On the “Degrees of Freedom” of the Lasso

Hui Zou[*]
Trevor Hastie[†]
Robert Tibshirani[‡]

Abstract

We study the degrees of freedom of the Lasso in the framework of Stein’s unbiased risk estimation (SURE). We show that the number of non-zero coefficients is an unbiased estimate for the degrees of freedom of the Lasso—a conclusion that requires no special assumption on the predictors. Our analysis also provides mathematical support for a related conjecture by Efron et al. (2004). As an application, various model selection criteria—C_p, AIC and BIC—are defined, which, along with the LARS algorithm, provide a principled and efficient approach to obtaining the optimal Lasso fit with the computational efforts of a single ordinary least-squares fit. We propose the use of BIC-Lasso shrinkage if the Lasso is primarily used as a variable selection method.

[*] Department of Statistics, Stanford University, Stanford, CA 94305. Email: hzou@stanford.edu.

[†] Department of Statistics and Department of Health Research & Policy, Stanford University, Stanford, CA 94305. Email: hastie@stanford.edu.

[‡] Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305. Email: tibs@stanford.edu.

1 Introduction

Modern data sets typically have a large number of observations and predictors. A typical goal in model fitting is to achieve good prediction accuracy with a sparse representation of the predictors in the model.

The Lasso is a promising automatic model building technique, simultaneously producing accurate and parsimonious models (Tibshirani 1996). Suppose $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector and $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T, j = 1, \ldots, p$ are the linearly independent predictors. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_p]$ be the predictor matrix. The Lasso estimates for the coefficients of a linear model solve

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t. \tag{1}$$

Or equivalently

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,, \tag{2}$$

where λ is a non-negative regularization parameter. Without loss of generality we assume that the data are centered, so the intercept is not included in the above model. There is a one-one correspondence (generally depending on the data) between t and λ making the optimization problems in (1) and (2) equivalent. The second term in (2) is called the 1-norm penalty and λ is called as the lasso regularization parameter. Since the $Loss + Penalty$ formulation is common in the statistical community, we use the representation (2) throughout this paper. Figure 1 displays the Lasso estimates as a function of λ using the diabetes data (Efron et al. 2004). As can be seen from Figure 1 (the left plot), the Lasso continuously shrinks the coefficients toward zero as λ increases; and some coefficients are shrunk to exact zero if λ is sufficiently large. In addition, the shrinkage often improves the prediction accuracy due to the bias-variance trade-off. Thus the Lasso simultaneously achieves accuracy and sparsity.

Generally speaking, the purpose of regularization is to control the complexity of the fitted model (Hastie et al. 2001). The least regularized Lasso ($\lambda = 0$) corresponds to Ordinary Least Squares (OLS); while the most regularized Lasso uses $\lambda = \infty$, yielding a constant fit. So the model complexity is reduced via shrinkage. However, the effect of the Lasso shrinkage is not very clear except for these two extreme cases. An informative measurement of model complexity is the *effective degrees of freedom* (Hastie & Tibshirani 1990). The profile of degrees of freedom clearly shows that how the model complexity is controlled by shrinkage. The degrees of freedom also plays an important role in estimating the prediction accuracy of the fitted model, which helps us pick an optimal model among all the possible candidates, e.g. the optimal choice of λ in the Lasso. Thus it is desirable to know what is the degrees of freedom of the lasso for a given regularization parameter λ, or $df(\lambda)$. This is an interesting problem of both theoretical and practical importance.

Degrees of freedom are well studied for linear procedures. For example, the degrees of freedom in multiple linear regression exactly equals the number of predictors. A generaliza-

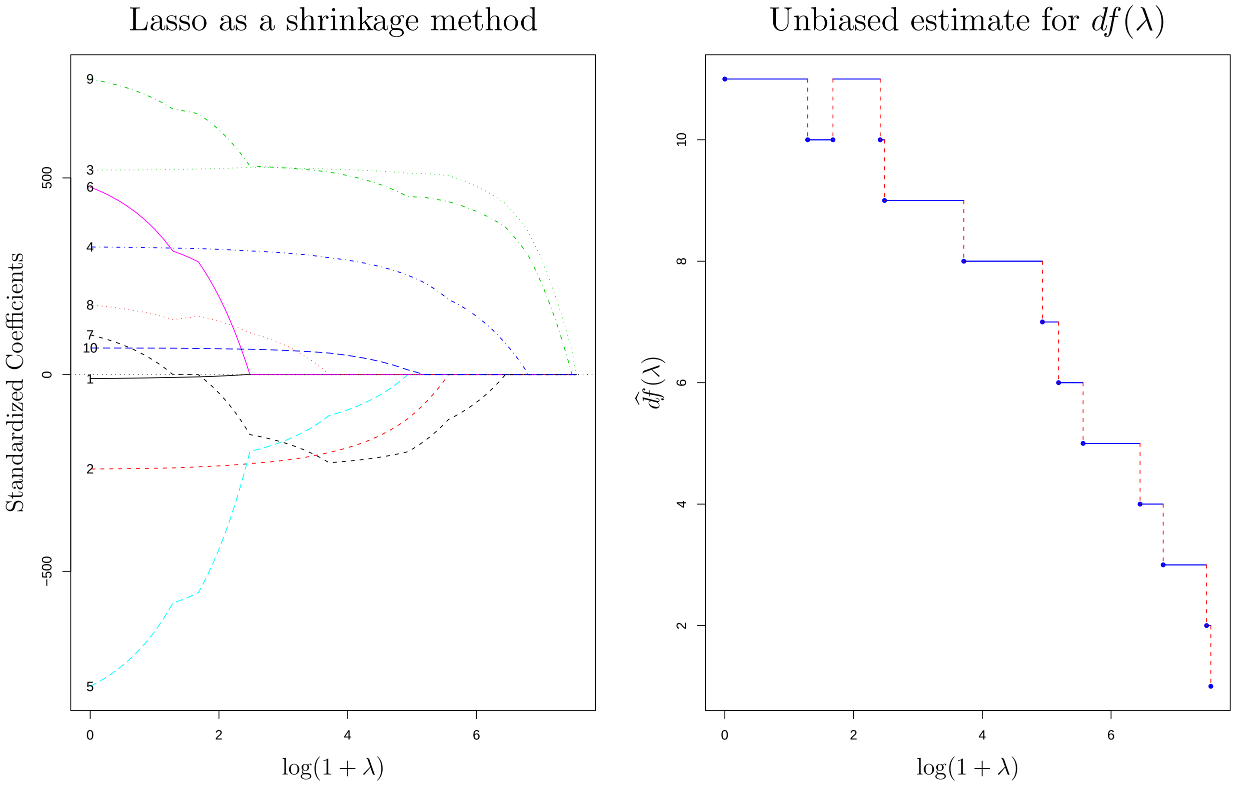


Figure 1: *Diabetes data with 10 predictors. The left panel shows the Lasso coefficients estimates* $\hat{\beta}_j, j = 1, 2, \ldots, 10$*, for the diabetes study. The diabetes data were standardized. The Lasso coefficients estimates are piece-wise linear functions of* λ *(Efron et al. 2004), hence they are piece-wise non-linear as functions of* $\log(1+\lambda)$*. The right panel shows the curve of the proposed unbiased estimate for the degrees of freedom of the Lasso, whose piece-wise constant property is basically determined by the piece-wise linearity of* $\hat{\beta}$*.*

tion is made for all linear smoothers (Hastie & Tibshirani 1990), where the fitted vector is written as $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ and the smoother matrix $\mathbf{S}$ is free of $\mathbf{y}$. Then $df(\mathbf{S}) = \mathrm{tr}(\mathbf{S})$ (see Section 2). A leading example is ridge regression (Hoerl & Kennard 1988) with $\mathbf{S} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}$. These results rely on the convenient expressions for representing linear smoothers. Unfortunately, the explicit expression of the Lasso fit is not available (at least so far) due to the nonlinear nature of the Lasso, thus the nice results for linear smoothers are not directly applicable.

Efron et al. (2004) (referred to as the LAR paper henceforth) propose *Least Angle Regression* (LARS), a new stage-wise model building algorithm. They show that a simple modification of LARS yields the entire Lasso solution path with the computational cost of a single OLS fit. LARS describes the Lasso as a forward stage-wise model fitting process. Starting at zero, the Lasso fits are sequentially updated till reaching the OLS fit, while being piece-wise linear between successive steps. The updates follow the current *equiangular direction*. Figure 2 shows how the Lasso estimates evolve step by step.

From the forward stage-wise point of view, it is natural to consider the number of steps as the meta parameter to control the model complexity. In the LAR paper, it is shown that under a *"positive cone" condition*, the degrees of freedom of LARS equals the number of steps, i.e., $df(\hat{\boldsymbol{\mu}}_k) = k$, where $\hat{\boldsymbol{\mu}}_k$ is the fit at step k. Since the Lasso and LARS coincide under the positive cone condition, the remarkable formula also holds for the Lasso. Under general situations $df(\hat{\boldsymbol{\mu}}_k)$ is still well approximated by k for LARS. However, this simple approximation cannot be true in general for the Lasso because the total number of Lasso steps can exceed the number of predictors. This usually happens when some variables are temporally dropped (coefficients cross zero) during the LARS process, and they are eventually included into the full OLS model. For instance, the LARS algorithm takes 12 Lasso steps to reach the OLS fit as shown in Figure 2, but the number of predictors is 10. For the degrees of freedom of the Lasso under general conditions, Efron et al. (2004) presented the following conjecture.

Conjecture 1 (EHJT04). *Starting at step 0, let m_k be the index of the last model in the Lasso sequence containing k predictors. Then $df(\hat{\boldsymbol{\mu}}_{m_k}) \doteq k$.*

In this paper we study the degrees of freedom of the Lasso using Stein's unbiased risk estimation (SURE) theory (Stein 1981). The Lasso exhibits the backward penalization and forward growth pictures, which consequently induces two different ways to describe its degrees of freedom. With the representation (2), we show that for any given λ the number of non-zero predictors in the model is an unbiased estimate for the degrees of freedom, and no special assumption on the predictors is required, e.g. the positive cone condition. The right panel in Figure 1 displays the unbiased estimate for the degrees of freedom as a function of λ on diabetes data (with 10 predictors). If the Lasso is viewed as a forward stage-wise process, our analysis provides mathematical support for the above conjecture.

The rest of the paper is organized as follows. We first briefly review the SURE theory in Section 2. Main results and proofs are presented in Section 3. In Section 4, model selection criteria are constructed using the degrees of freedom to adaptively select the optimal Lasso fit. We address the difference between two types of optimality: adaptive in prediction and

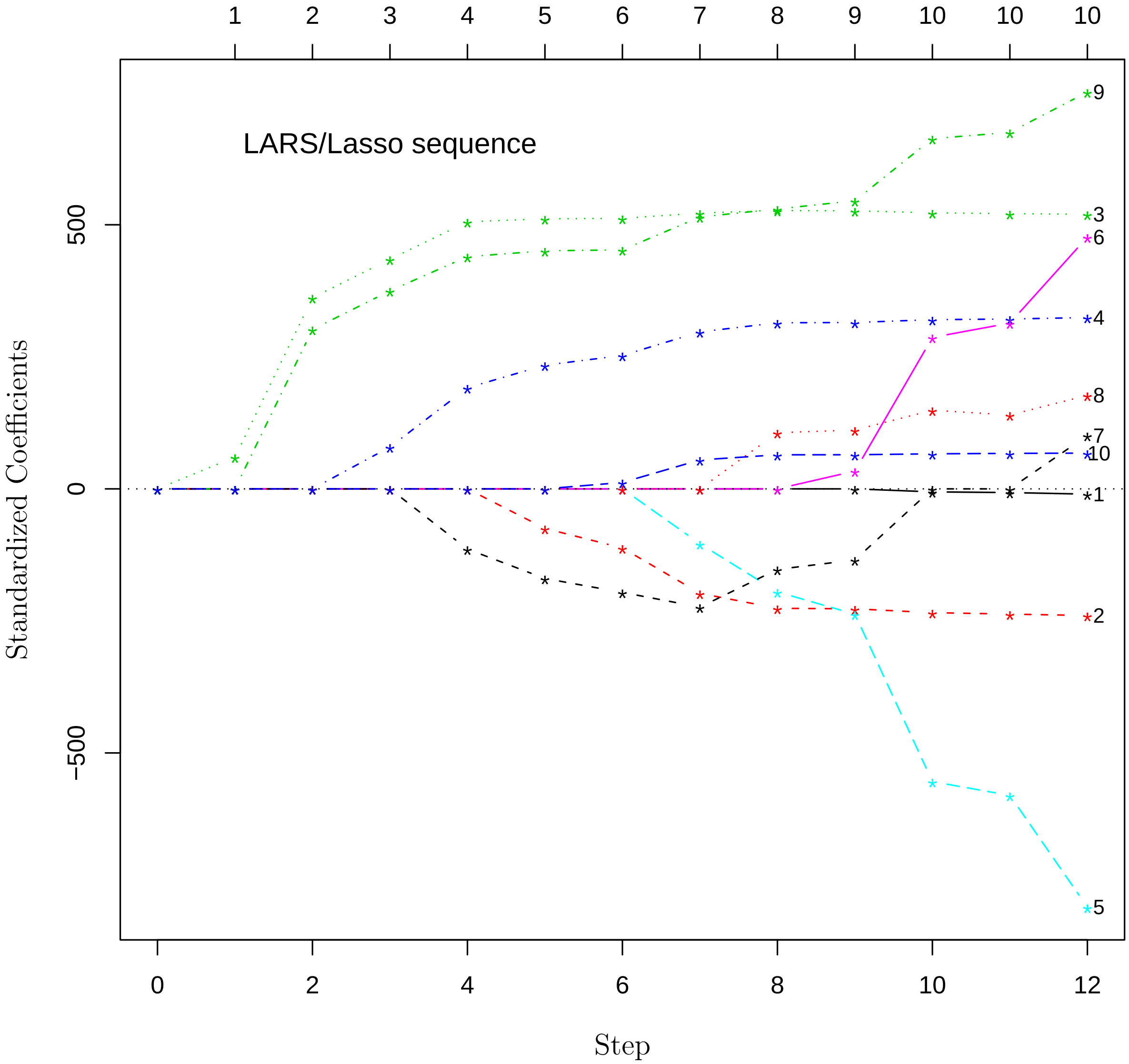


Figure 2: *Diabetes data with 10 predictors: the growth paths of the Lasso coefficients estimates as the LARS algorithm moves forward. On the top of the graph, we display the number of non-zero coefficients at each step.*

adaptive in variable selection. Discussions are in Section 5. Proofs of lemmas are presented in the appendix.

2 Stein's Unbiased Risk Estimation

We begin with a brief introduction to the Stein's unbiased risk estimation (SURE) theory (Stein 1981) which is the foundation of our analysis. The readers are referred to Efron (2004) for detailed discussions and recent references on SURE.

Given a model fitting method δ, let $\hat{\boldsymbol{\mu}} = \delta(\mathbf{y})$ represent its fit. We assume a homoskedastic model, i.e., given the $\mathbf{x}$'s, $\mathbf{y}$ is generated according to

$$\mathbf{y} \sim (\boldsymbol{\mu}, \sigma^2 \mathbf{I}), \tag{3}$$

where $\boldsymbol{\mu}$ is the true mean vector and σ^2 is the common variance. The focus is how accurate δ can be in predicting future data. Suppose $\mathbf{y}^{new}$ is a new response vector generated from (3), then under the squared-error loss, the prediction risk is $E\Big\{ \|\hat{\boldsymbol{\mu}} - \mathbf{y}^{new}\|^2 \Big\}/n$. Efron (2004) shows that

$$E\{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\} = E\{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 - n\sigma^2\} + 2\sum_{i=1}^{n} \operatorname{cov}(\hat{\boldsymbol{\mu}}_i, y_i). \tag{4}$$

The last term of (4) is called the *optimism* of the estimator $\hat{\boldsymbol{\mu}}$ (Efron 1986). Identity (4) also gives a natural definition of the *degrees of freedom* for an estimator $\hat{\boldsymbol{\mu}} = \delta(\mathbf{y})$,

$$df(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \operatorname{cov}(\hat{\boldsymbol{\mu}}_i, y_i)/\sigma^2. \tag{5}$$

If δ is a linear smoother, i.e., $\hat{\boldsymbol{\mu}} = \mathbf{S}\mathbf{y}$ for some matrix $\mathbf{S}$ independent of $\mathbf{y}$, then it is easy to verify that since $\operatorname{cov}(\hat{\boldsymbol{\mu}}, \mathbf{y}) = \sigma^2 S$, $df(\hat{\boldsymbol{\mu}}) = \operatorname{tr}(\mathbf{S})$, which coincides with the definition given by Hastie & Tibshirani (1990). By (4) we obtain

$$E\{\|\hat{\boldsymbol{\mu}} - \mathbf{y}^{new}\|^2\} = E\{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 + 2df(\hat{\boldsymbol{\mu}})\ \sigma^2\}. \tag{6}$$

Thus we can define a C_p-type statistic

$$C_p(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{n} + \frac{2df(\hat{\boldsymbol{\mu}})}{n}\sigma^2 \tag{7}$$

which is an unbiased estimator of the true prediction error. When σ^2 is unknown, it is replaced with an unbiased estimate.

Stein proves an extremely useful formula to simplify (5), which is often referred to as Stein's Lemma (Stein 1981). According to Stein, a function $g : \mathbb{R}^n \to \mathbb{R}$ is said to be *almost differentiable* if there is a function $f : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$g(x+u) - g(x) = \int_0^1 u^T f(x+tu)dt \tag{8}$$

for a.e. $x \in \mathbb{R}^n$, each $u \in \mathbb{R}^n$.

Lemma 1 (Stein's Lemma). *Suppose that* $\hat{\boldsymbol{\mu}} : \mathbb{R}^n \to \mathbb{R}^n$ *is almost differentiable and denote* $\nabla \cdot \hat{\boldsymbol{\mu}} = \sum_{i=1}^n \partial \hat{\boldsymbol{\mu}}_i / \partial y_i$. *If* $\mathbf{y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$, *then*

$$\sum_{i=1}^n \mathrm{cov}(\hat{\boldsymbol{\mu}}_i, y_i)/\sigma^2 = E[\nabla \cdot \hat{\boldsymbol{\mu}}]. \tag{9}$$

In many applications $\nabla \cdot \hat{\boldsymbol{\mu}}$ is shown to be a constant; for example, with $\hat{\boldsymbol{\mu}} = S\mathbf{y}$, $\nabla \cdot \hat{\boldsymbol{\mu}} = \mathrm{tr}(S)$. Thus the degrees of freedom is easily obtained. Even if $\nabla \cdot \hat{\boldsymbol{\mu}}$ depends on y, Stein's Lemma says

$$\widehat{df}(\hat{\boldsymbol{\mu}}) = \nabla \cdot \hat{\boldsymbol{\mu}} \tag{10}$$

is an unbiased estimate for the degrees of freedom $df(\hat{\boldsymbol{\mu}})$. In the spirit of SURE, we can use

$$C_p^*(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{n} + \frac{2\widehat{df}(\hat{\boldsymbol{\mu}})}{n}\,\sigma^2 \tag{11}$$

as an unbiased estimate for the true risk. It is worth mentioning that in some situations verifying the almost differentiability of $\hat{\boldsymbol{\mu}}$ is not easy.

Even though Stein's Lemma assumes normality, the essence of (9) only requires homoskedasticity (3) and the almost differentiability of $\hat{\boldsymbol{\mu}}$; its justification can be made by a "delta method" argument (Efron et al. 2004). After all, $df(\hat{\boldsymbol{\mu}})$ is about the self-influence of $\mathbf{y}$ on the fit, and $\nabla \cdot \hat{\boldsymbol{\mu}}$ is a natural candidate for that purpose. Meyer & Woodroofe (2000) discussed the degrees of freedom in shape-restricted regression and argued that the divergence formula (10) provides a measure of the effective dimension.

3 Main Theorems

We adopt the SURE framework with the Lasso fit. Let $\hat{\boldsymbol{\mu}}_\lambda$ be the Lasso fit using the representation (2). Similarly, let $\hat{\boldsymbol{\mu}}_m$ be the Lasso fit at step m in the LARS algorithm. For convenience, we also let $df(\lambda)$ and $df(m)$ stand for $df(\hat{\boldsymbol{\mu}}_\lambda)$ and $df(\hat{\boldsymbol{\mu}}_m)$, respectively.

The following matrix representation of Stein's Lemma is helpful. Let $\frac{\partial \hat{\boldsymbol{\mu}}}{\partial \mathbf{y}}$ be a $n \times n$ matrix whose elements are

$$\left(\frac{\partial \hat{\boldsymbol{\mu}}}{\partial \mathbf{y}}\right)_{i,j} = \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial y_j} \quad i, j = 1, 2, \ldots, n. \tag{12}$$

Then we can write

$$\nabla \cdot \hat{\boldsymbol{\mu}} = \mathrm{tr}\left(\frac{\partial \hat{\boldsymbol{\mu}}}{\partial \mathbf{y}}\right). \tag{13}$$

Suppose $\mathbf{M}$ is a matrix with p columns. Let $\mathcal{S}$ be a subset of the indices $\{1, 2, \ldots, p\}$. Denote by $\mathbf{M}_{\mathcal{S}}$ the sub-matrix

$$\mathbf{M}_{\mathcal{S}} = [\cdots m_j \cdots]_{j \in \mathcal{S}}, \tag{14}$$

where m_j is the j-th column of $\mathbf{M}$. Similarly, define $\beta_{\mathcal{S}} = (\cdots \beta_j \cdots)_{j \in \mathcal{S}}$ for any vector β of length p. Let Sgn$(\cdot)$ be the sign function:

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

3.1 Results and data examples

Our results are stated as follows. Denote the set of non-zero elements of $\hat{\beta}_\lambda$ as $\mathcal{B}(\lambda)$, then

$$df(\lambda) = E[|\mathcal{B}_\lambda|] \tag{15}$$

where $|\mathcal{B}_\lambda|$ means the size of $\mathcal{B}(\lambda)$. Hence $\widehat{df}(\lambda) = |\mathcal{B}_\lambda|$ is an unbiased estimate for $df(\lambda)$. The identity (15) holds for all $\mathbf{X}$, requiring no special assumption.

We also provide mathematical support for the conjecture in Section 1. Actually we argue that if m_k is a Lasso step containing k non-zero predictors, then $\widehat{df}(m_k) = k$ is a good estimate for $df(m_k)$. Note that m_k is not necessary the last Lasso step containing k non-zero predictors. So the result includes the conjecture as a special case. However, we show in Section 4 that the last step choice is superior in the Lasso model selection. We let m_k^{last} and m_k^{first} denote the last and first Lasso step containing exact k non-zero predictors, respectively.

Before delving into the detail of theoretical analysis, we check the validity of our arguments by a simulation study. Here is the outline of the simulation. We take the 64 predictors in the diabetes data which include the quadratic terms and interactions of the original 10 predictors. The positive cone condition is violated on the 64 predictors (Efron et al. 2004). The response vector $\mathbf{y}$ was used to fit a OLS model. We computed the OLS estimates $\hat{\beta}_{ols}$ and $\hat{\sigma}^2_{ols}$. Then we considered a synthetic model

$$\mathbf{y}^* = \mathbf{X}\beta + N(0,1)\sigma, \tag{16}$$

where $\beta = \hat{\beta}_{ols}$ and $\sigma = \hat{\sigma}_{ols}$.

Given the synthetic model, the degrees of freedom of the Lasso (both $df(\lambda)$ and $df(m_k)$) can be numerically evaluated by Monte Carlo methods. For $b = 1, 2, \ldots, B$, we independently simulated $\mathbf{y}^*(b)$ from (16). For a given λ, by the definition of $df(\lambda)$, we need to evaluate

$$\text{cov}_i = E[(\hat{\boldsymbol{\mu}}_{\lambda,i} - E[\hat{\boldsymbol{\mu}}_{\lambda,i}])(\mathbf{y}^*_i - (\mathbf{X}\beta)_i)]. \tag{17}$$

Then $df(\lambda) = \sum_{i=1}^n \text{cov}_i/\sigma^2$. Since $E[\mathbf{y}^*_i] = (\mathbf{X}\beta)_i$ and note that

$$\text{cov}_i = E[(\hat{\boldsymbol{\mu}}_{\lambda,i} - a_i)(\mathbf{y}^*_i - (\mathbf{X}\beta)_i)] \tag{18}$$

for any fixed known constant a_i. Then we compute

$$\widehat{\text{cov}}_i = \frac{\sum_{b=1}^B \left(\hat{\boldsymbol{\mu}}_{\lambda,i}(b) - a_i\right)\left(\mathbf{y}^*_i(b) - (\mathbf{X}\beta)_i\right)}{B} \tag{19}$$

and $df(\lambda) = \sum_{i=1}^{n} \widehat{\text{cov}}_i/\sigma^2$. Typically $a_i = 0$ is used in Monte Carlo calculation. In this work we use $a_i = (\mathbf{X}\beta)_i$, for it gives a Monte Carlo estimate for $df(\lambda)$ with smaller variance than that by $a_i = 0$. On the other hand, for a fixed λ, each $\mathbf{y}^*(b)$ gave the Lasso fit $\hat{\boldsymbol{\mu}}_\lambda(b)$ and the df estimate $\widehat{df}(\lambda)_b$. Then we evaluated $E[|\mathcal{B}_\lambda|]$ by $\sum_{b=1}^{B} \widehat{df}(\lambda)_b/B$. Similarly, we computed $df(m_k)$ by replacing $\hat{\boldsymbol{\mu}}_\lambda(b)$ with $\hat{\boldsymbol{\mu}}_{m_k}(b)$. We are interested in $E[|\mathcal{B}_\lambda|] - df(\lambda)$ and $k - df(m_k)$. Standard errors were calculated based on the B replications.

Figure 3 is a very convincing picture for the identity (15). Figure 4 shows that $df(m_k)$ is well approximated by k even when the positive cone condition is failed. The simple approximation works pretty well for both m_k^{last} and m_k^{first}.

In Figure 4, it appears that $k - df(m_k)$ is not exactly zero for some k. We would like to check if the bias is real. Furthermore, if the bias is real, then we would like to explore the relation between the bias $k - df(m_k)$ and the signal/noise ratio. In the synthetic model (16) the signal/noise ratio $\frac{\text{Var}(\mathbf{X}\hat{\beta}_{ols})}{\hat{\sigma}^2_{ols}}$ is about 1.25. We repeated the same simulation procedure with $(\beta = 0, \sigma = 1)$ and $(\beta = \hat{\beta}_{ols}, \sigma = \frac{\hat{\sigma}_{ols}}{10})$ in the synthetic model. The corresponding signal/noise ratios are zero and 125, respectively.

As shown clearly in Figure 5, the bias $k - df(m_k)$ is truly non-zero for some k. Thus the positive cone condition seems to be sufficient and necessary for turning the approximation into an exact result. However, even if the bias exists, its maximum magnitude is less than one, regardless the size of the signal/noise ratio. So k is a very good estimate for $df(m_k)$. An interesting observation is that k tends to underestimate $df(m_k^{\text{last}})$ and overestimate $df(m_k^{\text{first}})$. In addition, we observe that $k - df(m_k^{\text{last}}) \doteq df(m_k^{\text{first}}) - k$.

3.2 Theorems on $df(\lambda)$

Let $\mathcal{B} = \{j : \text{Sgn}(\beta)_j \neq 0\}$ be the *active set* of β where $\text{Sgn}(\beta)$ is the sign vector of β given by $\text{Sgn}(\beta)_j = \text{Sgn}(\beta_j)$. We denote the active set of $\hat{\beta}(\lambda)$ as $\mathcal{B}(\lambda)$ and the corresponding sign vector $\text{Sgn}(\hat{\beta}(\lambda))$ as $\text{Sgn}(\lambda)$. We do not distinguish the index of a predictor and the predictor itself.

Firstly, let us review some characteristics of the Lasso solution. For a given response vector $\mathbf{y}$, there are a sequence of λ's:

$$\lambda_0 > \lambda_1 > \lambda_2 \cdots > \lambda_K = 0 \quad \text{such that:} \tag{20}$$

- For all $\lambda > \lambda_0$, $\hat{\beta}(\lambda) = 0$.
- In the interior of the interval $(\lambda_{m+1}, \lambda_m)$, the active set $\mathcal{B}(\lambda)$ and the sign vector $\text{Sgn}(\lambda)_{\mathcal{B}(\lambda)}$ are constant with respect to λ. Thus we write them as $\mathcal{B}_m$ and Sgn_m for convenience.

The active set changes at each λ_m. When λ decreases from $\lambda = \lambda_m - 0$, some predictors with zero coefficients at λ_m are about to have non-zero coefficients, thus they join the active set $\mathcal{B}_m$. However, as λ approaches $\lambda_{m+1} + 0$ there are possibly some predictors in $\mathcal{B}_m$ whose coefficients reach zero. Hence we call $\{\lambda_m\}$ the *transition points*.

We shall proceed by proving the following lemmas (proofs are given in the appendix).

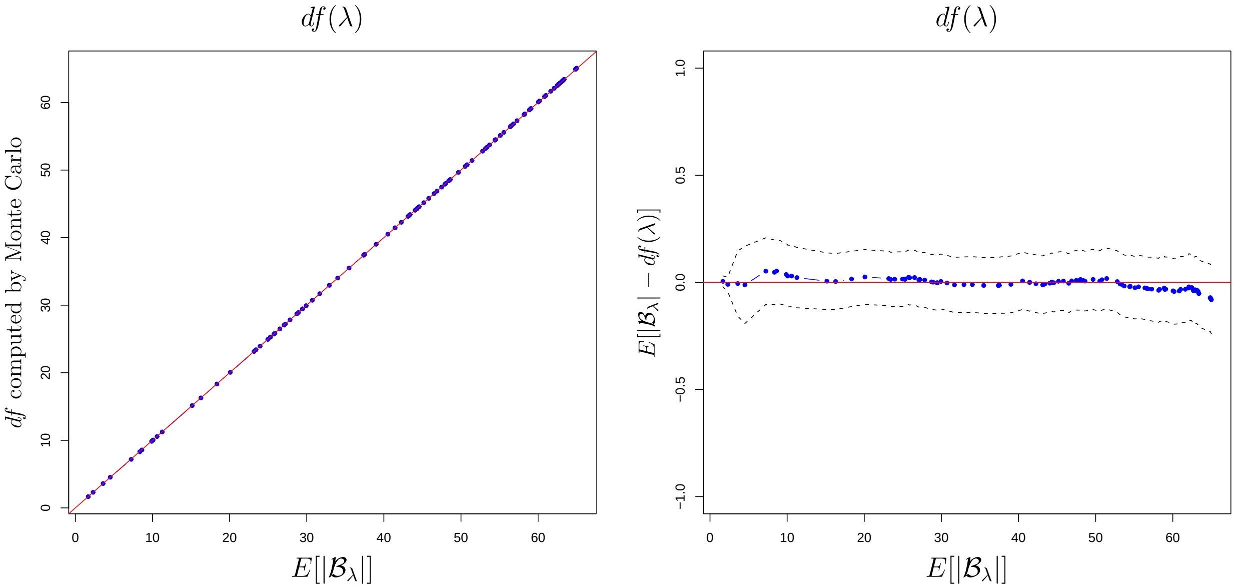
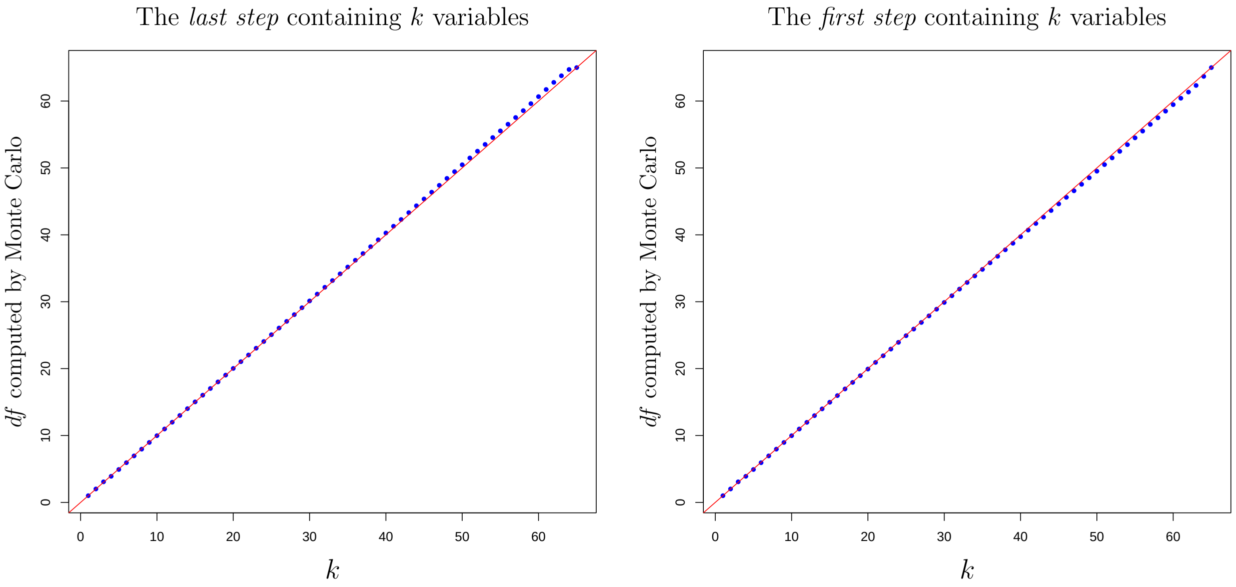

$df(\lambda)$ $df(\lambda)$

Figure 3: *The synthetic model with the 64 predictors in the diabetes data. In the left panel we compare* $E[|\mathcal{B}_\lambda|]$ *with the true degrees of freedom* $df(\lambda)$ *based on* $B = 20000$ *Monte Carlo simulations. The solid line is the* 45^0 *line (the perfect match). The right panel shows the estimation bias and its point-wise* 95% *confidence intervals indicated by the thin dashed lines. Note that the zero horizontal line is well inside the confidence intervals.*


The *last step* containing k variables

The *first step* containing k variables

Figure 4: *The synthetic model with the 64 predictors in the diabetes data. We compare* $\widehat{df}(m_k)$ *with the true degrees of freedom* $df(m_k)$ *based on* $B = 20000$ *Monte Carlo simulations. We consider two choices of* m_k*: in the left panel* m_k *is the last Lasso step containing exact* k *non-zero variables, while the right panel chooses the first Lasso step containing exact* k *non-zero variables. As can be seen from the plots, our formula works pretty well in both cases.*

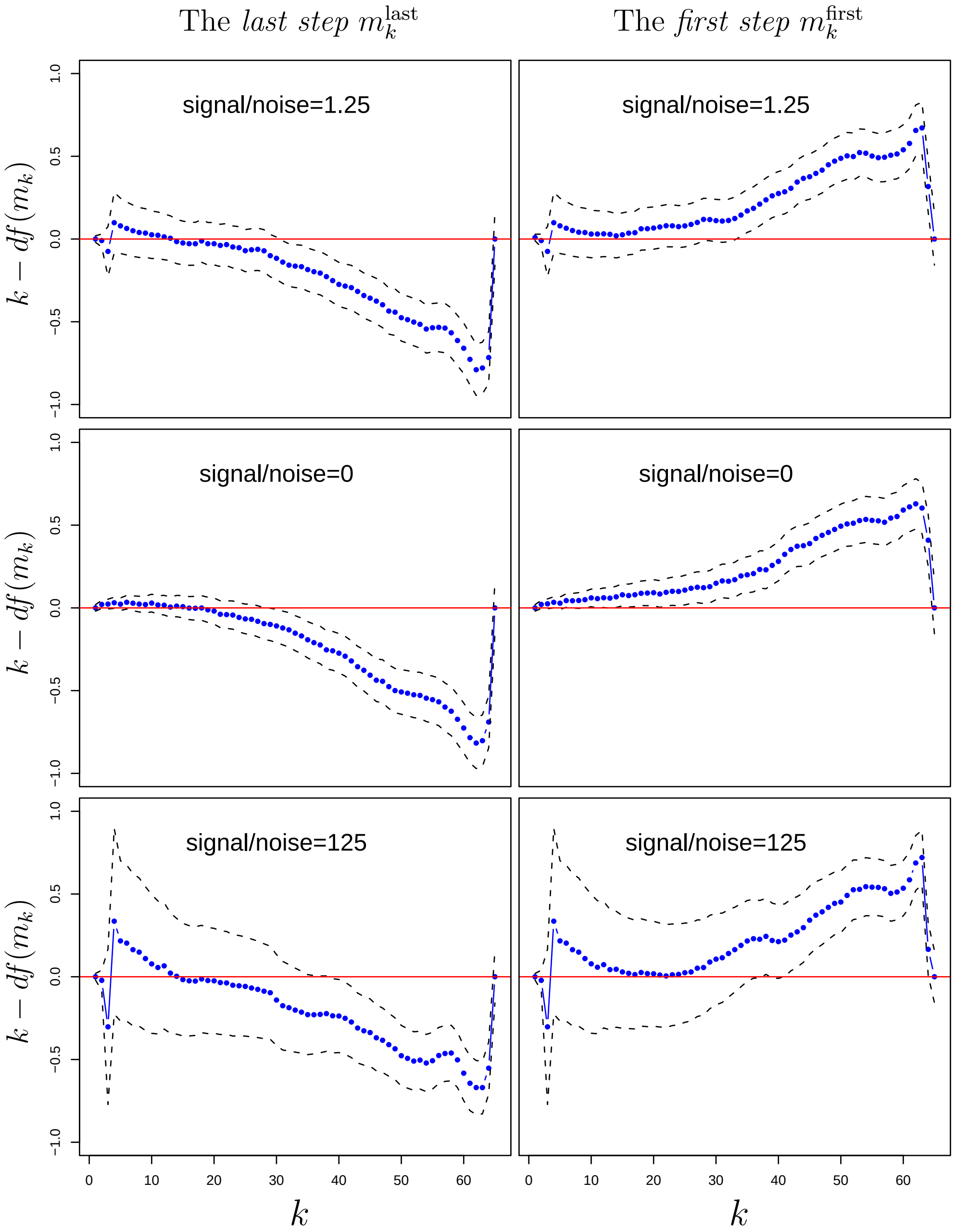


Figure 5: *$B = 20000$ replications were used to assess the bias of $\widehat{df}(m_k) = k$. The 95% point-wise confidence intervals are indicated by the thin dashed lines. Under the positive cone condition, it is exactly the true degrees of freedom $df(m_k)$. This simulation suggests that when the positive cone condition is violated, $df(m_k) \neq k$ for some k. However, the bias is small (the maximum absolute bias is about 0.8). It seems that k tends to underestimate $df(m_k^{\text{last}})$ and overestimate $df(m_k^{\text{first}})$. In addition, we observe that $k - df(m_k^{\text{last}}) \doteq df(m_k^{\text{first}}) - k$. The most important message is that the magnitude of the bias is always less than one, regardless the size of the signal/noise ratio.*

Lemma 2. *Suppose $\lambda \in (\lambda_{m+1}, \lambda_m)$. $\hat{\beta}(\lambda)$ are the Lasso coefficient estimates. Then we have*

$$\hat{\beta}(\lambda)_{\mathcal{B}_m} = \left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m}\right)^{-1} \left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{y} - \frac{\lambda}{2} \operatorname{Sgn}_m\right). \tag{21}$$

Lemma 3. *Consider the transition points λ_m and λ_{m+1}, $\lambda_{m+1} \geq 0$. $\mathcal{B}_m$ is the active set in $(\lambda_{m+1}, \lambda_m)$. Suppose i_{add} is an index added into $\mathcal{B}_m$ at λ_m and its index in $\mathcal{B}_m$ is i^*, i.e., $i_{add} = (\mathcal{B}_m)_{i^*}$. Denote by $(a)_k$ the k-th element of the vector a. We can express the transition point λ_m as follows:*

$$\lambda_m = \frac{2\left(\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m}\right)^{-1} \mathbf{X}_{\mathcal{B}_m}^T \mathbf{y}\right)_{i^*}}{\left(\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m}\right)^{-1} \operatorname{Sgn}_m\right)_{i^*}} \tag{22}$$

Moreover, if j_{drop} is a dropped (if there is any) index at λ_{m+1} and $j_{drop} = (\mathcal{B}_m)_{j^*}$, then λ_{m+1} can be written as:

$$\lambda_{m+1} = \frac{2\left(\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m}\right)^{-1} \mathbf{X}_{\mathcal{B}_m}^T \mathbf{y}\right)_{j^*}}{\left(\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m}\right)^{-1} \operatorname{Sgn}_m\right)_{j^*}} \tag{23}$$

Lemma 4. *$\forall\ \lambda > 0$, $\exists$ a null set $\mathcal{N}_\lambda$ which is a finite collection of hyperplanes in $\mathbb{R}^n$. Let $\mathcal{G}_\lambda = \mathbb{R}^n \setminus \mathcal{N}_\lambda$. Then $\forall \mathbf{y}\ \in \mathcal{G}_\lambda$, λ is not any of the transition points, i.e., $\lambda \notin \{\lambda(\mathbf{y})_m\}$.*

Lemma 5. *$\forall \lambda$, $\hat{\beta}_\lambda(\mathbf{y})$ is a continuous function of $\mathbf{y}$.*

Lemma 6. *Fix any $\lambda > 0$, consider $\mathbf{y} \in \mathcal{G}_\lambda$ as defined in Lemma 4. The active set $\mathcal{B}(\lambda)$ and the sign vector $\operatorname{Sgn}(\lambda)$ are locally constant with respect to $\mathbf{y}$.*

Theorem 1. *Let $\mathcal{G}_0 = \mathbb{R}^n$. Fix an arbitrary $\lambda \geq 0$. On the set $\mathcal{G}_\lambda$ with full measure as defined in Lemma 4, the Lasso fit $\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y})$ is uniformly Lipschitz. Precisely,*

$$\|\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y} + \Delta\mathbf{y}) - \hat{\boldsymbol{\mu}}_\lambda(\mathbf{y})\| \leq \|\Delta\mathbf{y}\| \quad \textit{for sufficiently small} \ \ \Delta\mathbf{y} \tag{24}$$

Moreover, we have the divergence formula

$$\nabla \cdot \hat{\boldsymbol{\mu}}_\lambda(\mathbf{y}) = |\mathcal{B}_\lambda|. \tag{25}$$

Proof. If $\lambda = 0$, then the Lasso fit is just the OLS fit. The conclusions are easy to verify. So we focus on $\lambda > 0$. Fix a $\mathbf{y}$. Choose a small enough ϵ such that $\operatorname{Ball}(\mathbf{y}, \epsilon) \subset \mathcal{G}_\lambda$.

Since λ is not any transition point, using (21) we observe

$$\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y}) = \mathbf{X}\hat{\beta}(\mathbf{y}) = \mathbf{H}_\lambda(\mathbf{y})\mathbf{y} - \lambda\boldsymbol{\omega}_\lambda(\mathbf{y}), \tag{26}$$

where $\mathbf{H}_\lambda(\mathbf{y}) = \mathbf{X}_{\mathcal{B}_\lambda}(\mathbf{X}_{\mathcal{B}_\lambda}^T \mathbf{X}_{\mathcal{B}_\lambda})^{-1}\mathbf{X}_{\mathcal{B}_\lambda}^T$ is the projection matrix on the space $\mathbf{X}_{\mathcal{B}_\lambda}$ and $\omega_\lambda(\mathbf{y}) = \frac{1}{2}\ \mathbf{X}_{\mathcal{B}_\lambda}(\mathbf{X}_{\mathcal{B}_\lambda}^T \mathbf{X}_{\mathcal{B}_\lambda})^{-1}\operatorname{Sgn}_{\mathcal{B}_\lambda}$. Consider $\|\Delta\mathbf{y}\| < \epsilon$. Similarly, we get

$$\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y} + \Delta\mathbf{y}) = \mathbf{H}_\lambda(\mathbf{y} + \Delta\mathbf{y})(\mathbf{y} + \Delta\mathbf{y}) - \lambda\boldsymbol{\omega}_\lambda(\mathbf{y} + \Delta\mathbf{y}). \tag{27}$$

Lemma 6 says that we can further let ϵ be sufficiently small such that both the effective set $\mathcal{B}_\lambda$ and the sign vector Sgn_λ stay constant in $\text{Ball}(\mathbf{y}, \epsilon)$. Now fix ϵ. Hence if $\|\Delta\mathbf{y}\| < \epsilon$, then

$$\mathbf{H}_\lambda(\mathbf{y} + \Delta\mathbf{y}) = \mathbf{H}_\lambda(\mathbf{y}) \quad \text{and} \quad \boldsymbol{\omega}_\lambda(\mathbf{y} + \Delta\mathbf{y}) = \boldsymbol{\omega}_\lambda(\mathbf{y}). \tag{28}$$

Then (26) and (27) give

$$\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y} + \Delta\mathbf{y}) - \hat{\boldsymbol{\mu}}_\lambda(\mathbf{y}) = \mathbf{H}_\lambda(\mathbf{y})\Delta\mathbf{y}. \tag{29}$$

But since $\|\mathbf{H}_\lambda(\mathbf{y})\Delta\mathbf{y}\| \leq \|\Delta\mathbf{y}\|$, (24) is proved.

By the local constancy of $H(\mathbf{y})$ and $\omega(\mathbf{y})$, we have

$$\frac{\partial \hat{\boldsymbol{\mu}}_\lambda(\mathbf{y})}{\partial \mathbf{y}} = \mathbf{H}_\lambda(\mathbf{y}). \tag{30}$$

Then the trace formula (13) implies

$$\nabla \cdot \hat{\boldsymbol{\mu}}_\lambda(\mathbf{y}) = \text{tr}\left(\mathbf{H}_\lambda(\mathbf{y})\right) = |\mathcal{B}_\lambda|. \tag{31}$$

□

By standard analysis arguments, it is easy to check the following proposition
Proposition *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and suppose f is uniformly Lipschitz on $\mathcal{G} = \mathbb{R}^n \setminus \mathcal{N}$ where $\mathcal{N}$ is a finite set of hyperplanes. If f is continuous, then f is uniformly Lipschitz on $\mathbb{R}^n$.*

Theorem 2. $\forall\ \lambda$ *the Lasso fit $\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y})$ is uniformly Lipschitz. The degrees of freedom of $\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y})$ equal the expectation of the effective set $\mathcal{B}_\lambda$, i.e.,*

$$df(\lambda) = E\left[|\mathcal{B}_\lambda|\right]. \tag{32}$$

Proof. The proof is trivial for $\lambda = 0$. We only consider $\lambda > 0$. By Theorem 1 and the proposition, we conclude that $\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y})$ is uniformly Lipschitz. Therefore $\hat{\boldsymbol{\mu}}_\lambda(\mathbf{y})$ is almost differentiable, see Meyer & Woodroofe (2000) and Efron et al. (2004). Then (32) is obtained by Stein's Lemma and the divergence formula (25). □

3.3 $df(m_k)$ and the conjecture

In this section we provide mathematical support for the conjecture in Section 1. The conjecture becomes a simple fact for two trivial cases $k = 0$ and $k = p$, thus we only need to consider $k = 1, \ldots, (p-1)$. Our arguments rely on the details of the LARS algorithm. For the sake of clarity, we first briefly describe the LARS algorithm. The readers are referred to the LAR paper (Efron et al. 2004) for the complete description.

The LARS algorithm sequentially updates the Lasso estimate in a predictable way. Initially (the 0 step), let $\hat{\beta}_0 = 0$, $A_0 = \varnothing$. Suppose that $\hat{\beta}_m$ is the vector of current Lasso coefficient estimates. Then $\hat{\boldsymbol{\mu}}_m = \mathbf{X}\hat{\beta}_m$ and $\hat{r}_m = \mathbf{y} - \hat{\boldsymbol{\mu}}_m$ are the current fit and residual vectors. We say $\hat{c} = \mathbf{X}^T\hat{r}_m$ is the vector of current correlations. Define

$$\widehat{C} = \max_j\{|\hat{c}|\} \quad \mathcal{W}_m = \{j : |\hat{c}_j| = \widehat{C} \text{ and } j \in A_m^c\}. \tag{33}$$

Then $\lambda_m = 2\widehat{C}$. Define the current active set $\mathcal{A} = A_m \cup \mathcal{W}_m$ and the signed matrix

$$X_{\mathcal{A}}^{\text{sign}} = (\cdots \operatorname{Sgn}(\hat{c}_j)\mathbf{x}_j \cdots)_{j\in\mathcal{A}}. \tag{34}$$

Let $\mathcal{G}_{\mathcal{A}} = \left(X_{\mathcal{A}}^{\text{sign}}\right)^T X_{\mathcal{A}}^{\text{sign}}$. $1_{\mathcal{A}}$ is a vector of 1's of length $|\mathcal{A}|$. Then we compute the *equiangular vector*

$$\mathbf{u}_{\mathcal{A}} = X_{\mathcal{A}}^{\text{sign}} w_{\mathcal{A}} \quad \text{with} \quad w_{\mathcal{A}} = DG_{\mathcal{A}}^{-1}1_{\mathcal{A}}, \tag{35}$$

where $D = (1_{\mathcal{A}}^T\mathcal{G}_{\mathcal{A}}^{-1}1_{\mathcal{A}})^{-\frac{1}{2}}$. Let the inner product vector $\mathbf{a} = \mathbf{X}^T\mathbf{u}_{\mathcal{A}}$ and

$$\widehat{\gamma} = \min{}^{+}_{j\in\mathcal{A}^c}\left\{\frac{\widehat{C}-\widehat{c}_j}{D-a_j}, \frac{\widehat{C}+\widehat{c}_j}{D+a_j}\right\}. \tag{36}$$

For $j \in \mathcal{A}$ we compute $d_j = \operatorname{Sgn}(\hat{c}_j)w_{\mathcal{A}_j}$ and $\gamma_j = -(\hat{\beta}_m)_j/d_j$. Define

$$\widetilde{\gamma} = \min_{\gamma_j>0}\{\gamma_j\} \quad \text{and} \quad \mathcal{V}_m = \{j : \gamma_j = \widetilde{\gamma}\ j \in \mathcal{A}\}. \tag{37}$$

The Lasso coefficient estimates are updated by

$$(\hat{\beta}_{m+1})_j = (\hat{\beta}_m)_j + \min\{\widehat{\gamma}, \widetilde{\gamma}\}d_j \quad \text{for } j \in \mathcal{A}. \tag{38}$$

The set A_m is also updated. If $\widehat{\gamma} < \widetilde{\gamma}$ then $A_{m+1} = \mathcal{A}$. Otherwise $A_{m+1} = \mathcal{A}\backslash\mathcal{V}_m$.

Let q_m be the indicator of whether $\mathcal{V}_m$ is dropped or not. Define $q_m\mathcal{V}_m = \mathcal{V}_m$ if $q_m = 1$, otherwise $q_m\mathcal{V}_m = \varnothing$; and conventionally let $\mathcal{V}_{-1} = \varnothing$ and $q_{-1}\mathcal{V}_{-1} = \varnothing$. Considering the active set $\mathcal{B}_\lambda$ as a function of λ, we summarize the following facts

$$|\mathcal{B}_\lambda| = |\mathcal{B}_{\lambda_m}| + |\mathcal{W}_m| \quad \text{if } \lambda_m < \lambda < \lambda_{m+1}, \tag{39}$$

$$|\mathcal{B}_{\lambda_{m+1}}| = |\mathcal{B}_{\lambda_m}| + |\mathcal{W}_m| - |q_m\mathcal{V}_m|. \tag{40}$$

In the LARS algorithm, the Lasso is regarded as one kind of forward stage-wise method for which the number of steps is often used as an effective regularization parameter. For each k, $k \in \{1, 2, \ldots, (p-1)\}$, we seek the models with k non-zero predictors. Let

$$\Lambda_k = \{m : |\mathcal{B}_{\lambda_m}| = k\}. \tag{41}$$

The conjecture is asking for the fit using $m_k^{\text{last}} = \sup(\Lambda_k)$. However, it may happen that for some k there is no such m with $|\mathcal{B}_{\lambda_m}| = k$. For example, if $\mathbf{y}$ is an equiangular vector of all $\{\mathbf{X}_j\}$, then the Lasso estimates become the OLS estimates after just one step. So $\Lambda_k = \varnothing$ for $k = 2, \ldots, (p-1)$. The next Lemma concerns this type of situation. Basically, it shows that the *"one at a time"* condition (Efron et al. 2004) holds almost everywhere, therefore Λ_k is not empty for all k a.s.

Lemma 7. $\exists$ *a set* $\widetilde{\mathcal{N}}_0$ *which is a collection of finite many hyperplanes in* $\mathbb{R}^n$. $\forall \mathbf{y} \in \mathbb{R}^n \setminus \widetilde{\mathcal{N}}_0$,

$$|\mathcal{W}_m(\mathbf{y})| = 1 \quad \textit{and} \quad |q_m\mathcal{V}_m(\mathbf{y})| \le 1 \quad \forall\ m = 0, 1, \ldots, K(\mathbf{y}). \tag{42}$$

Corollary 1. *$\forall\ \mathbf{y} \in \mathbb{R}^n \setminus \widetilde{\mathcal{N}}_0$, Λ_k is not empty for all k, $k = 0, 1, \ldots, p$.*

Proof. This is a direct consequence of Lemma 7 and (39), (40). □

The next theorem presents an expression for the Lasso fit at each transition point, which helps us compute the divergence of $\hat{\boldsymbol{\mu}}_{m_k}(\mathbf{y})$.

Theorem 3. *Let $\hat{\boldsymbol{\mu}}_m(\mathbf{y})$ be the Lasso fit at the transition point λ_m, $\lambda_m > 0$. Then for any $i \in \mathcal{W}_m$, we can write $\hat{\boldsymbol{\mu}}(m)$ as follows*

$$\hat{\boldsymbol{\mu}}_m(\mathbf{y}) = \left\{ \mathbf{H}_{\mathcal{B}(\lambda_m)} - \frac{\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{X}_{\mathcal{B}(\lambda_m)} \right) \mathrm{Sgn}(\lambda_m) \mathbf{x}_i^T (\mathbf{I} - \mathbf{H}_{\mathcal{B}(\lambda_m)})}{\mathrm{Sgn}_i - \mathbf{x}_i^T \mathbf{X}^T_{\mathcal{B}(\lambda_m)} \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{X}_{\mathcal{B}(\lambda_m)} \right) \mathrm{Sgn}(\lambda_m)} \right\} \mathbf{y} \tag{43}$$

$$=: \mathbf{S}_m(\mathbf{y})\mathbf{y} \tag{44}$$

where $\mathbf{H}_{\mathcal{B}(\lambda_m)}$ is the projection matrix on the subspace of $\mathbf{X}_{\mathcal{B}(\lambda_m)}$. Moreover

$$\mathrm{tr}\left(\mathbf{S}_m(\mathbf{y})\right) = |\mathcal{B}(\lambda_m)|. \tag{45}$$

Proof. Note that $\hat{\beta}(\lambda)$ is continuous on λ. Using (18) in Lemma 2 and taking the limit of $\lambda \to \lambda_m$, we have

$$-2\mathbf{x}_j^T \left(\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \hat{\beta}(\lambda_m)_j \right) + \lambda_m \ \mathrm{Sgn}(\hat{\beta}(\lambda_m)_j) = 0, \quad \text{for } j \in \mathcal{B}(\lambda_m). \tag{46}$$

However, $\sum_{j=1}^{p} \mathbf{x}_j \hat{\beta}(\lambda_m)_j = \sum_{j \in \mathcal{B}(\lambda_m)} \mathbf{x}_j \hat{\beta}(\lambda_m)_j$. Thus we have

$$\hat{\beta}(\lambda_m) = \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{X}_{\mathcal{B}(\lambda_m)} \right)^{-1} \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{y} - \frac{\lambda_m}{2} \ \mathrm{Sgn}(\lambda_m) \right). \tag{47}$$

Hence

$$\begin{aligned} \hat{\boldsymbol{\mu}}_m(\mathbf{y}) &= \mathbf{X}_{\mathcal{B}(\lambda_m)} \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{X}_{\mathcal{B}(\lambda_m)} \right)^{-1} \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{y} - \frac{\lambda_m}{2} \ \mathrm{Sgn}(\lambda_m) \right) \\ &= \mathbf{H}_{\mathcal{B}(\lambda_m)} \mathbf{y} - \mathbf{X}_{\mathcal{B}(\lambda_m)} \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{X}_{\mathcal{B}(\lambda_m)} \right)^{-1} \mathrm{Sgn}(\lambda_m) \ \frac{\lambda_m}{2}. \end{aligned} \tag{48}$$

Since $i \in \mathcal{W}_m$, we must have the *equiangular* condition

$$\mathrm{Sgn}_i \ \mathbf{x}_i^T \left(\mathbf{y} - \hat{\boldsymbol{\mu}}(m) \right) = \frac{\lambda_m}{2}. \tag{49}$$

Substituting (48) into (49), we solve $\frac{\lambda_m}{2}$ and obtain

$$\frac{\lambda_m}{2} = \frac{\mathbf{x}_i^T \left(\mathbf{I} - \mathbf{H}_{\mathcal{B}(\lambda_m)} \right) \mathbf{y}}{\mathrm{Sgn}_i - \mathbf{x}_i^T \mathbf{X}^T_{\mathcal{B}(\lambda_m)} \left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)} \mathbf{X}_{\mathcal{B}(\lambda_m)} \right) \mathrm{Sgn}(\lambda_m)}. \tag{50}$$

Then putting (50) back to (48) yields (43).

Using the identity $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we observe

$$\begin{aligned}\mathrm{tr}\left(\mathbf{S}_m(\mathbf{y}) - \mathbf{H}_{\mathcal{B}(\lambda_m)}\right) &= \mathrm{tr}\left(\frac{\mathbf{X}^T_{\mathcal{B}(\lambda_m)}\left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)}\mathbf{X}_{\mathcal{B}(\lambda_m)}\right)\mathrm{Sgn}(\lambda_m)\mathbf{x}_i^T(\mathbf{I} - \mathbf{H}_{\mathcal{B}(\lambda_m)})}{\mathrm{Sgn}_i - \mathbf{x}_i^T\mathbf{X}^T_{\mathcal{B}(\lambda_m)}\left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)}\mathbf{X}_{\mathcal{B}(\lambda_m)}\right)\mathrm{Sgn}(\lambda_m)}\right) \\ &= \mathrm{tr}\left(\frac{\left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)}\mathbf{X}_{\mathcal{B}(\lambda_m)}\right)\mathrm{Sgn}(\lambda_m)\mathbf{x}_i^T(\mathbf{I} - H_{\mathcal{B}(\lambda_m)})\mathbf{X}^T_{\mathcal{B}(\lambda_m)}}{\mathrm{Sgn}_i - \mathbf{x}_i^T\mathbf{X}^T_{\mathcal{B}(\lambda_m)}\left(\mathbf{X}^T_{\mathcal{B}(\lambda_m)}\mathbf{X}_{\mathcal{B}(\lambda_m)}\right)\mathrm{Sgn}(\lambda_m)}\right) \\ &= \mathrm{tr}\,(0) = 0.\end{aligned}$$

So $\mathrm{tr}\,(\mathbf{S}_m(\mathbf{y})) = \mathrm{tr}\left(\mathbf{H}_{\mathcal{B}(\lambda_m)}\right) = |\mathcal{B}(\lambda_m)|$. □

Definition 1. $\mathbf{y} \in \mathbb{R}^n \setminus \widetilde{\mathcal{N}}_0$ *is said to be a locally stable point for* Λ_k*, if* $\forall$ $\mathbf{y}'$ *such that* $\|\mathbf{y}' - \mathbf{y}\| \leq \epsilon(\mathbf{y})$ *for a small enough* $\epsilon(\mathbf{y})$*, the effective set* $\mathcal{B}_{\lambda_m}(\mathbf{y}') = \mathcal{B}_{\lambda_m}(\mathbf{y})$*, for all* $m \in \Lambda_k$*. Let* $LS(\Lambda_k)$ *be the set of all locally stable points for* Λ_k*.*

Theorem 4. *If* $\mathbf{y} \in LS(\Lambda_k)$*, then we have the divergence formula* $\nabla \cdot \hat{\boldsymbol{\mu}}_m(\mathbf{y}) = k$ *which holds for all* $m \in \Lambda_k$ *including* $m_k = \sup(\Lambda_k)$*, the choice in the conjecture.*

Proof. The conclusion immediately follows definition 1 and Theorem 3. □

Points in $LS(\Lambda_k)$ are the majority of $\mathbb{R}^n$. Under the positive cone condition, $LS(\Lambda_k)$ is a set of full measure for all k. In fact the positive cone condition implies a stronger conclusion.

Definition 2. $\mathbf{y}$ *is said to be a strong locally stable point if* $\forall$ $\mathbf{y}'$ *such that* $\|\mathbf{y}' - \mathbf{y}\| \leq \epsilon(\mathbf{y})$ *for a small enough* $\epsilon(\mathbf{y})$*, the effective set* $\mathcal{B}_{\lambda_m}(\mathbf{y}') = \mathcal{B}_{\lambda_m}(\mathbf{y})$*, for all* $m = 0, 1, \ldots, K(\mathbf{y})$*.*

Lemma 8. *Let* $\widetilde{\mathcal{N}}_1 = \left\{\mathbf{y} : \widehat{\gamma}(\mathbf{y}) = \widetilde{\gamma}(\mathbf{y}) \quad \text{for some } m, m \in \{0, 1, \ldots, K(\mathbf{y})\}\right\}$*.* $\forall$ $\mathbf{y} \in$ *the interior of* $\mathbb{R}^n \setminus (\widetilde{\mathcal{N}}_0 \cup \widetilde{\mathcal{N}}_1)$*,* $\mathbf{y}$ *is a strong locally stable point. In particular, the positive cone condition implies* $\widetilde{\mathcal{N}}_1 = \varnothing$*.*

LARS is a discrete procedure by its definition, but the Lasso is a continuous shrinkage method. So it also makes sense to talk about fractional Lasso steps in the LARS algorithm, e.g. what is the Lasso fit at 3.5 steps? Under the positive cone condition, we can generalize the result of Theorem 4 in the LAR paper to the case of non-integer steps.

Corollary 2. *Under the positive cone condition* $df(\hat{\boldsymbol{\mu}}_s) = s$ *for all real valued* s*:* $0 \leq s \leq p$*.*

Proof. Let $k \leq s < k+1$, $s = k + r$ for some $r \in [0, 1)$. According to the LARS algorithm, the Lasso fit is linearly interpolated between steps k and $k+1$. So $\hat{\boldsymbol{\mu}}_s = \hat{\boldsymbol{\mu}}_k \cdot (1-r) + \hat{\boldsymbol{\mu}}_{k+1} \cdot r$. Then by definition (5) and the fact cov is a linear operator, we have

$$\begin{aligned} df(\hat{\boldsymbol{\mu}}_s) &= df(\hat{\boldsymbol{\mu}}_k) \cdot (1-r) + df(\hat{\boldsymbol{\mu}}_{k+1}) \cdot r \\ &= k \cdot (1-r) + (k+1) \cdot r = s. \end{aligned} \tag{51}$$

In (51) we have used the positive cone condition and Theorem 4 in the LAR paper. □

4 Adaptive Lasso Shrinkage

4.1 Model selection criteria

For any regularization method an important issue is to find a good choice of the regularization parameter such that the corresponding model is optimal according to some criterion, e.g. minimizing the prediction risk. For this purpose, model selection criteria have been proposed in the literature to compare different models. Famous examples are AIC (Akaike 1973) and BIC (Schwartz 1978). Mallows's C_p (Mallows 1973) is very similar to AIC and a whole class of AIC or C_p-type criteria are provided by SURE theory (Stein 1981). In Efron (2004) C_p and SURE are summarized as covariance penalty methods for estimating the prediction error, and are shown to offer substantially better accuracy than cross-validation and related nonparametric methods, if one is willing to assume the model is correct.

In the previous section we have derived the degrees of freedom of the Lasso for both types of regularization: λ and m_k. Although the exact value of $df(\lambda)$ is still unknown, our formula provides a convenient unbiased estimate. In the spirit of SURE theory, the unbiased estimate for $df(\lambda)$ suffices to provide an unbiased estimate for the prediction error of $\hat{\boldsymbol{\mu}}_\lambda$. If we choose m_k as the regularization parameter, the good approximation $df(\hat{\boldsymbol{\mu}}_{m_k}) \doteq k$ also well serves the SURE purpose. Therefore an estimate for the prediction error of $\hat{\boldsymbol{\mu}}$ ($pe(\hat{\boldsymbol{\mu}})$) is

$$\widehat{pe}(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{n} + \frac{2}{n}\widehat{df}(\hat{\boldsymbol{\mu}})\ \sigma^2, \tag{52}$$

where $\widehat{df}$ is either $\widehat{df}(\lambda)$ or $\widehat{df}(m_k)$, depending on the type of regularization. When σ^2 is unknown, it is usually replaced with an estimate based on the largest model.

Equation (52) equivalently derives AIC for the Lasso

$$\text{AIC}(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{n\sigma^2} + \frac{2}{n}\ \widehat{df}(\hat{\boldsymbol{\mu}}). \tag{53}$$

Selecting the Lasso model by AIC is called *AIC-Lasso shrinkage*. Following the usual definition of BIC, we propose BIC for the Lasso as follows

$$\text{BIC}(\hat{\boldsymbol{\mu}}) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{n\sigma^2} + \frac{\log(n)}{n}\ \widehat{df}(\hat{\boldsymbol{\mu}}). \tag{54}$$

Similarly the Lasso model selection by BIC is called *BIC-Lasso shrinkage.*

AIC and BIC possess different asymptotic optimality. It is well known that if the true regression function is not in the candidate models, the model selected by AIC asymptotically achieves the smallest average squared error among the candidates; and the AIC estimator of the regression function converges at the minimax optimal rate whether the true regression function is in the candidate models or not, see Shao (1997), Yang (2003) and references therein. On the other hand, BIC is well known for its consistency in selecting the true model (Shao 1997). If the true model is in the candidate list, the probability of selecting the true model by BIC approaches one as the sample size $n \to \infty$. Considering the case where the true underlying model is sparse, BIC-Lasso shrinkage is adaptive in variable selection.

n	AIC	BIC
100	0.162	0.451
500	0.181	0.623
1000	0.193	0.686
2000	0.184	0.702

Table 1: *The simulation example: the probability of discovering the exact true model by AIC and BIC Lasso shrinkage. The calculation is based on 2000 replications. Compared with AIC-Lasso shrinkage, BIC-Lasso shrinkage has a much higher probability of identifying the ground truth.*

n	AIC	BIC
100	5	4
500	5	3
1000	5	3
2000	5	3

Table 2: *The simulation example: the median of the number of non-zero variables selected by AIC and BIC Lasso shrinkage based on 2000 replications. One can see that AIC-Lasso shrinkage is conservative in variable selection and BIC-Lasso shrinkage tends to find models with the right size.*

However, AIC-Lasso shrinkage tends to include more non-zero predictors than the truth. The conservative nature of AIC is a familiar result in linear regression. Hence BIC-Lasso shrinkage is more appropriate than AIC-Lasso shrinkage when variable selection is the primary concern in applying the Lasso.

Here we show a simulation example to demonstrate the above argument. We simulated response vectors $\mathbf{y}$ from a linear model: $\mathbf{y} = \mathbf{X}\beta + N(0,1)$ where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$. Predictors $\{\mathbf{x}_i\}$ are multivariate normal vectors with pairwise correlation $\text{cor}(i,j) = (0.1)^{|i-j|}$ and the variance of each $\mathbf{x}_i$ is one. For each estimate $\hat{\beta}$, it is said to discover the exact true model if $\{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_5\}$ are non-zero and the rest coefficients are all zero. Table 1 shows the probability of discovering the exact true model using AIC-Lasso shrinkage and BIC-Lasso shrinkage. In this example both AIC and BIC always select the true predictors $\{1, 2, 5\}$ in all the 2000 replications, but AIC tends to include other variables as real factors as shown in Table 2. In contrast to AIC, BIC has a much lower false positive rate.

One may think of combining the good properties of AIC and BIC into a new criterion. Although this proposal sounds quite reasonable, a surprising result is proved that any model selection criterion cannot be consistent and optimal in average squared error at the same time (Yang 2003). In other words, any model selection criterion must sacrifice either prediction optimality or consistency.

4.2 Computation

Using either AIC or BIC to find the optimal Lasso model, we are facing an optimization problem

$$\lambda(\text{optimal}) = \arg\min_{\lambda} \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda\|^2}{n\sigma^2} + \frac{w_n}{n}\,\widehat{df}(\lambda), \tag{55}$$

where $w_n = 2$ for AIC and $w_n = \log(n)$ for BIC. Since the LARS algorithm efficiently solves the Lasso solution for all λ, finding $\lambda(\text{optimal})$ is attainable in principle. In fact, we show that $\lambda(\text{optimal})$ is one of the transition points, which further facilitates the searching procedure.

Theorem 5. *To find* $\lambda(\text{optimal})$*, we only need to solve*

$$m^* = \arg\min_{m} \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\lambda_m}\|^2}{n\sigma^2} + \frac{w_n}{n}\,\widehat{df}(\lambda_m) \tag{56}$$

then $\lambda(\text{optimal}) = \lambda_{m^*}$*.*

Proof. Let us consider $\lambda \in (\lambda_{m+1}, \lambda_m)$. By (21) we have

$$\begin{aligned} \mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda &= (\mathbf{I} - H_{\mathcal{B}_m})\mathbf{y} + \frac{\lambda}{2}\,\mathbf{X}_{\mathcal{B}_m}(\mathbf{X}_{\mathcal{B}_m}^T\mathbf{X}_{\mathcal{B}_m})^{-1}\text{Sgn}_m & (57)\\ \|\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda\|^2 &= \mathbf{y}^T(\mathbf{I} - H_{\mathcal{B}_m})\mathbf{y} + \frac{\lambda^2}{4}\,\text{Sgn}_m^T(\mathbf{X}_{\mathcal{B}_m}^T\mathbf{X}_{\mathcal{B}_m})^{-1}\text{Sgn}_m & (58)\end{aligned}$$

where $H_{\mathcal{B}_m} = \mathbf{X}_{\mathcal{B}_m}(\mathbf{X}_{\mathcal{B}_m}^T\mathbf{X}_{\mathcal{B}_m})^{-1}\mathbf{X}_{\mathcal{B}_m}^T$. Thus we can conclude that $\|\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda\|^2$ is strictly increasing in the interval $(\lambda_{m+1}, \lambda_m)$. Moreover, the Lasso estimates are continuous on λ, hence

$$\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\lambda_m}\|^2 > \|\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda\|^2 > \|\mathbf{y} - \hat{\boldsymbol{\mu}}_{\lambda_{m+1}}\|^2. \tag{59}$$

On the other hand, note that $\widehat{df}(\lambda) = |\mathcal{B}_m|\ \forall\ \lambda \in (\lambda_{m+1}, \lambda_m)$ and $|\mathcal{B}_m| \geq |\mathcal{B}(\lambda_{m+1})|$. Therefore the optimal choice of λ in $[\lambda_{m+1}, \lambda_m)$ is λ_{m+1}, which means $\lambda(\text{optimal}) \in \{\lambda_m\}, m = 0, 1, 2, \ldots, K$. □

According to Theorem 5, the optimal Lasso model is immediately selected once the entire Lasso solution path is solved by the LARS algorithm, which has the cost of a single least squares fit.

If we consider the best Lasso fit in the forward stage-wise modeling picture (like Figure 2), inequality (59) explains the superiority of the choice of m_k in the conjecture. Let m_k be the last Lasso step containing k non-zero predictors. Suppose m_k' is another Lasso step containing k non-zero predictors, then $\widehat{df}(\hat{\boldsymbol{\mu}}(m_k')) = k = \widehat{df}(\hat{\boldsymbol{\mu}}(m_k))$. However, $m_k' < m_k$ gives $\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{m_k}\|^2 < \|\mathbf{y} - \hat{\boldsymbol{\mu}}_{m_k'}\|^2$. Then by the C_p statistic, we see that $\hat{\boldsymbol{\mu}}(m_k)$ is always more accurate than $\hat{\boldsymbol{\mu}}(m_k')$, while using the same number of non-zero predictors. Using k as the tuning parameter of the Lasso, we need to find $k(\text{optimal})$ such that

$$k(\text{optimal}) = \arg\min_{k} \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}_{m_k}\|^2}{n\sigma^2} + \frac{w_n}{n}k. \tag{60}$$

Once $\lambda^* = \lambda(\text{optimal})$ and $k^* = k(\text{optimal})$ are found, we fix them as the regularization parameters for fitting the Lasso on future data. Using the fixed k^* means the fit on future data is $\hat{\boldsymbol{\mu}}_{m_{k^*}}$, while the fit using the fixed λ^* is $\hat{\boldsymbol{\mu}}_{\lambda^*}$. It is easy to see that the selected models by (55) and (60) coincide on the training data, i.e., $\hat{\boldsymbol{\mu}}_{\lambda^*} = \hat{\boldsymbol{\mu}}_{m_{k^*}}$.

Figure 6 displays the C_p (equivalently AIC) and BIC estimates of risk using the diabetes data. The models selected by C_p are the same as those selected in the LAR paper. With 10 predictors, C_p and BIC select the same model using 7 non-zero covariates. With 64 predictors, C_p selects a model using 15 covariates, while BIC selects a model with 11 covariates.

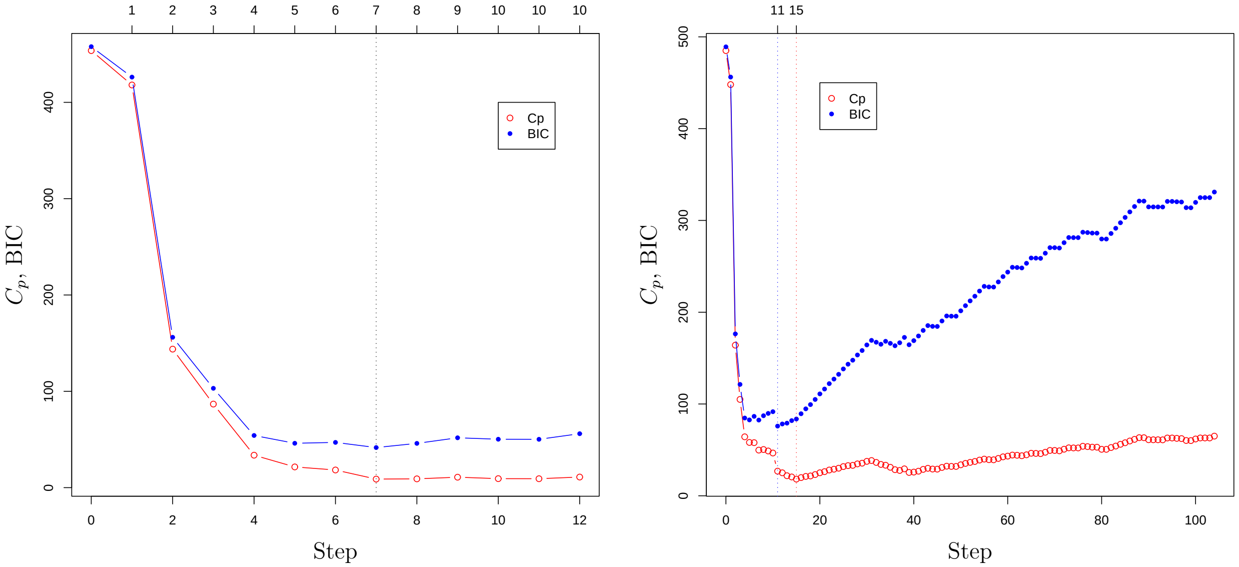


Figure 6: *The diabetes data. C_p and BIC estimates of risk with 10 (left) and 64 (right) predictors. In the left panel C_p and BIC select the same model with 7 non-zero coefficients. In the right panel, C_p selects a model with 15 non-zero coefficients and BIC selects a model with 11 non-zero coefficients.*

5 Discussion

It is interesting to note that the true degrees of freedom is a strictly decreasing function of λ, as shown in Figure 7. However, the unbiased estimate $\widehat{df}(\lambda)$ is not necessarily monotone, although its global trend is monotonically decreasing. The same phenomenon is also shown in the right panel of Figure 1. The non-monotonicity of $\widehat{df}(\lambda)$ is due to the fact that some variables can be dropped during the LARS/Lasso process.

An interesting question is that whether there is a smoothed estimate $\widehat{df}^*(\lambda)$ such that $\widehat{df}^*(\lambda)$ is a smooth decreasing function and keeps the unbiased property, i.e.,

$$df(\lambda) = E[\widehat{df}^*(\lambda)] \tag{61}$$

holds for all λ. This is a future research topic.

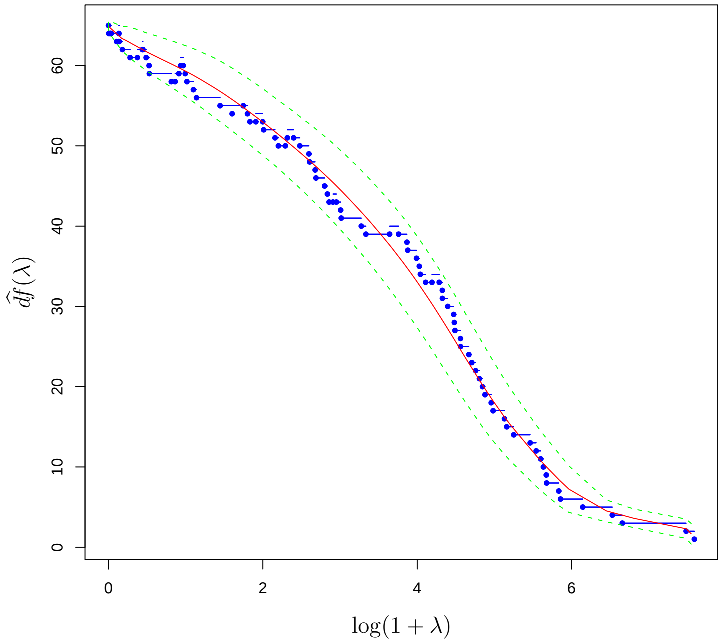


Figure 7: *The dotted line is the curve of estimated degrees of freedom ($\widehat{df}(\lambda)$ vs. $\log(1+\lambda)$), using a typical realization $\mathbf{y}^*$ generated by the synthetic model (16). The smooth curve shows the true degrees of freedom $df(\lambda)$ obtained by averaging* 20000 *estimated curves. One can see that the estimated df curve is piece-wise constant and non-monotone, while the true df curve is smooth and monotone. The two thin broken lines correspond to $df(\lambda)\pm 2\sqrt{Var(\widehat{df}(\lambda))}$, where $Var(\widehat{df}(\lambda))$ is calculated from the $B = 20000$ replications.*

6 Appendix: proofs of lemmas 2-8

Proof. Lemma 2

Let

$$\ell(\beta, \mathbf{y}) = \|\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|\,. \tag{62}$$

Given $\mathbf{y}$, $\hat{\beta}(\lambda)$ is the minimizer of $\ell(\beta, \mathbf{y})$. For those $j \in \mathcal{B}_m$ we must have $\frac{\partial \ell(\beta,\mathbf{y})}{\partial \beta_j} = 0$, i.e.,

$$-2\mathbf{x}_j^T \left(\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \hat{\beta}(\lambda)_j \right) + \lambda\ \mathrm{Sgn}(\hat{\beta}(\lambda)_j) = 0, \quad \text{for } j \in \mathcal{B}_m. \tag{63}$$

Since $\hat{\beta}(\lambda)_i = 0$ for all $i \notin \mathcal{B}_m$, then $\sum_{j=1}^{p} \mathbf{x}_j \hat{\beta}(\lambda)_j = \sum_{j \in \mathcal{B}_\lambda} \mathbf{x}_j \hat{\beta}(\lambda)_j$. Thus equations in (63) become

$$-2\mathbf{X}_{\mathcal{B}_m}^T \left(\mathbf{y} - \mathbf{X}_{\mathcal{B}_m} \hat{\beta}(\lambda)_{\mathcal{B}_m} \right) + \lambda \ \mathrm{Sgn}_m = 0 \tag{64}$$

which gives (21). □

Proof. Lemma 3

We adopt the matrix notation used in $\texttt{S}$: $\mathbf{M}[i,]$ means the i-th row of $\mathbf{M}$. i_{add} joins $\mathcal{B}_m$ at λ_m, then

$$\hat{\beta}(\lambda_m)_{i_{add}} = 0. \tag{65}$$

Consider $\hat{\beta}(\lambda)$ for $\lambda \in (\lambda_{m+1}, \lambda_m)$. Lemma 2 gives

$$\hat{\beta}(\lambda)_{\mathcal{B}_m} = \left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} \left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{y} - \frac{\lambda}{2} \ \mathrm{Sgn}_m \right). \tag{66}$$

By the continuity of $\hat{\beta}(\lambda)_{i_{add}}$, taking the limit of the i^*-th element of (66) as $\lambda \to \lambda_m - 0$, we have

$$2 \Bigg\{ \left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [i^*,] \mathbf{X}_{\mathcal{B}_m}^T \Bigg\} \mathbf{y} = \lambda_m \Bigg\{ \left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [i^*,] \mathrm{Sgn}_m \Bigg\}. \tag{67}$$

The second $\{\cdot\}$ is a non-zero scalar, otherwise $\hat{\beta}(\lambda)_{i_{add}} = 0$ for all $\lambda \in (\lambda_{m+1}, \lambda_m)$, which contradicts the assumption that i_{add} becomes a member of the active set $\mathcal{B}_m$. Thus we have

$$\lambda_m = \left\{ 2 \frac{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [i^*,]}{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [i^*,] \mathrm{Sgn}_m} \right\} \mathbf{X}_{\mathcal{B}_m}^T \mathbf{y} =: v(\mathcal{B}_m, i^*) \mathbf{X}_{\mathcal{B}_m}^T \mathbf{y}, \tag{68}$$

where $v(\mathcal{B}_m, i^*) = \left\{ 2 \frac{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [i^*,]}{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [i^*,] \mathrm{Sgn}_m} \right\}$. Rearranging (68), we get (22).

Similarly, if j_{drop} is a dropped index at λ_{m+1}, we take the limit of the j^*-th element of (66) as $\lambda \to \lambda_{m+1} + 0$ to conclude that

$$\lambda_{m+1} = \left\{ 2 \frac{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [j^*,]}{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [j^*,] \mathrm{Sgn}_m} \right\} \mathbf{X}_{\mathcal{B}_m}^T \mathbf{y} =: v(\mathcal{B}_m, j^*) \mathbf{X}_{\mathcal{B}_m}^T \mathbf{y}, \tag{69}$$

where $v(\mathcal{B}_m, j^*) = \left\{ 2 \frac{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [j^*,]}{\left(\mathbf{X}_{\mathcal{B}_m}^T \mathbf{X}_{\mathcal{B}_m} \right)^{-1} [j^*,] \mathrm{Sgn}_m} \right\}$. Rearranging (69), we get (23). □

Proof. Lemma 4

Suppose for some $\mathbf{y}$ and m, $\lambda = \lambda(\mathbf{y})_m$. $\lambda > 0$ means m is not the last Lasso step. By Lemma 3 we have

$$\lambda = \lambda_m = \{ v(\mathcal{B}_m, i^*) \mathbf{X}_{\mathcal{B}_m}^T \} \mathbf{y} =: \alpha(\mathcal{B}_m, i^*) \mathbf{y}. \tag{70}$$

Obviously $\alpha(\mathcal{B}_m, i^*) = v(\mathcal{B}_m, i^*) \mathbf{X}_{\mathcal{B}_m}^T$ is a non-zero vector. Now let α_λ be the totality of $\alpha(\mathcal{B}_m, i^*)$ by considering all the possible combinations of $\mathcal{B}_m$, i^* and the sign vector Sgn_m.

α_λ only depends on $\mathbf{X}$ and is a finite set, since at most p predictors are available. Thus $\forall \alpha \in \alpha_\lambda$, $\alpha \mathbf{y} = \lambda$ defines a hyperplane in $\mathbb{R}^n$. We define

$$\mathcal{N}_\lambda = \{\mathbf{y} : \alpha \mathbf{y} = \lambda \text{ for some } \alpha \in \alpha_\lambda\} \quad \text{and} \quad \mathcal{G}_\lambda = \mathbb{R}^n \setminus \mathcal{N}_\lambda.$$

Then on $\mathcal{G}_\lambda$ (70) is impossible.

□

Proof. Lemma 5

For writing convenience we omit the subscript λ. Let

$$\hat{\beta}(\mathbf{y})_{ols} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{71}$$

be the OLS estimates. Note that we always have the inequality

$$|\hat{\beta}(\mathbf{y})|_1 \leq |\hat{\beta}(\mathbf{y})_{ols}|_1 \ . \tag{72}$$

Fix an arbitrary $\mathbf{y}_0$ and consider a sequence of $\{\mathbf{y}_n\}$ $(n = 1, 2, \ldots)$ such that $\mathbf{y}_n \to \mathbf{y}_0$. Since $\mathbf{y}_n \to \mathbf{y}_0$, we can find a Y such that $\|\mathbf{y}_n\| \leq Y$ for all $n = 0, 1, 2, \ldots$. Consequently $\|\hat{\beta}(\mathbf{y}_n)_{ols}\| \leq B$ for some upper bound B (B is determined by $\mathbf{X}$ and Y). By Cauchy's inequality and (72), we have

$$|\hat{\beta}(\mathbf{y}_n)|_1 \leq \sqrt{p B} \quad \text{for all } n = 0, 1, 2, \ldots \tag{73}$$

(73) implies that to show $\hat{\beta}(\mathbf{y}_n) \to \hat{\beta}(\mathbf{y}_0)$, it is equivalent to show for every converging subsequence of $\{\hat{\beta}(\mathbf{y}_n)\}$, say $\{\hat{\beta}(\mathbf{y}_{n_k})\}$, the subsequence converge to $\hat{\beta}(\mathbf{y})$.

Now assume $\hat{\beta}(\mathbf{y}_{n_k})$ converges to $\hat{\beta}_\infty$ as $n_k \to \infty$. We show $\hat{\beta}_\infty = \hat{\beta}(\mathbf{y}_0)$. The Lasso criterion $\ell(\beta, \mathbf{y})$ is written in (62). Let

$$\Delta\ell(\beta, \mathbf{y}, \mathbf{y}') = \ell(\beta, \mathbf{y}) - \ell(\beta, \mathbf{y}'). \tag{74}$$

By the definition of $\hat{\beta}_{n_k}$, we must have

$$\ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_{n_k}) \geq \ell(\hat{\beta}(\mathbf{y}_{n_k}), \mathbf{y}_{n_k}). \tag{75}$$

Then (75) gives

$$\begin{aligned} \ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_0) &= \ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_{n_k}) + \Delta\ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{n_k}) \\ &\geq \ell(\hat{\beta}(\mathbf{y}_{n_k}), \mathbf{y}_{n_k}) + \Delta\ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{n_k}) \\ &= \ell(\hat{\beta}(\mathbf{y}_{n_k}), \mathbf{y}_0) + \Delta\ell(\hat{\beta}(\mathbf{y}_{n_k}), \mathbf{y}_{n_k}, \mathbf{y}_0) + \Delta\ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{n_k}). \end{aligned} \tag{76}$$

We observe

$$\Delta\ell(\hat{\beta}(\mathbf{y}_{n_k}), \mathbf{y}_{n_k}, \mathbf{y}_0) + \Delta\ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_0, \mathbf{y}_{n_k}) = 2(\mathbf{y}_0 - \mathbf{y}_{n_k})\mathbf{X}^T(\hat{\beta}(\mathbf{y}_{n_k}) - \hat{\beta}(\mathbf{y}_0)). \tag{77}$$

Let $n_k \to \infty$, the right hand side of (77) goes to zero. Moreover, $\ell(\hat{\beta}(\mathbf{y}_{n_k}), \mathbf{y}_0) \to \ell(\hat{\beta}_\infty, \mathbf{y}_0)$. Therefore (76) reduces to

$$\ell(\hat{\beta}(\mathbf{y}_0), \mathbf{y}_0) \geq \ell(\hat{\beta}_\infty, \mathbf{y}_0).$$

However, $\hat{\beta}(\mathbf{y}_0)$ is the unique minimizer of $\ell(\beta, \mathbf{y}_0)$, thus $\hat{\beta}_\infty = \hat{\beta}(\mathbf{y}_0)$.

□

Proof. Lemma 6

Fix an arbitrary $\mathbf{y}_0 \in \mathcal{G}_\lambda$. Denote $\text{Ball}(\mathbf{y}, r)$ the n-dimensional ball with center $\mathbf{y}$ and radius r. Note that $\mathcal{G}_\lambda$ is an open set, so we can choose a small enough ϵ such that $\text{Ball}(\mathbf{y}_0, \epsilon) \subset \mathcal{G}_\lambda$. Fix ϵ. Suppose $\mathbf{y}_n \to \mathbf{y}$ as $n \to \infty$, then without loss of generality we can assume $\mathbf{y}_n \in \text{Ball}(\mathbf{y}_0, \epsilon)$ for all n. So λ is not a transition point for any $\mathbf{y}_n$.

By definition $\hat{\beta}(\mathbf{y}_0)_j \neq 0$ for all $j \in \mathcal{B}(\mathbf{y}_0)$. Then Lemma 5 says that $\exists$ a N, as long as $n > N_1$, we have $\hat{\beta}(\mathbf{y}_n)_j \neq 0$ and $\text{Sgn}(\hat{\beta}(\mathbf{y}_n)) = \text{Sgn}(\hat{\beta}(\mathbf{y}_n))$, for all $j \in \mathcal{B}(\mathbf{y}_0)$. Thus $\mathcal{B}(\mathbf{y}_0) \subseteq \mathcal{B}(\mathbf{y}_n)$ $\forall n > N_1$.

On the other hand, we have the following *equiangular* conditions (Efron et al. 2004)

$$\begin{aligned} \lambda &= 2|\mathbf{x}_j^T(\mathbf{y}_0 - \mathbf{X}\hat{\beta}(\mathbf{y}_0))| \quad \forall\ j \in \mathcal{B}(\mathbf{y}_0), & (78) \\ \lambda &> 2|\mathbf{x}_j^T(\mathbf{y}_0 - \mathbf{X}\hat{\beta}(\mathbf{y}_0))| \quad \forall\ j \notin \mathcal{B}(\mathbf{y}_0). & (79) \end{aligned}$$

Using Lemma 5 again, we conclude that $\exists$ a $N > N_1$ such that $\forall\ j \notin \mathcal{B}(\mathbf{y}_0)$ the strict inequalities (79) hold for $\mathbf{y}_n$ provided $n > N$. Thus $\mathcal{B}^c(\mathbf{y}_0) \subseteq \mathcal{B}^c(\mathbf{y}_n)$ $\forall n > N$. Therefore we have $\mathcal{B}(\mathbf{y}_n) = \mathcal{B}(\mathbf{y}_0)$ $\forall n > N$. Then the local constancy of the sign vector follows the continuity of $\hat{\beta}(\mathbf{y})$. □

Proof. Lemma 7

Suppose at step m, $|\mathcal{W}_m(\mathbf{y})| \geq 2$. Let i_{add} and j_{add} be two of the predictors in $\mathcal{W}_m(\mathbf{y})$, and let i^*_{add} and j^*_{add} be their indices in the current active set $\mathcal{A}$. Note the current active set $\mathcal{A}$ is $\mathcal{B}_m$ in Lemma 3. Hence we have

$$\begin{aligned} \lambda_m &= v[\mathcal{A}, i^*]\mathbf{X}_{\mathcal{A}}^T\mathbf{y}, & (80) \\ \lambda_m &= v[\mathcal{A}, j^*]\mathbf{X}_{\mathcal{A}}^T\mathbf{y}. & (81) \end{aligned}$$

Therefore

$$0 = \left\{ [v(\mathcal{A}, i^*_{add}) - v(\mathcal{A}, j^*_{add})]\mathbf{X}_{\mathcal{A}}^T \right\} \mathbf{y} =: \alpha_{add}\mathbf{y}. \tag{82}$$

We claim $\alpha_{add} = [v(\mathcal{A}, i^*_{add}) - v(\mathcal{A}, j^*_{add})]\mathbf{X}_{\mathcal{A}}^T$ is not a zero vector. Otherwise, since $\{\mathbf{X}_j\}$ are linearly independent, $\alpha_{add} = 0$ forces $v(\mathcal{A}, i^*_{add}) - v(\mathcal{A}, j^*_{add}) = 0$. Then we have

$$\frac{\left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\right)^{-1}[i^*,]}{\left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\right)^{-1}[i^*,]\text{Sgn}_{\mathcal{A}}} = \frac{\left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\right)^{-1}[j^*,]}{\left(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}\right)^{-1}[i^*,]\text{Sgn}_{\mathcal{A}}}, \tag{83}$$

which contradicts the fact $(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})^{-1}$ is a full rank matrix.

Similarly, if i_{drop} and j_{drop} are dropped predictors, then

$$0 = \left\{ [v(\mathcal{A}, i^*_{drop}) - v(\mathcal{A}, j^*_{drop})]\mathbf{X}_{\mathcal{A}}^T \right\} \mathbf{y} =: \alpha_{drop}\mathbf{y}, \tag{84}$$

and $\alpha_{drop} = [v(\mathcal{A}, i^*_{drop}) - v(\mathcal{A}, j^*_{drop})]\mathbf{X}_{\mathcal{A}}^T$ is a non-zero vector.

Let M_0 be the totality of α_{add} and α_{drop} by considering all the possible combinations of $\mathcal{A}$, (i_{add}, j_{add}), (i_{drop}, j_{drop}) and $\mathrm{Sgn}_{\mathcal{A}}$. Clearly M_0 is a finite set and only depends on $\mathbf{X}$. Let

$$\widetilde{\mathcal{N}}_0 = \Big\{ y : \alpha y = 0 \ \text{ for some } \ \alpha \in M_0 \Big\}. \tag{85}$$

Then on $\mathbb{R}^n \setminus \widetilde{\mathcal{N}}_0$, the conclusion holds. □

Proof. Lemma 8

First we can choose a sufficiently small ϵ^* such that $\forall\ \mathbf{y}' : \|\mathbf{y}' - \mathbf{y}\| < \epsilon^*$, $\mathbf{y}'$ is an interior point of $\mathbb{R}^n \setminus (\widetilde{\mathcal{N}}_0 \cup \widetilde{\mathcal{N}}_1)$. Suppose K is the last step of the Lasso solution given $\mathbf{y}$. We show that for each $m \leq K$, there is a $\epsilon_m < \epsilon^*$ such that $q_{m-1}\mathcal{V}_{m-1}$ and $\mathcal{W}_m$ are locally fixed in the $\mathrm{Ball}(\mathbf{y}, \epsilon_m)$; also λ_m and $\hat{\beta}_m$ are locally continuous in the $\mathrm{Ball}(\mathbf{y}, \epsilon_m)$.

We proceed by induction. For $m = 0$ we only need to verify the local constancy of $\mathcal{W}_0$. Lemma 7 says $\mathcal{W}_0(\mathbf{y}) = \{j\}$. By the definition of $\mathcal{W}$, we have $|\mathbf{x}_j^T \mathbf{y}| > |\mathbf{x}_i^T \mathbf{y}|$ for all $i \neq j$. Thus the strict inequality holds if $\mathbf{y}'$ is sufficiently close to $\mathbf{y}$, which implies $\mathcal{W}_0(\mathbf{y}') = \{j\} = \mathcal{W}_0(\mathbf{y})$.

Assuming the conclusion holds for m, we consider points in the $\mathrm{Ball}(\mathbf{y}, \epsilon_{m+1})$ with $\epsilon_{m+1} < \min_{\ell \leq m}\{\epsilon_\ell\}$. By the induction assumption, $A_m(\mathbf{y})$ is locally fixed since it only depends on $\{(q_\ell \mathcal{V}_\ell, \mathcal{W}_\ell), \ell \leq (m-1)\}$. $q_m\mathcal{V}_m = \varnothing$ is equivalent to $\widehat{\gamma}(\mathbf{y}) < \widetilde{\gamma}(\mathbf{y})$. Once A_m and $\mathcal{W}_m$ are fixed, both $\widehat{\gamma}(\mathbf{y})$ and $\widetilde{\gamma}(\mathbf{y})$ are continuous on $\mathbf{y}$. Thus if $\mathbf{y}'$ is sufficiently close to $\mathbf{y}$, the strict inequality still holds, which means $q_m(\mathbf{y}')\mathcal{V}_m(\mathbf{y}') = \varnothing$. If $q_m\mathcal{V}_m = \mathcal{V}_m$, then $\widehat{\gamma}(\mathbf{y}) > \widetilde{\gamma}(\mathbf{y})$ since the possibility of $\widehat{\gamma}(\mathbf{y}) = \widetilde{\gamma}(\mathbf{y})$ is ruled out. By Lemma 7, we let $\mathcal{V}_m(\mathbf{y}) = \{j\}$. By the definition of $\widetilde{\gamma}(\mathbf{y})$, we can see that if $\mathbf{y}'$ is sufficiently close to $\mathbf{y}$, $\mathcal{V}_m(\mathbf{y}') = \{j\}$, and $\widehat{\gamma}(\mathbf{y}') > \widetilde{\gamma}(\mathbf{y}')$ by continuity. So $q_m(\mathbf{y}')\mathcal{V}_m(\mathbf{y}') = \mathcal{V}_m(\mathbf{y}') = \mathcal{V}_m(\mathbf{y})$.

Then $\hat{\beta}_{m+1}$ and λ_{m+1} are locally continuous, because their updates are continuous on $\mathbf{y}$ once A_m,$\mathcal{W}_m$ and $q_m\mathcal{V}_m$ are fixed. Moreover, since $q_m\mathcal{V}_m$ is fixed, A_{m+1} is also locally fixed. Let $\mathcal{W}_{m+1}(\mathbf{y}) = \{j\}$ for some $j \in A_{m+1}^c$. Then we have

$$|\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}^T\hat{\beta}_{m+1}(\mathbf{y}))| > |\mathbf{x}_i^T(\mathbf{y} - \mathbf{X}^T\hat{\beta}_{m+1}(\mathbf{y})| \quad \forall\ i \neq j,\ i \in A_{m+1}^c$$

By the continuity argument, the above strict inequality holds for all $\mathbf{y}'$ provided $\|\mathbf{y}' - \mathbf{y}\| \leq \epsilon_{m+1}$ for a sufficiently small ϵ_{m+1}. So $\mathcal{W}_{m+1}(\mathbf{y}') = \{j\} = \mathcal{W}_{m+1}(\mathbf{y})$. In conclusion, we can choose a small enough ϵ_{m+1} to make sure that $q_m\mathcal{V}_m$ and $\mathcal{W}_{m+1}$ are locally fixed, and $\hat{\beta}_{m+1}$ and λ_{m+1} are locally continuous. □

7 Acknowledgements

We thank Brad Efron and Yuhong Yang for their helpful comments.

References

Akaike, H. (1973), ‘Information theory and an extension of the maximum likelihood principle’, *Second International Symposium on Information Theory* pp. 267–281.

Efron, B. (1986), ‘How biased is the apparent error rate of a prediction rule?’, *Journal of the American Statistical Association* **81**, 461–470.

Efron, B. (2004), ‘The estimation of prediction error: covariance penalties and cross-validation’, *Journal of the American Statistical Association* **99**, 619–632.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), ‘Least angle regression’, *Annals of Statistics* **32**, 407–499.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning; Data mining, Inference and Prediction*, Springer Verlag, New York.

Hoerl, A. & Kennard, R. (1988), Ridge regression, *in* ‘Encyclopedia of Statistical Sciences’, Vol. 8, Wiley, New York, pp. 129–136.

Mallows, C. (1973), ‘Some comments on c_p’, *Technometrics* **15**, 661–675.

Meyer, M. & Woodroofe, M. (2000), ‘On the degrees of freedom in shape-restricted regression’, *Annals of Statistcs* **28**, 1083–1104.

Schwartz, G. (1978), ‘Estimating the dimension of a model’, *Annals of Statistics* **6**, 461–464.

Shao, J. (1997), ‘An asymptotic theory for linear model selection (with discussion)’, *Statistica Sinica* **7**, 221–242.

Stein, C. (1981), ‘Estimation of the mean of a multivariate normal distribution’, *Annals of Statistics* **9**, 1135–1151.

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

Yang, Y. (2003), Can the strengths of AIC and BIC be shared? (submitted. http://www.stat.iastate.edu/preprint/articles/2003-10.pdf).