subject, emitting positrons; when the tissue gathers the metabolite, the radioactive material concentrates around the tissue.

The positron emissions can be modeled by a Poisson point process in three dimensions with an intensity rate proportional to the varying concentrations of the biochemical metabolite. Therefore, an estimate of the intensity rate is an estimate of the level of biochemcial metabolite. However, the positron emissions are not directly observed. After each positron is emitted it very quickly annihilates a nearby electron, sending two X-ray photons in nearly opposite directions (at the speed of light) Vardi, Shepp, and Kaufman (1985). These X-rays are observed by several sensors in a ring surrounding the subject.

A physical model of this system informs the estimation of the intensity level of the Poisson process from the observed data. It can be expressed as a Poisson model where the sample size n represents the number of sensors, Y is a vector of observed values, 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, and p is the number of voxels wanted, the more voxels, the finer the resolution of the final image.

Because the positron emissions are modeled by a Poisson point process, the variance of each observed value Y_i is equal to the expected value $E(Y_i)$. Motivated by the Poissonian model, this paper studies the Lasso under the sparse Poisson-like model

$$\begin{aligned} Y &= \mathbf{X}\beta^* + \epsilon, \\ E(\epsilon \mid \mathbf{X}) &= 0, \\ Cov(\epsilon \mid \mathbf{X}) &= \sigma^2 \times diag(|\mathbf{X}\beta^*|), \\ \epsilon &\perp\!\!\!\perp X(S^c) \mid X(S), \end{aligned} \tag{1.3}$$

where $\sigma^2 > 0$ and the sparsity index set is defined as $S = \{1 \leq j \leq p : \beta_j \neq 0\}$, with the cardinality $q = \#S$ such that $0 < q < p$. 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 $\mathbf{X}$, is $\sigma^2 \times diag(|\mathbf{X}\beta^*|)$, an $n \times n$ diagonal matrix with the vector $\sigma^2 \times |\mathbf{X}\beta^*|$ down the diagonal, and $X(S)$ and $X(S^c)$ denote two matrices consisting of the relevant column vectors (with nonzero coefficients) and irrelevant column vectors (with zero coefficients), respectively. This is a heteroscedastic model.

We do not develop a penalized maximum likelihood estimator for the Poisson-like model, our main interest is to study how the Lasso performs under a departure from the standard linear model. Such results are useful when the "true" model is unknown. The Poisson-like model's likelihood function is non-convex and this presents computational challenges. Before tackling these, we believe it is advantageous to first understand the performance of the computationally tractable Lasso. We do carry out some simulation studies to compare the pure Lasso and the penalized maximum likelihood method in Example 3. To tackle the non-convex problem of penalized maximum likelihood method, we put the true variance of response in the likelihood function. Still, we find that the pure Lasso outperforms the penalized maximum likelihood method for our simulated data in terms of sign consistency.

Since the Lasso provides a computationally feasible way to select a model (Osborne, Presnell, and Turlach (2000); Efron et al. (2004); Rosset (2004); Zhao and Yu (2007)), it can be applied in non-standard settings to give sparse solutions. In this paper we show that the Lasso is robust to the heteroscedastic noise of the

sparse Poisson-like model. Under the Poisson-like model, for general scalings of p, q, n, and β^*, this paper investigates when the Lasso is sign consistent, and when it is not, with theoretical and simulation studies. Our results are comparable to the results for the standard model: when a measure of the signal to noise ratio is large, the Lasso is sign consistent.

1.1. Overview of previous work

The Lasso (Tibshirani (1996)) is well established as a popular technique to simultaneously select a model and provide regularized estimated coefficients. Under the standard homoscedastic linear model, we give a brief overview of this literature.

In noiseless setting, $\epsilon = 0$, with contributions from a broad range of researchers (Chen, Donoho, and Saunders (1998); Donoho and Huo (2001); Elad and Bruckstein (2002); Feuer and Nemirovski (2003); Tropp (2004); Candes and Tao (2006)), there is now an 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 $\mathbf{X}$.

There is also a substantial body of work focusing on the noisy setting. Knight and Fu (2000) analyze the asymptotic behavior of the optimal solution for fixed dimension (p) for L_r regularization with $r \in (0, 2]$. Both Tropp (2006) and Donoho, Elad, and Temlyakov (2006) provide sufficient conditions for the support of the optimal solution to (1.2) to be contained within the support of β^*. Meinshausen and Buhlmann (2006) focuse on model selection for Gaussian graphical models; Zhao and Yu (2006) consider linear regression and more general noise distributions. For the case of Gaussian noise and Gaussian predictors, these papers establish 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) used a particular mutual incoherence condition, the 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 or sub-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. The conditions on the Lasso can be stringent in high dimensions. If the correlations between variables are small, the Irrepresentable Condition usually holds and not otherwise. See also Fan and Lv (2010).

1.2. Our work

Some definitions are needed. Let

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

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

This paper studies the sign consistency of the Lasso applied to data from the sparse Poisson-like model, and also gives 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, and the sign consistency results follow. We also give necessary conditions for the Lasso to be sign consistent under the sparse Poisson-like model. It is shown that the Irrepresentable Condition is necessary for the Lasso's sign consistency under this model; it is also necessary under the standard model (Zhao and Yu (2006); Zou (2006); Wainwright (2009)).

The sufficient conditions for sign consistency for the deterministic and random Gaussian designs require that the variance of the noise be not too large and that the smallest nonzero element of $|\beta^*|$ be not too small. Write the smallest nonzero element of $|\beta^*|$ as

$$M(\beta^*) = \min_{j \in S} |\beta_j^*|.$$

For a 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 a 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. An essential quantity for determining the probability of sign recovery is an (unconventional) signal to noise ratio

$$uSNR = \frac{n[M(\beta^*)]^2}{\sigma^2 \|\beta^*\|_2}. \tag{1.4}$$

The numerator corresponds to the signal strength for sign recovery. The most difficult sign to estimate in β^* is the element that corresponds to $M(\beta^*)$. When the smallest element is larger, estimating the signs is easier, and the signal is more powerful. For the noise term in the denominator, $\|\beta^*\|_2$ is fundamental in the scaling of the noise. The typical definition of SNR is

$$SNR = \frac{\beta^T X' X \beta}{\sum_{i=1}^n var(\epsilon_i)}.$$

Roughly speaking, increasing some $\beta_j's$ makes the signals stronger, so SNR increases, but $uSNR$ might decrease because the denominator increases while the numerator stays the same.

When $uSNR$ is large, the Lasso is sign consistent. Specifically, the sufficient condition for a deterministic design requires that

$$uSNR = \Omega\left(q \log(p+1) \ \max_i \|x_i(S)\|_2\right),$$

where $a_n = \Omega(b_n)$ means that $a_n/b_n \to \infty$. The sufficient conditions for a random Gaussian design requires that

$$uSNR \geq 8 \log n \frac{\sqrt{2\tilde{C}_{\max} \max(4q, \log n)}}{\tilde{C}_{\min}},$$

and

$$uSNR = \Omega\left(q \log(p-q+1)\right).$$

These conditions all require that the (unconventional) signal to noise ratio be large. Thus the Poisson-like model and the standard model require that the variance of the noise be small compared to the size of the signal. Simulations in Section 4 support this.

The remainder of the paper is organized as follows. Section 2 analyzes the Lasso estimator when the design matrix is deterministic. Section 3 considers the case in which the rows of $\mathbf{X}$ are i.i.d. Gaussian vectors. Both sections give necessary and sufficient conditions for the Lasso to be sign consistent. In Section 4, simulations demonstrate the fundamental role of the $uSNR$ and show that the Lasso performs similarly on both homoscedastic and Poisson-like data. Section 5 gives some concluding thoughts. Proofs are presented in the supplementary materials.

2. Deterministic Design

This section examines sign consistency of the Lasso under the sparse Poisson-like model for a nonrandom design matrix $\mathbf{X}$. Let $x_i(S) = e_i^T X(S)$, 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 q. Take $\beta^*(S) = (\beta_j^*)_{j\in S}$ and $\overrightarrow{b} = \text{sign}(\beta^*(S))$, and suppose the Irrepresentable Condition holds: for some constant $\eta \in (0, 1]$,

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \leq 1 - \eta. \tag{2.1}$$

In addition, assume that

$$\Lambda_{\min} \left(\frac{1}{n} X(S)^T X(S) \right) \geq C_{\min} > 0, \tag{2.2}$$

where $\Lambda_{\min}$ denotes the minimal eigenvalue and $C_{\min}$ is some positive constant. Condition (2.2) 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. Let

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

Theorem 1. *Suppose that* $(\mathbf{X}, Y)$ *follows the sparse Poisson-like model at* (1.3) *with each column of* $\mathbf{X}$ *normalized to* l_2*-norm* $\sqrt{n}$*. Assume that* (2.1) *and* (2.2) *hold. If* λ *satisfies* $M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda)$*, then with probability greater than*

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

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

Theorem 1 can be thought as a straightforward result from Theorem 1 in Wainwright (2009). In Wainwright (2009), sign consistency of the Lasso estimate is given for a standard model with sub-Gaussian noise with parameter σ^2. In the Poisson-like model, since $var(\epsilon_i|x_i) = \sigma^2|x_i^T\beta^*| \leq \sigma^2 \max_i \|x_i(S)\|_2 \|\beta^*\|_2$, the noise can be thought of as sub-Gaussian variables with parameter $\sigma^2 \max_i \|x_i(S)\|_2 \|\beta^*\|_2$. A proof of Theorem 1 is in the supplementary materials.

Theorem 1 gives a non-asymptotic result on the Lasso's sparsity pattern recovery property. The next corollary specifies a sequence of λ's that can asymptotically recover the true sparsity pattern. The essential requirements are that

$$\frac{n\lambda^2}{\max_i \|x_i(S)\|_2 \|\beta^*\|_2 \log(p+1)} \to \infty \quad \text{and} \quad M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda).$$

Take

$$\Gamma(\mathbf{X}, \beta^*, \sigma^2) = \frac{\eta^2 \ uSNR}{8 \max_i \|x_i(S)\|_2 (\eta \ C_{\min}^{-1/2} + \sqrt{q} \ C_{\min}^{-1})^2 \log(p+1)}.$$

Corollary 1. *Suppose that* $(\mathbf{X}, Y)$ *follows the sparse Poisson-like model at* (1.3) *with each column of* $\mathbf{X}$ *normalized to* l_2*-norm* $\sqrt{n}$*. Assume that* (2.1) *and* (2.2) *hold. Take* λ *such that*

$$\lambda = \frac{M(\beta^*)}{2\left(\eta\ C_{\min}^{-1/2} + \sqrt{q}\ C_{\min}^{-1}\right)}. \tag{2.3}$$

Then $\hat{\beta}(\lambda) =_s \beta^*$ *with probability greater than*

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

If $\Gamma(\mathbf{X}, \beta^*, \sigma^2) \to \infty$*, then* $P[\hat{\beta}(\lambda) =_s \beta^*]$ *converges to one.*

A proof of Corollary 1 is in the supplementary materials.

The corollary gives a class of heteroscedastic models for which the Lasso gives a sign consistent estimate of β^*. This class requires that $\Gamma(\mathbf{X}, \beta^*, \sigma^2) \to \infty$ which means that

$$uSNR = \frac{n[M(\beta^*)]^2}{\sigma^2\|\beta^*\|_2} = \Omega\left(q\log(p+1)\max_i \|x_i(S)\|_2\right), \tag{2.4}$$

or that $uSNR$ grows fast enough.

The next corollary addresses the classical settings, where p, q and β^* are all fixed and n goes to infinity. This is a straightforward result from Corollary 1. Since $M(\beta^*)$ and $\|\beta^*\|_2$ do not change with n, $\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha) \to \infty$ in Corollary 1 when $(1/n)\max_{1\le i\le n}\|x_i(S)\|_2 \to 0$.

Corollary 2. *Suppose that* $(\mathbf{X}, Y)$ *follows the sparse Poisson-like model* (1.3) *with each column of* $\mathbf{X}$ *normalized to* l_2*-norm* $\sqrt{n}$*. Assume that* (2.1) *and* (2.2) *hold. In the classical case when* p, q *and* β^* *are fixed, if*

$$\frac{1}{n}\max_{1\le i\le n}\|x_i(S)\|_2 \to 0, \tag{2.5}$$

then with λ *at* (2.3)*, as* $n \to \infty$*,* $P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \to 1$.

Condition (2.5) is not a strong one. Suppose

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

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

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

$$\frac{1}{n}\max_{1\le i\le n}\|x_i(S)\|_2 = \frac{1}{\sqrt{n}}\max_{1\le i\le n}\left\|\frac{1}{\sqrt{n}}x_i(S)\right\|_2 \le \frac{1}{\sqrt{n}}C_{\max}^{1/2} \to 0.$$

Theorem 2 (Necessary Conditions)**.** *Suppose that* $(\mathbf{X}, Y)$ *follows the sparse Poisson-like model at* (1.3)*. Assume that* (2.2) *holds.*

(a) *Suppose* $(1/n)X(S)^T X(S) = I_{q\times q}$*. For any* j*, let*

$$c_{n,j}^2 = \frac{n^2 \beta_j^{*2}}{\sigma^2 e_j^T \left[X(S)^T diag(|X\beta^*|) X(S)\right] e_j} \tag{2.6}$$

and $c_n = \min_j c_{n,j}$*. Then*

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

(b) *If*

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S)\right)^{-1} \overrightarrow{b} \right\|_\infty = 1 + \zeta > 1, \tag{2.7}$$

$$\text{then } P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \le 1/2.$$

A proof of Theorem 2 is in the supplementary materials.

Statement (a) holds for the homoscedastic model by removing $diag(|\mathbf{X}\beta^*|)$ from the denominator of (2.6).

From Statement (b), the Irrepresentable Condition (2.1) is necessary for the Lasso's sign consistency. This is known from both Zhao and Yu (2006) and Wainwright (2009). Zhao and Yu (2006) point out that the Irrepresentable Condition is almost necessary and sufficient for the Lasso to be sign consistent under the standard homosedastic model when p and q are fixed. Wainwright (2009) has it as necessary under the standard model for any p and q.

3. Gaussian Random Design

Suppose the rows of $\mathbf{X}$ are i.i.d. from a p-dimensional multivariate Gaussian distribution with mean 0 and variance-covariance matrix Σ. Take the variance-covariance matrix of the relevant predictors and the covariance between the irrelevant predictors and the relevant predictors to be, respectively,

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

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, the following regularity conditions are needed. First, for some positive constants $\tilde{C}_{\min}$ and $\tilde{C}_{\max}$ that do not depend on n,

$$\Lambda_{\min}(\Sigma_{11}) \ge \tilde{C}_{\min} > 0 \quad \text{and} \quad \Lambda_{\max}(\Sigma) \le \tilde{C}_{\max} < \infty; \tag{3.1}$$

second, the Irrepresentable Condition

$$\|\Sigma_{21}(\Sigma_{11})^{-1}\overrightarrow{b}\|_\infty \leq 1-\eta, \tag{3.2}$$

for some constant $\eta \in (0,1]$. Assumptions (3.1) and (3.2) are standard assumptions. Define

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

$$A(n,\beta^*,\sigma^2) = \sqrt{\frac{4\sigma^2\|\beta^*\|_2 \log n\sqrt{2\tilde{C}_{\max}\max(16q,4\log n)}}{n\tilde{C}_{\min}}}, \text{ and}$$

$$\tilde{\Psi}(n,\beta^*,\lambda,\sigma^2) = A(n,\beta^*,\sigma^2) + \frac{2\lambda\sqrt{q}}{\tilde{C}_{\min}}.$$

Theorem 3. *Consider the sparse Poisson-like model at* (1.3) *under Gaussian random design. Suppose that the variance-covariance matrix* Σ *satisfies conditions* (3.1) *and* (3.2). *Further, suppose that* $q/n \to 0$. *Then for any* λ *such that* $M(\beta^*) > \tilde{\Psi}(n,\beta^*,\lambda,\sigma^2)$, *when* n *is big enough* $\hat{\beta}(\lambda) =_s \beta^*$ *holds with probability greater than*

$$1-2\exp\left\{-\frac{\lambda^2\eta^2}{2V^*(n,\beta^*,\lambda,\sigma^2)\tilde{C}_{\max}} + \log(p-q)\right\} - (2q+3)\exp\{-0.01n\} - \frac{1+3q}{n}.$$

A proof of Theorem 3 is in the supplementary materials.

Theorem 3 gives a non-asymptotic result on the Lasso's sparsity pattern recovery property. The next corollary specifies a sequence of λ's that asymptotically recovers the true sparsity pattern on a well behaved class of models. This class of models restricts the relationship between the data ($\mathbf{X}$), the coefficients (β^*), and the distribution of the noise (ϵ). Thus, λ should be chosen such that

$$\frac{\lambda^2\eta^2}{2V^*(n,\beta^*,\lambda,\sigma^2)\tilde{C}_{\max}} - \log(p-q) \to \infty \text{ and } M(\beta^*) > \tilde{\Psi}(n,\beta^*,\lambda,\sigma^2).$$

The results for deterministic design are similar to the results for Gaussian random design. Conditions (3.1) and (3.2) for a random design case can be viewed as the population version of those for a deterministic design.

Take

$$\tilde{\Gamma}(n,\beta^*,\sigma^2) =$$
$$n\eta^2\left[4q\log(p-q+1)\frac{\tilde{C}_{\max}}{\tilde{C}_{\min}} + \frac{96\sigma^2 q\|\beta^*\|_2\log(p-q+1)\sqrt{\tilde{C}^3_{\max}}}{[M(\beta^*)-A(n,\beta^*,\sigma^2)]^2\tilde{C}^2_{\min}}\right]^{-1}.$$

Corollary 3. *Consider the sparse Poisson-like model at* (1.3) *under Gaussian random design. Suppose* Σ *satisfies conditions* (3.1) *and* (3.2)*, and suppose* $M(\beta^*) > A(n, \beta^*, \sigma^2)$ *and* $q/n \to 0$*. If*

$$\lambda = \frac{[M(\beta^*) - A(n, \beta^*, \sigma^2)]\tilde{C}_{\min}}{4\sqrt{q}},$$

then when n *is big enough* $\hat{\beta}(\lambda) =_s \beta^*$ *holds with probability greater than*

$$1 - 2\exp\left\{-\log(p-q+1)[\tilde{\Gamma}(n, \beta^*, \sigma^2) - 1]\right\} - (2q+3)\exp\{-0.01n\} - \frac{1+3q}{n}.$$

If

$$\frac{n}{q\log(p-q+1)} \to \infty \quad \textit{and} \quad \frac{n[M(\beta^*) - A(n, \beta^*, \sigma^2)]^2}{\sigma^2\|\beta^*\|_2 \; q\log(p-q+1)} \to \infty, \tag{3.3}$$

then $P[\hat{\beta}(\lambda) =_s \beta^*]$ *converges to one.*

A proof of Corollary 3 is in the supplementary materials.

The condition that $M(\beta^*) \geq A(n, \beta^*, \sigma^2)$ is equivalent to

$$uSNR \geq 8\tilde{C}_{\min}^{-1} \log n \sqrt{2\tilde{C}_{\max} \max(4q, \log n)}, \tag{3.4}$$

and the conditions at (3.3) imply that

$$uSNR = \Omega\left(q\log(p-q+1)\right). \tag{3.5}$$

Thus when $uSNR$ is large, the Lasso can identify the sign of the true predictors.

Theorem 4 (Necessary Conditions)**.** *Consider the sparse Poisson-like model at* (1.3) *under Gaussian random design, and suppose the variance-covariance matrix* Σ *satisfies* (3.1)*. If*

$$\|\Sigma_{21}(\Sigma_{11})^{-1}\overrightarrow{b}\|_\infty = 1 + \zeta > 1, \tag{3.6}$$

$$\textit{then } P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \leq \tfrac{1}{2};$$

A proof of Corollary 4 is in the supplementary materials.

In the next section, simulations are used to directly compare the performance of the Lasso between the Poisson-like model and the standard homoscedastic model.

4. Simulation Studies

Our first example investigates a peculiarity of the $uSNR$ at (1.4): functions of β^* appear in both the signal and in the noise. The second example compares the model selection performance of the Lasso under the standard sparse linear model to the model selection performance of the Lasso under the sparse Poisson-like model. The third example compares the performance of the Lasso against the performance of penalized maximum likelihood in the sparse Poisson-like model. In the first and third example, all data were generated from the sparse Poisson-like model; In the second, the performance of the Lasso is compared between homoscedastic noise and Poisson-like noise. The parameterizations of the standard homoscedastic models differ only in that the noise terms are homoscedastic. To ensure a fair comparison, the variance of the noise terms in the standard model is set equal to the average variance of the noise terms in the corresponding Poisson-like model.

All simulations were done in R with the LARS package (Efron et al. (2004)).

Example 1. Consider an initial model with $n = 400$, $p = 1,000$, $q = 20$, $\sigma^2 = 1$, and each element of the design matrix $\mathbf{X}$ drawn independently from $N(0, 1)$. Once $\mathbf{X}$ was drawn, it was fixed through all of the simulations. It was also used in Example 2. We took

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

The first simulation design had $M(\beta^*) = \beta_{\min} = 5$ and changed the value of $\beta_{\max}$; the second simulation design fixed $\|\beta\|_2$ and changed the value of $M(\beta^*)$. One model was present in both designs: $\beta_{\max} = 40$ and $\beta_{\min} = 5$. Here $\|\beta^*\|_2 = 127$ and $uSNR = 400 \times 5^2/127 \approx 78$.

The first simulation design had ten different parameterizations: $\beta_{\min} = 5$ and $\beta_{\max} \in \{100, 90, 80, 70, 60, 50, 40, 30, 20, 10\}$; the second simulation design had ten different parameterizations, each fixing $\|\beta^*\|_2$ and altering $\beta_{\min}$ such that $uSNR$ did not change from the first simulation design (to keep $\|\beta^*\|_2$ fixed, $\beta_{\max}$ had to change accordingly). The values of the parameters for the two designs are in Tables 1 and 2. We also list conventional SNR in these two tables, from which we see that there is almost no relation between SNR and $uSNR$.

For each simulation design, the Monte Carlo estimate for the probability of correctly estimating the signs is plotted against $uSNR$ in Figure 1: each point along the solid line in Figure 1 corresponds to the first simulation design and each point along the dashed line corresponds to the second. Success is defined as

Table 1. The first simulation. Numbers in the table are rounded to the nearest integer.

$\beta_{\max}$	100	90	80	70	60	50	40	30	20	10
$\|\beta^*\|_2$	317	285	253	222	190	159	127	96	65	35
$uSNR$	32	35	39	45	53	63	78	104	153	283
SNR	375	338	300	262	225	188	151	114	78	43

Table 2. The second simulation. Numbers in the table are rounded.

$\beta_{\min} = M(\beta^*)$	3.2	3.3	3.6	3.8	4.1	4.5	5.0	5.8	7.0	9.5
$\beta_{\max}$	40	40	40	40	40	40	40	40	40	39
$uSNR$	32	35	39	45	53	63	78	104	153	283
SNR	151	151	151	151	151	151	151	151	152	152

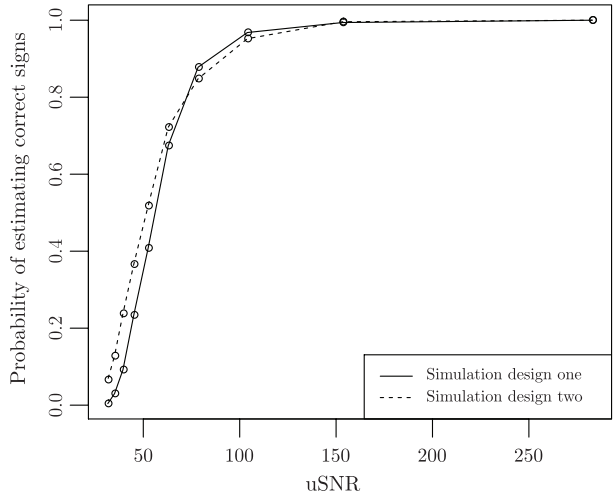


Figure 1. Probability of Success vs. $uSNR$ in Example 1. Each probability was estimated with 500 simulations.

the existence of a λ that makes $\hat{\beta}(\lambda) =_s \beta^*$. The probability of success for each point was estimated with 500 trials.

Figure shows that as $uSNR$ increases, the probability of success increases. What is remarkable is the similarity between the solid and dashed lines. This simulation demonstrates that increasing the elements of β^* can have either a positive or a negative effect on the probability of successfully estimating the signs, and shows that these effects are well characterized by $uSNR$.

Example 2. The only difference from Example 1 is that here the noise terms are homoscedastic. To ensure a fair comparison, the variance of the noise was always set equal to the average variance of the noise terms in the corresponding Poisson-like model.

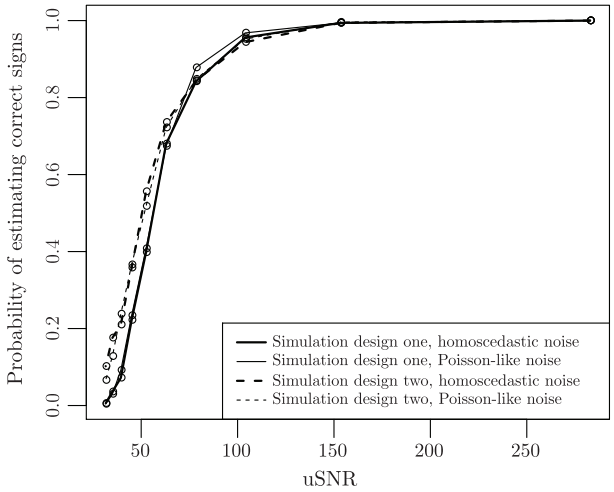
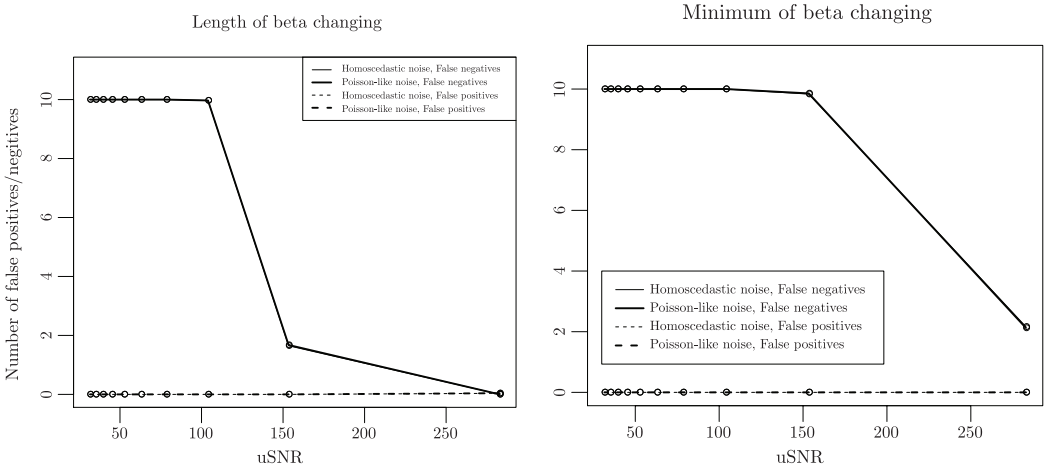


Figure 2. The performance of the Lasso on homoscedastic data and on Poisson-like data.



(a) Design 1: $\|\beta^*\|_2$ shrinking

(b) Design 2: β_{min} growing

Figure 3. The relationship between $uSNR$ and the number of false negatives / false positives.

In Figure 2, $uSNR$ is plotted against the probability of success. As in Example 1, success is defined as the existence of a λ which makes $\hat{\beta}(\lambda) =_s \beta^*$, and the probability of success for each point was estimated with 500 trials. In the figure, the dashed lines are nearly indistinguishable from the solid lines, suggesting that the Lasso is robust to one type of heteroscedastic noise.

To further understand the estimation errors, Figure 3 plots the number of

false positives,

$$FP = \#\{j : \hat{\beta}_j \neq 0 \text{ but } \beta_j = 0\},$$

and the number of false negatives,

$$FN = \#\{j : \hat{\beta}_j = 0 \text{ but } \beta_j \neq 0\},$$

for the Lasso with both homoscedastic data and Poisson-like data. Figure 3(a) follows the first simulation design, and Figure 3(b) follows the second. In each of the simulations, the tuning parameter λ was chosen to minimize the cross-validated mean squared error prediction accuracy with the function cv.lars() in the LARS package (Efron et al. (2004)).

In Figures 3(a) and 3(b) the dashed lines are indistinguishable from each other, suggesting that the number of false positives does not depend on the noise model. Additionally, the solid lines are indistinguishable.

Example 3. When the noise terms are normally distributed with equal variance, the Lasso is the penalized maximum likelihood estimator; this is not the case in the sparse Poisson-like model. This example demonstrates that, in terms of model selection consistency, the Lasso appears to outperform the penalized maximum likelihood estimator for the sparse Poisson-like model. We also compare the prediction error of the methods. Results show that the penalized MLE has a smaller prediction error when the weights are well chosen.

The likelihood function for the Poisson-like model is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2|x_i^T\beta|}} \exp\left\{-\frac{(y_i - x_i\beta)^2}{\sigma^2|x_i^T\beta|}\right\}.$$

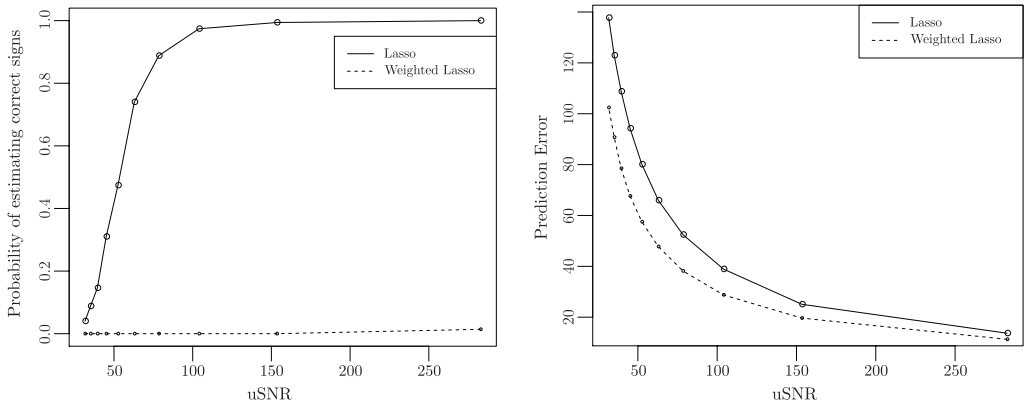
So the ℓ_1 penalized maximum likelihood is

$$\operatorname{argmax}_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} \left[-\frac{(y_i - x_i\beta)^2}{\sigma^2|x_i^T\beta|} - \frac{1}{2}\sum_{i=1}^{n} \log[\sigma^2|x_i^T\beta|]\right] - \lambda\|\beta\|_1. \tag{4.1}$$

The objective function in (4.1) is not convex. For this simulation, given the true parameters, we replaced $\sigma^2|x_i^T\beta|$ with the known variance of the noise $\sigma^2|x_i^T\beta^*|$, and solved the (convex) weighted Lasso problem

$$\arg\min_{\beta} \frac{1}{2n}\|W(Y - X\beta)\|_2^2 + \lambda\|\beta\|_1, \tag{4.2}$$

where W is a diagonal matrix with $W_{ii} = (\sigma^2|x_i^T\beta^*|)^{-1/2}$. This example shows the standard Lasso outperformed the weighted Lasso (4.2), suggesting that for the



(a) Probability of success (b) Prediction Error

Figure 4. Comparisons of the standard Lasso and the Penalized MLE.

Poisson-like data, the standard Lasso outperforms penalized maximum likelihood when considering variable selection.

In the simulation, we set $n = 400$, $p = 1,000$, $q = 20$, $\sigma^2 = 1$, and

$$\beta^* = [\underbrace{5, \ldots, 5}_{10}, \underbrace{\beta_{\max}, \ldots, \beta_{\max}}_{10}, \underbrace{0, \ldots, 0}_{980}]',$$

where $\beta_{\max}$ took values in $\{1, 2, \ldots, 10\} \times 10$. Each element of the design matrix $\mathbf{X}$ was drawn independently from $N(0, 1)$. Once $\mathbf{X}$ was drawn, it was fixed through all of the simulations. Then, both the standard Lasso and ℓ_1 penalized likelihood were fit to each pair (X, Y), and each fit was examined to see if it successfully estimated the sign of β^*. The comparison were performed 500 times and the results are shown in Figure 4(a).

Figure 4(a) displays the results for the standard Lasso and the weighted Lasso. This figure shows that the standard Lasso greatly outperformed the weighted Lasso. In this contrived example, the weighted Lasso has the advantage of knowing the true variance of noise, and this led to a drastic decrease in performance compared to the standard Lasso. In practice, when the variance of the noise terms is not known, the penalized maximum likelihood estimator would be difficult to optimize and would likely have even worse statistical performance.

In this example, the standard Lasso outperformed the weighted Lasso on sign estimation because the standard Lasso had a greater chance of satisfying the irrepresentable condition. We examined this for $n = 400$, $p = 1,000$, $X_{ij} \sim N(0, 1)$, and

$$\beta^* = [\underbrace{5, \ldots, 5}_{10}, \underbrace{10, \ldots, 10}_{10}, \underbrace{0, \ldots, 0}_{980}]'.$$

Here the standard Lasso satisfied the irrepresentable condition 98% of the time, but the weighted Lasso satisfiedit only 15% of the time.

Figure 4(b) compares the prediction error of the standard Lasso estimator and the penalized maximum likelihood estimator. When calculating prediction error, we independently drew 1,000 new samples and calculated

$$MSE = \frac{\sum_{i=1}^{1,000}(y_i^* - x_i^{*T}\hat{\beta})^2}{1,000},$$

where $(x_i^{*T}, y_i^*), i = 1, \ldots, 1,000$, were new samples independent of the training data; $\hat{\beta}$ is the estimation of regression coefficients via standard Lasso or weighted Lasso. In each run of the simulation, the tuning parameter λ was selected by 10-fold cross validation with cv.lars() in the LARS package (Efron et al. (2004)). The square root of average MSE is reported in Figure 4(b). It shows that the weighted Lasso (penalized maximum likelihood) had a smaller prediction error than the standard Lasso, unlike for sign estimation.

To see why the weighted Lasso performed better in terms of prediction, Figure 5 gives boxplots of estimated coefficients $\hat{\beta}$ for both standard Lasso and the weighted Lasso when $\beta_{\min} = 5$ and $\beta_{\max} = 10$. Figure 5 shows that the standard Lasso estimated the zeros in β^* very well, while the weighted Lasso estimated many of the zero elements of β^* to be *non*zero. At the same time, the weighted Lasso estimated the nonzero elements of β^* with less bias and variance than the standard Lasso, explaining why the weighted Lasso has poor sign estimation, but slightly better prediction performance.

5. Conclusion

This paper aims to understand if the sign consistency of the Lasso is robust to the heteroscedastic errors in a Poisson-like model motivated by certain problems in high-dimensional medical imaging (Fessler (2000)). We found that, for sign consistency, the Lasso is robust to this violation of homoscedasticity. Theoretical results for the sparse Poisson-like model are similar to results for the standard model. Simulations suggest that the Lasso performs similarly in terms of model selection performance on both Poisson-like data and homoscedastic data when the variance of the noise is scaled appropriately.

Our results do not extend to general heteroskedastic models as our techniques are highly tuned to the specific the Poisson-like model. High dimensional regression under misspecified models is an important and extensive area for future research. It is our hope that this paper prompts others to study the challenging problems in this area.

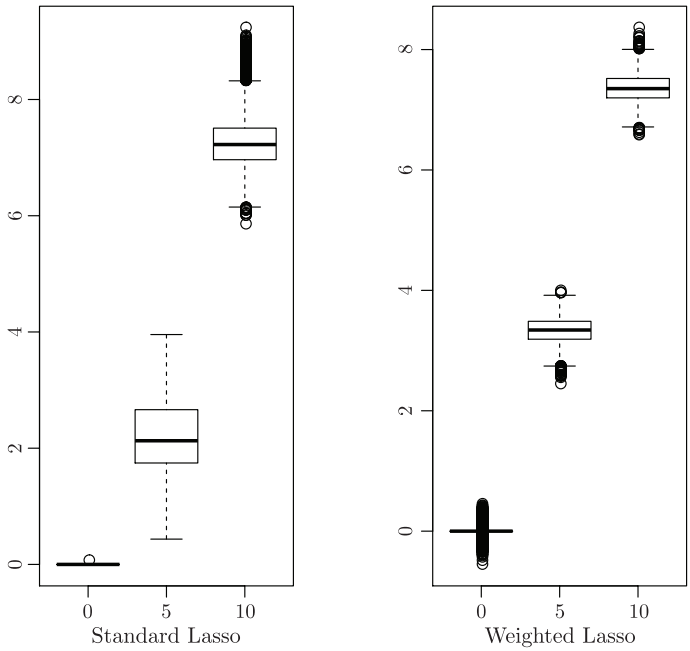


Figure 5. Comparisons of the standard Lasso and the weighted Lasso. The horizontal axes correspond to the true coefficients. The vertical axes correspond to the estimated coefficients. Each plot is the results of 500 simulations.

Acknowledgements

This work was 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. We would like to thank anonymous reviewers whose comments have led to an improved paper. Jinzhu Jia is partially supported by the National Basic Research Program of China (973 Program 2011CB809105) and the National Science Foundation of China (61121002). Jinzhu Jia's work was carried out when he was a postdoc in UC Berkeley supported by NSF grant SES-0835531 (CDI). Karl Rohe's work was carried out when he was a graduate student in UC Berkeley. Bin Yu is partially supported by NSF grants DMS-0907632 and DMS-1107000, CCF-0939370, and ARO grant W911NF-11-1-0114. Karl Rohe is partially supported by NSF VIGRE Graduate Fellowship.

References

Candes, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51**, 4203-4215.

Candes, E. and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52**, 489 - 509.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

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

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

Donoho, D. L. (2004). 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.

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

Donoho, D. L. and Huo., X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47**, 2845-2862.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-451.

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

Feuer, A. and Nemirovski, A. (2003). On sparse representation in pairs of bases. *IEEE Trans. Inform. Theory* **49**, 1579-1581.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101-148.

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

Fuchs, J. (2005). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Trans. Inform. Theory* **51**, 3601-3608.

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

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

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436-1462.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the lasso and its dual. *J. Comput. Graph. Statist.* **9**, 319-37.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

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

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

Vardi, Y., Shepp, L. and Kaufman, L. (1985). A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.* **80**, 8-20.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55**, 2183-2202.

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

Zhao, P. and Yu, B. (2007). Stagewise lasso. *J. Mach. Learn. Res.* **8**, 2701-2726.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871, P. R. China.

E-mail: jzjia@math.pku.edu.cn

Department of Statistics, University of Wisconsin-Madison, WI 53706, USA.

E-mail: karlrohe@stat.wisc.edu

Department of Statistics, and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720, USA.

E-mail: binyu@stat.berkeley.edu

(Received November 2010; accepted April 2012)