A sensitivity analysis of the adaptive lasso

Tathagata Basu*
Jochen Einbeck
and
Matthias C. M. Troffaes
Department of Mathematical Sciences, Durham University

August 31, 2019

Abstract

Sparse regression is an efficient statistical modelling technique which is of major relevance for high dimensional statistics. There are several ways of achieving sparse regression, the well-known lasso being one of them. However, lasso variable selection may not be consistent in selecting the true sparse model. Zou (2006) proposed an adaptive form of the lasso which overcomes this issue, and showed that data driven weights on the penalty term will result in a consistent variable selection procedure. We are interested in the case that the weights are informed by a prior execution of ridge regression. We carry out a sensitivity analysis of the Adaptive lasso through the power parameter γ of the weights, and demonstrate that, in effect, this parameter γ takes over the role of the usual lasso penalty parameter. In addition, we use the γ parameter as an input variable to obtain an error bound on the Adaptive lasso.

Keywords: Adaptive lasso, sensitivity analysis, oracle properties, variable selection, ridge regression.

*This work is funded by the European Commissions H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 722734.

1 Introduction

Let $\boldsymbol{X} \coloneqq (X_1, \cdots, X_p)$ with $X_j = (X_{j1}, \ldots X_{jn})^T$ for $1 \leq j \leq p$, and $Y = (Y_1, \ldots, Y_n)^T$. We can characterise their relation in the linear regression setting

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

where $\boldsymbol{\beta} \coloneqq (\beta_1, \cdots \beta_p)$ is a vector of regression coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$, with $\boldsymbol{I}_n$ denoting the n-dimensional identity matrix. We assume $\boldsymbol{X}$ and Y to be scaled to mean 0.

The least square method is the conventional way to estimate these regression coefficients. However, in high dimension (i.e $p > n$), the least square method, which involves inversion of $\boldsymbol{X}^T\boldsymbol{X}$, cannot be used. Several estimators have been proposed which solve the issue by introducing bias in the estimation process. Tikhonov (1963) introduced ℓ_2 penalised regression or Ridge regression. The ℓ_2 penalty achieves a stable solution through the eigen value decay method, which, however, fails to be sparse. Sparsity of estimates is a desirable property in high dimensional statistics. Tibshirani (1996) introduced the lasso or least absolute shrinkage and selection operator, which attains sparsity through a ℓ_1 penalty. Lasso is a well practised regression technique in high dimensional statistics. However, Fan & Li (2001) showed that lasso variable selection can be inconsistent as the ℓ_1 penalty term produces biased estimates for large effects. They introduced the idea of oracle properties for high dimensional problems which ensure the consistency in true model selection as well as asymptotic properties of the estimators. Later on, Zhao & Yu (2006) showed irrepresentable conditions (necessary conditions) for lasso consistency. Zou (2006) proposed an adaptive form of lasso based on data-driven weights in the penalty term. Geer & Bühlmann (2009), Van de Geer et al. (2011) gave restricted eigen value conditions for the lasso and provided an error bound for the adaptive lasso for misspecified models.

In this paper, we build on the framework given by Zou (2006) to investigate and under-

stand the sensitivity of the adaptive lasso with respect to the weight parameter γ. For this we apply a two-step approach. We use ridge estimates, say $\hat{\beta}_j$, to initialise the adaptive lasso, yielding weights of type $1/|\hat{\beta}_j|^\gamma$ which are then embedded in the penalty term. The resulting adaptive lasso estimates are considered as functions of the parameter γ. We use these estimates to obtain an error bound based on γ.

The rest of the paper is organised as follows; we first discuss some properties of the Adaptive Lasso in Section 2. We then discuss oracle properties and irrepresentable condition in Section 3. We provide our main result in Section 4. We illustrate our results for a simulated dataset in Section 5 and gaia data in Section 6. Finally we conclude our work in Section 7.

2 Lasso and Adaptive Lasso

Let us consider the linear model (1) which can be written in alternative form as

$$\mathbb{E}[Y \mid \boldsymbol{X}] = \boldsymbol{X}\boldsymbol{\beta} = \beta_1 X_1 + \cdots + \beta_p X_p \tag{2}$$

Note that $\boldsymbol{X}^T\boldsymbol{X}$ is guaranteed to be positive semi-definite but not necessarily positive definite, even for $p < n$. We make the following two assumptions on the design $\boldsymbol{X}$:

(A1) $\mathbb{E}[\boldsymbol{X}^t\epsilon \mid \boldsymbol{X}] = 0$

(A2) We assume that $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \Sigma$ exists, for a positive definite matrix Σ.

The lasso estimator (Tibshirani 1996) is defined as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) \coloneqq \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2}\|Y - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right). \tag{3}$$

Let, $\hat{\boldsymbol{\beta}} \coloneqq (\hat{\beta}_1, \cdots, \hat{\beta}_p)$ be any root-n consistent estimator of $\boldsymbol{\beta}$. Then the adaptive lasso estimates (Zou 2006) are given by

$$\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma) \coloneqq \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j(\gamma) |\beta_j| \right) \tag{4}$$

where,

$$w_j(\gamma) = \frac{1}{|\hat{\beta}_j|^\gamma}, \quad \text{for} \quad \gamma > 0. \tag{5}$$

Zou (2006) showed that the Adaptive lasso can be computed as regular lasso by using transformation of variables. We rewrite Eq. (4) as

$$\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma) = \boldsymbol{K}(\gamma) \arg\min_{\boldsymbol{\beta}^*(\gamma)} \left(\frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{K}(\gamma)\boldsymbol{\beta}^*(\gamma)\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j^*(\gamma)| \right) \tag{6}$$

where,

$$\boldsymbol{K}(\gamma) \coloneqq \text{diag}\left(\frac{1}{w_1(\gamma)}, \cdots, \frac{1}{w_p(\gamma)} \right) = \text{diag}\left(|\hat{\beta}_1|^\gamma, \cdots, |\hat{\beta}_p|^\gamma \right) \tag{7}$$

and

$$\boldsymbol{\beta}^*(\gamma) \coloneqq (w_1(\gamma)\beta_1, \cdots, w_p(\gamma)\beta_p) = [\boldsymbol{K}(\gamma)]^{-1}\boldsymbol{\beta}. \tag{8}$$

Therefore, with $\boldsymbol{X}^*(\gamma) \coloneqq \boldsymbol{X}\boldsymbol{K}(\gamma)$,

$$\hat{\boldsymbol{\beta}}^*(\lambda, \gamma) \coloneqq \arg\min_{\boldsymbol{\beta}^*(\gamma)} \left(\frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{X}^*(\gamma)\boldsymbol{\beta}^*(\gamma)\|_2^2 + \lambda \|\boldsymbol{\beta}^*(\gamma)\|_1 \right), \tag{9}$$

from which we can compute adaptive lasso estimate by $\hat{\boldsymbol{\beta}}_{\text{alasso}} = \boldsymbol{K}(\gamma)\hat{\boldsymbol{\beta}}^*(\gamma)$.

In general, we cannot find an analytical solution to Eq. (3), Eq. (4) or Eq. (9) and we need to use numerical optimisation techniques to get a solution. In order to get some intuition for the problem, it helps to consider the soft-thresholding operator,

$$\text{Soft}(\boldsymbol{\beta}; \lambda) = \text{sign}(\boldsymbol{\beta}) \cdot (|\boldsymbol{\beta}| - \lambda)_+ .$$

In the special case of an orthogonal design or only a single predictor, one can obtain the lasso estimate as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) = \text{Soft}(\hat{\boldsymbol{\beta}}_{\text{ols}}; \lambda) \tag{10}$$

where $\hat{\boldsymbol{\beta}}_{\text{ols}}$ is least square estimate (Hastie et al. 2015).

For the Adaptive lasso, the weights and the parameter γ in Eq. (5), modify this expression as follows:

$$\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma, \hat{\boldsymbol{\beta}}) = \text{Soft}(\hat{\boldsymbol{\beta}}_{\text{ols}}; \lambda/|\hat{\boldsymbol{\beta}}|^{\gamma}) = \text{sign}(\hat{\boldsymbol{\beta}}_{\text{ols}}) \cdot \left(|\hat{\boldsymbol{\beta}}_{\text{ols}}| - \frac{\lambda}{|\hat{\boldsymbol{\beta}}|^{\gamma}} \right)_{+}, \tag{11}$$

where $\hat{\boldsymbol{\beta}}$ is any root n-consistent estimate of $\boldsymbol{\beta}$ in Eq. (1). In Fig. 1, we illustrate soft-thresholding operators using Eq. (12) for different values of γ. Since ordinary least square estimates are n-consistent, therefore using $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\text{ols}}$ in Eq. (11), we get

$$\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda, \gamma) = \text{Soft}(\hat{\boldsymbol{\beta}}_{\text{ols}}; \lambda/\hat{\boldsymbol{\beta}}_{\text{ols}}^{\gamma}). \tag{12}$$

3 Consistency and Oracle Properties

Let the lasso estimator be defined by Eq. (3). We define the subset $\mathcal{S}$ such that,

$$\mathcal{S} \coloneqq \{j : \beta_j \neq 0\} \quad \text{and} \quad |\mathcal{S}| = p^* < p. \tag{13}$$

That is, the true model can be specified by p^* predictors. Then we can rearrange the input matrix $\boldsymbol{X}$ such that first p^* predictors correctly identify the model. Since, $|S| = p^* < p$, then without loss of generality we can rewrite $\Sigma = \lim_{n\to\infty} \frac{1}{n} \boldsymbol{X}^T \boldsymbol{X}$ as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{14}$$

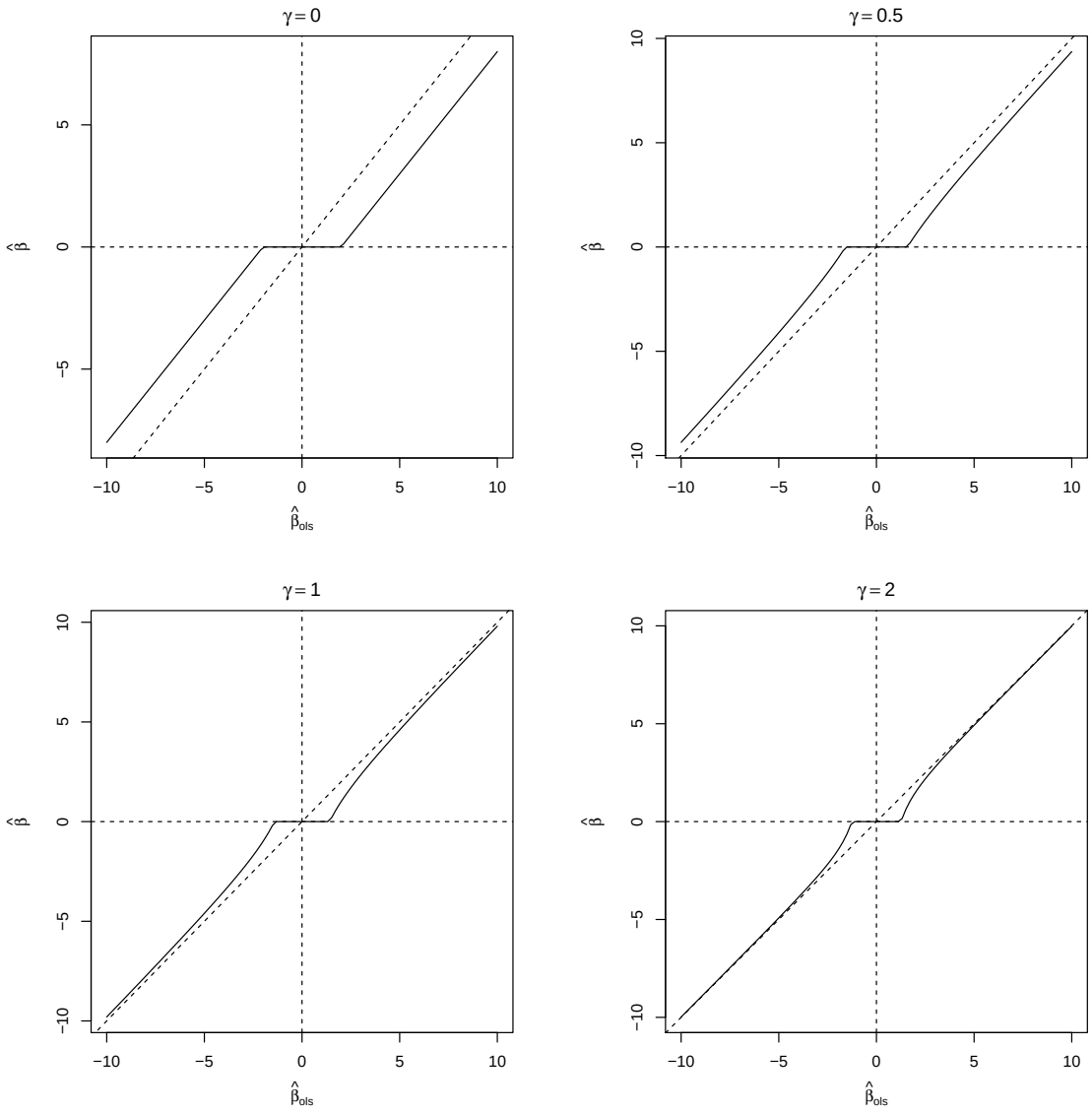

Figure 1: Soft thresholding operator for different values of γ and fixed λ $(= 2)$.

such that Σ_{11} is a $p^* \times p^*$ matrix. Now, let for n number of samples $\mathcal{S}_n \coloneqq \{j : \hat{\beta}_j \neq 0\}$. A necessary condition for the consistency of the lasso estimator is given by following theorem (Zhao & Yu 2006, Zou 2006).

Theorem 3.1. *Let,* $\lim_{n\to\infty} P(\mathcal{S}_n = \mathcal{S}) = 1$. *Then there exists some sign vector* $\boldsymbol{s} \coloneqq (s_1, s_2, \cdots, s_{p^*})$ *such that,*

$$|\Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{s}| \leq 1 \tag{15}$$

for each component of the left hand side.

Let $\mathcal{S}^{(n)}_{\text{alasso}}$ be the selected subset adaptive lasso when the sample size is n, ie.

$$\mathcal{S}^{(n)}_{\text{alasso}} \coloneqq \{j : \hat{\beta}^{(n)}_{\text{alasso; } j} \neq 0\}. \tag{16}$$

Then the Adaptive lasso estimates satisfy following asymptotic properties (Zou 2006):

Definition 3.1 (Oracle Properties)**.** *Let* $\frac{\lambda^{(n)}}{\sqrt{n}} \to 0$ *and* $\lambda^{(n)} n^{(\gamma-1)/2} \to \infty$.

1. *Consistent variable selection:* $\lim_{n\to\infty} P\left(\mathcal{S}^{(n)}_{alasso} = \mathcal{S}\right) = 1$
2. *Asymptotic normality:* $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(n)}_{alasso} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2\Sigma_{11}^{-1})$

Zou (2006) noted that adaptive lasso estimates can follow oracle properties under even weaker conditions on the convergence of λ.

4 Main Result

Consider a linear model given by Eq. (1), such that $\boldsymbol{X}$ is real valued, ie. $X_j : \Omega \to \mathbb{R}$ for $1 \leq j \leq p$, and satisfies **(A1)** and **(A2)**.

Let $\|\cdot\|$ denote the matrix norm in the space $\mathbb{R}^{p\times p}$ such that, for any matrix A

$$\|A\| \coloneqq \sup_{\|x\|_2=1} \{\|Ax\|_2 : x \in \mathbb{R}^p\}, \tag{17}$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm in $\mathbb{R}^p$. Note that, $\|A\|$ is the largest eigen value of A

Let $\{A_n\}_n$ be the sequence of matrices

$$A_n = \frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p\right), \tag{18}$$

where $0 < \lambda < \infty$.

Lemma 4.1. *The* $\lim_{n\to\infty} A_n^{-1}$ *exists and it equals to* Σ^{-1}.

Proof. To proof Lem. 4.1, we first show that, $\lim_{n\to\infty} A_n$ exists and is equal to Σ.

$$\|A_n - \Sigma\| = \left\|\frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p\right) - \Sigma\right\| \tag{19}$$

$$= \left\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} - \Sigma + \frac{\lambda}{n}\boldsymbol{I}_p\right\| \tag{20}$$

by applying triangle inequality in Eq. (20), ie. $\|a+b\| \le \|a\| + \|b\|$, we get,

$$\|A_n - \Sigma\| \le \left\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} - \Sigma\right\| + \left\|\frac{\lambda}{n}\boldsymbol{I}_p\right\| \tag{21}$$

$$= \left\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} - \Sigma\right\| + \frac{\lambda}{n}. \tag{22}$$

Now, as $n \to \infty$, $\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \Sigma$ Therefore,

$$\|A_n - \Sigma\| \to 0 \tag{23}$$

$$\implies \lim_{n\to\infty} A_n = \Sigma. \tag{24}$$

Since, $\{A_n\}_n$ is convergent, therefore it is a Cauchy sequence, that is, for every $\delta > 0$ there exists a positive natural number N such that for all natural numbers $m_1, m_2 > N$

$$\|A_{m_1} - A_{m_2}\| < \delta. \tag{25}$$

Now, since, A_n is sum of a positive semi-definite matrix ($\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$) and a diagonal matrix with positive entries ($\lambda \boldsymbol{I}_p$), it is easy to see that A_n is positive definite. Then, the inverse A_n^{-1} exists. Let, $A_n = U_n D_n U_n^T$ where, D_n is a diagonal matrix and U_n is orthogonal. Now,

$$\|A_n^{-1}\| = \left\|(U_n D_n U_n^T)^{-1}\right\| \tag{26}$$

$$= \left\|(U_n D_n^{-1} U_n^T)\right\| \tag{27}$$

since, U_n is orthogonal and D_n is diagonal, we get,

$$\|A_n^{-1}\| = \sup_{1\le j\le p}\{[D_n^{-1}]_{jj}\} \tag{28}$$

$$= \frac{1}{\inf_{1\le j\le p}\{[D_n]_{jj}\}}. \tag{29}$$

As, $A_n = \frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)$ is positive definite, therefore all of its eigen values are greater than or equal to λ. Therefore,

$$\|A_n^{-1}\| \le \frac{1}{\lambda}. \tag{30}$$

Then,

$$A_{m_1}^{-1} - A_{m_2}^{-1} = A_{m_1}^{-1} A_{m_2} A_{m_2}^{-1} - A_{m_1}^{-1} A_{m_1} A_{m_2}^{-1} \tag{31}$$

$$= A_{m_1}^{-1}\left(A_{m_2} - A_{m_1}\right) A_{m_2}^{-1} \tag{32}$$

$$\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| = \left\|A_{m_1}^{-1}\left(A_{m_2} - A_{m_1}\right) A_{m_2}^{-1}\right\| \tag{33}$$

applying Cauchy-Schwartz inequality we get,

$$\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| \le \left\|A_{m_1}^{-1}\right\| \|A_{m_2} - A_{m_1}\| \left\|A_{m_2}^{-1}\right\| \tag{34}$$

using Eq. (25),

$$\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| \leq \delta \left\|A_{m_1}^{-1}\right\| \left\|A_{m_2}^{-1}\right\| \tag{35}$$

$$\leq \frac{\delta}{\lambda^2}. \tag{36}$$

Therefore, for every $\frac{\delta}{\lambda^2} > 0$, we can find a positive natural number N, such that for every $m_1, m_2 > 0$, $\left\|A_{m_1}^{-1} - A_{m_2}^{-1}\right\| \leq \frac{\delta}{\lambda^2}$. Hence, $\{A_n^{-1}\}_n$ is a Cauchy sequence. Since, $\mathbb{R}^p$ is a Banach space under the Euclidean norm $\|\cdot\|_2$, therefore every Cauchy sequence is convergent. Then there exist L such that, $\lim_{n\to\infty} A_n^{-1} = L$. Now,

$$A_n A_n^{-1} = \boldsymbol{I}_p = A_n^{-1} A_n \tag{37}$$

$$\lim_{n\to\infty} A_n A_n^{-1} = \boldsymbol{I}_p = \lim_{n\to\infty} A_n^{-1} A_n \tag{38}$$

since both A_n and A_n^{-1} is convergent,

$$\lim_{n\to\infty} A_n \cdot \lim_{n\to\infty} A_n^{-1} = \boldsymbol{I}_p = \lim_{n\to\infty} A_n^{-1} \cdot \lim_{n\to\infty} A_n \tag{39}$$

$$\Sigma \cdot L = \boldsymbol{I}_p = L \cdot \Sigma \tag{40}$$

Therefore, $\lim_{n\to\infty} A_n^{-1} = \Sigma^{-1}$. □

We define Ridge estimates in the following way:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) \coloneqq \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2} \|Y - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right). \tag{41}$$

Lemma 4.2. *Ridge estimates are root n-consistent.*

Proof. From Eq. (41), we have the Ridge estimates as

$$\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^T Y \tag{42}$$

using $A_n = \frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)$

$$= (nA_n)^{-1}\,\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \epsilon) \tag{43}$$

$$= (nA_n)^{-1}\,\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + (nA_n)^{-1}\,\boldsymbol{X}^T\epsilon \tag{44}$$

We know that, $\mathbb{E}[\boldsymbol{X}^T\epsilon \mid \boldsymbol{X}] = 0$. Therefore, conditioning on $\boldsymbol{X}$, we get

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta} \mid \boldsymbol{X}] = (nA_n)^{-1}\,\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta} \tag{45}$$

$$= (nA_n)^{-1}\,(nA_n - \lambda \boldsymbol{I}_p)\,\boldsymbol{\beta} - \boldsymbol{\beta} \tag{46}$$

$$= \boldsymbol{\beta} - (nA_n)^{-1}\,\lambda\boldsymbol{\beta} - \boldsymbol{\beta} \tag{47}$$

$$= -\lambda\,(nA_n)^{-1}\,\boldsymbol{\beta}. \tag{48}$$

Multiplying $\sqrt{n}$ on both sides,

$$\mathbb{E}[\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}) \mid \boldsymbol{X}] = -\sqrt{n}\lambda\,(nA_n)^{-1}\,\boldsymbol{\beta} \tag{49}$$

$$= -\frac{\sqrt{n}}{n}\lambda A_n^{-1}\boldsymbol{\beta} \tag{50}$$

$$= -\frac{\lambda}{\sqrt{n}}A_n^{-1}\boldsymbol{\beta} \tag{51}$$

Now, as $n \to \infty$, from Eq. (51), we get:

$$\lim_{n\to\infty} \mathbb{E}[\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}) \mid \boldsymbol{X}] = \lim_{n\to\infty} -\frac{\lambda}{\sqrt{n}}A_n^{-1}\boldsymbol{\beta} \tag{52}$$

since, by Lem. 4.1, $\lim_{n\to\infty} A_n^{-1}$ exists and $\boldsymbol{\beta}$ is independent of n, therefore using product rule of limits we get

$$= -\boldsymbol{\beta} \lim_{n\to\infty} \frac{\lambda}{\sqrt{n}} \lim_{n\to\infty} A_n^{-1} \tag{53}$$

$$= -\boldsymbol{\beta} \cdot 0 \cdot \Sigma^{-1} \tag{54}$$

$$= 0. \tag{55}$$

This proves $\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\right)$ is asymptotically unbiased. Along the same lines it would be possible to show that $n^s\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\right)$ is asymptotically unbiased for any $0 \leq s < 1$.

As before, conditioning on $\boldsymbol{X}$, we get:

$$\begin{aligned}
\text{Var}\left[\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\right) \mid \boldsymbol{X}\right] &= \text{Var}\left[\sqrt{n}\left((nA_n)^{-1}\boldsymbol{X}^T\epsilon\right)\right] && (56)\\
&= \text{Var}\left[\frac{\sqrt{n}}{n}A_n^{-1}\boldsymbol{X}^T\epsilon\right] && (57)\\
&= \text{Var}\left[A_n^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{X}^T\epsilon\right)\right] && (58)\\
&= A_n^{-1}\text{Var}\left[\left(\frac{1}{\sqrt{n}}\boldsymbol{X}^T\epsilon\right)\right]\cdot\left(A_n^{-1}\right)^T && (59)
\end{aligned}$$

since, $A_n^{-1} = \left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{n}\lambda\boldsymbol{I}_p\right)^{-1}$ is symmetric

$$= A_n^{-1}\cdot\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\cdot\text{Var}[\epsilon]\cdot A_n^{-1}. \quad (60)$$

Now, as $n \to \infty$,

$$\lim_{n\to\infty}\text{Var}\left[\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\right) \mid \boldsymbol{X}\right] = \lim_{n\to\infty} A_n^{-1}\cdot\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\cdot\text{Var}[\epsilon]\cdot A_n^{-1}. \quad (61)$$

Since, by Lem. 4.1, $\lim_{n\to\infty} A_n^{-1}$ exists and $\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ exists by assumption, therefore applying product rule of limits, we get:

$$\begin{aligned}
&= \text{Var}[\epsilon]\cdot\lim_{n\to\infty} A_n^{-1}\cdot\lim_{n\to\infty}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\cdot\lim_{n\to\infty} A_n^{-1} && (62)\\
&= \sigma^2\Sigma^{-1}\Sigma\Sigma^{-1} && (63)\\
&= \sigma^2\Sigma^{-1}. && (64)
\end{aligned}$$

Now, by central limit theorem we know,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\right) \xrightarrow{d} \mathcal{N}\left(\mathbb{E}\left[\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\right)\right], \text{Var}\left[\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda) - \boldsymbol{\beta}\right)\right]\right). \quad (65)$$

Now, applying Eq. (44), Eq. (51) and Eq. (64) we get,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)-\boldsymbol{\beta}\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\Sigma^{-1}\right). \tag{66}$$

This proves that Ridge estimates are root n-consistent. □

Let,

$$Y=\boldsymbol{X}\boldsymbol{\beta}^*+\boldsymbol{\epsilon} \tag{67}$$

be the model with true regression coefficients $\boldsymbol{\beta}^*$, such that $|\beta^*|\gg 1$. Then clearly, $\boldsymbol{X}$ has full column rank, ie. $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.

Let, $\boldsymbol{K}(\gamma)$ is defined by Eq. (7) and for the sake of notation, we write it as $\boldsymbol{K}$.

Theorem 4.3. *For large effects models (ie. $|\beta_j|\gg 1$ and $0<\lambda<\min\{\boldsymbol{K}\boldsymbol{X}^TY\}$, we have,*

$$\left\|\hat{\boldsymbol{\beta}}_{alasso}(\lambda,\gamma)-\boldsymbol{\beta}^*\right\|_2^2\leq\frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\|+\frac{\lambda^2p}{n^2}\left\|\Sigma_n^{-1}\right\|^2\min_{1\leq j\leq p}|\hat{\beta}_j|^{-2\gamma} \tag{68}$$

$$\left\|Y-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{alasso}(\lambda,\gamma)\right\|_2^2\leq\frac{\lambda^2p}{n}\left\|\Sigma_n^{-1}\right\|\min_{1\leq j\leq p}|\hat{\beta}_j|^{-2\gamma} \tag{69}$$

Therefore as $\gamma\to\infty$ the adaptive estimates are unbiased.

Proof. Let the adaptive lasso model be defined by Eq. (4). We use Ridge estimates as the weights of adaptive lasso. Let $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ be the ridge estimates such that

$$\hat{\boldsymbol{\beta}}_{\text{ridge}}=(\hat{\beta}_1,\cdots,\hat{\beta}_p). \tag{70}$$

Then the weights are given by:

$$w=\left(\frac{1}{|\hat{\beta}_1|^\gamma},\cdots\frac{1}{|\hat{\beta}_p|^\gamma}\right). \tag{71}$$

Then, applying w as weights in adaptive lasso estimates we get,

$$\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma) = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right). \tag{72}$$

Now, applying Karush-Kahn-Tucker condition in Eq. (72), we have

$$\mathbf{0} \in -\boldsymbol{X}^T(Y - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{K}^{-1} \partial \|\boldsymbol{\beta}\|_1. \tag{73}$$

We write $\partial\|\boldsymbol{\beta}\|_1$ in Eq. (73) in the following way: (Nesterov 2014, §3.1.5)

$$\partial\|\boldsymbol{\beta}\|_1 = \text{sign}(\beta_1) \times \cdots \times \text{sign}(\beta_p) \tag{74}$$

where

$$\text{sign}(\beta_j) \coloneqq \begin{cases} \{-1\} & \text{if } \beta_j < 0 \\ [-1,1] & \text{if } \beta_j = 0 \\ \{1\} & \text{if } \beta_j > 0. \end{cases} \tag{75}$$

Note that, for any fixed $\lambda < min\{\boldsymbol{K}\boldsymbol{X}^T Y\}$, $\beta_j \neq 0$ for $1 \leq j \leq p$. Then, from Eq. (75) we have:

$$\text{sign}(\beta_j) \coloneqq \begin{cases} \{-1\} & \text{if } \beta_j < 0 \\ \{1\} & \text{if } \beta_j > 0. \end{cases} \tag{76}$$

Therefore, we write Adaptive lasso estimates as:

$$\boldsymbol{X}^T \left(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma) \right) = \lambda \boldsymbol{K}^{-1}\boldsymbol{s} \tag{77}$$

$$\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma) \right) = \lambda \boldsymbol{K}^{-1}\boldsymbol{s} \tag{78}$$

where $\boldsymbol{s} = (s_1, s_2, \ldots, s_{p^*})$ are auxiliary variables subject to the constraint $s_j \in \text{sign}(\beta_j)$.

Now, from Eq. (78), we get

$$\frac{1}{n}\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma) \right) = \frac{\lambda}{n}\boldsymbol{K}^{-1}\boldsymbol{s} \tag{79}$$

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \left(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma) \right) = \frac{\lambda}{n}\boldsymbol{K}^{-1}\boldsymbol{s} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{\epsilon} \tag{80}$$

Since, inverse of $\boldsymbol{X}^T\boldsymbol{X}$ exists. Then,

$$\left(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma)\right) = \left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\frac{\lambda}{n}\boldsymbol{K}^{-1}\boldsymbol{s} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{\epsilon}\right) \tag{81}$$

taking norm in both sides,

$$\left\|\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma) - \boldsymbol{\beta}^*\right\|_2^2 = \left\|\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\epsilon} - \frac{\lambda}{n}\boldsymbol{K}^{-1}\boldsymbol{s}\right)\right\|_2^2 \tag{82}$$

$$\leq \left\|\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\epsilon}\right\|_2^2 + \left\|\frac{\lambda}{n}\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{K}^{-1}\boldsymbol{s}\right\|_2^2 \tag{83}$$

$$\leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\| + \frac{\lambda^2}{n^2}\left\|\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{K}^{-1}\right\|^2\|\boldsymbol{s}\|_2^2. \tag{84}$$

Here, $\|.\|$ is the induced matrix norm in $\mathbb{R}^p$. Now, since, $\|\boldsymbol{s}\|_2^2 = p$

$$\leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\| + \frac{p\cdot\lambda^2}{n^2}\left\|\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right\|^2\left\|\boldsymbol{K}^{-1}\right\|^2 \tag{85}$$

$$\leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\| + \frac{\lambda^2 p}{n^2}\left\|\Sigma_n^{-1}\right\|^2\left\|\boldsymbol{K}^{-1}\right\|^2 \tag{86}$$

$$\leq \frac{\sigma^2}{n}\left\|\Sigma_n^{-1}\right\| + \frac{\lambda^2 p}{n^2}\left\|\Sigma_n^{-1}\right\|^2\min_{1\leq i\leq p}|\hat{\beta}_i|^{-2\gamma}. \tag{87}$$

Similarly, from Eq. (77), we have,

$$\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{X}^T\left(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma)\right) = \lambda\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{K}^{-1}\boldsymbol{s} \tag{88}$$

$$\left(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma)\right) = \lambda\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{K}^{-1}\boldsymbol{s} \tag{89}$$

Taking norm on both sides of Eq. (89), we get

$$\left\|Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma)\right\|_2^2 = \lambda^2\left\|\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{K}^{-1}\boldsymbol{s}\right\|_2^2 \tag{90}$$

applying Cauchy-Schwartz inequality

$$\leq \frac{\lambda^2}{n} \left\|\Sigma_n^{-1}\right\| \left\|\boldsymbol{K}^{-1}\right\|^2 \|\boldsymbol{s}\|_2^2. \tag{91}$$

Therefore, we get the following:

$$\left\|Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\text{alasso}}(\lambda,\gamma)\right\|_2^2 \leq \frac{\lambda^2 p}{n} \left\|\Sigma_n^{-1}\right\| \min_{1\leq i\leq p} |\hat{\beta}_i|^{-2\gamma}. \tag{92}$$

So, we reduce the mean square error by increasing the value of γ. It also indicates that as for higher values of γ, λ does not controls any shrinkage over large effects and produce unbiased estimates. □

5 Simulation Study

Model: We simulate the predictors from a standard normal distribution such that, $X_{i,j} \sim N(0,1)$ for $j = 1, \cdots, 6$ and $i = 1, \cdots, n$. We assign the regression coefficients to be, $(\beta_1, \cdots, \beta_6) \coloneqq (4, 3, 2, -2, -3, -4)$. Further, we add redundant variables such that, $\beta_j = 0$ for $j > 6$. We consider standard normal noise to construct the response vector $y_i = \sum_{j=1}^{6} X_{i,j}\beta_j + 0.01\epsilon_i$ where, $\epsilon_i \sim N(0,1)$ for $i = 1, \cdots, n$. We repeat this experiment for $n = 10, 100, 1000$ and $p = 7, 20, 50$.

Results: We analyse the sensitivity of the model for $0 \leq \gamma \leq 10$ ($\gamma = 0$ allows us to obtain regular lasso estimates). We use Ridge regression estimates as the weights of Adaptive lasso. We show the coefficient paths in Fig. 2, for $p = 7$. On the left hand side in Fig. 2, we show coefficient path in terms of γ for a fixed $\lambda(= 0.5)$. For $n = 10$, we see an interesting feature of the model. The use of fixed λ forcefully shrinks some true large effects to zero for the smaller values of γ. However, as γ increases, the effect of λ becomes less significant and the method recovers actual values of $\boldsymbol{\beta}$. The use of fixed λ is

also more stable than the adaptive lasso with cross-validated λ for $n = 10$. We also notice an interesting behaviour of adaptive lasso for cross validated λ (right hand side in Fig. 2), as the values of γ increases, some estimates become smaller. This possibly happens due to the local shrinkage effects of cross-validated λ.

As we increase the amount information ($n = 100, 1000$), our method becomes more consistent for smaller values of γ unlike the previous case ($n = 10$). In both cases, our estimates converge to the true values for larger values of γ. However, for cross-validated λ (right side of Fig. 2), adaptive lasso tends to shrink the smaller coefficients towards zero and increase the values of larger effects similar to the previous case.

We use root mean square error to measure prediction accuracy. We compute the root mean square error as:

$$\text{rmse} = \sqrt{\frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2}. \tag{93}$$

We illustrate our results in Fig. 3. The left side in Fig. 3 shows the root mean square error for fixed λ ($= 0.5$). The fixed λ causes poor accuracy for small γ. However, it performs better for larger values and converges to a fixed value. For cross-validated case (right side of Fig. 3), we see that, as γ increases, some true effects go to zero increasing the value of root mean square. We also notice, the addition of information eventually increases the accuracy.

For 20 and 50 predictors, we use λ from initial ridge estimates as both are high dimensional problem for $n = 10$ and we may get poor result due to the scaling factor. We compare prediction accuracy of different variants of lasso in Table 1 using RMSE. We observe that for $p = 20, 50$ the RMSE is higher for $\gamma = 10$ than $\gamma = 5$. This happens as the ridge estimates are not sparse and it contributes to negligible (10^{-5}) false positive effects. We can improve this by introducing a truncation constant or increasing the computation power.

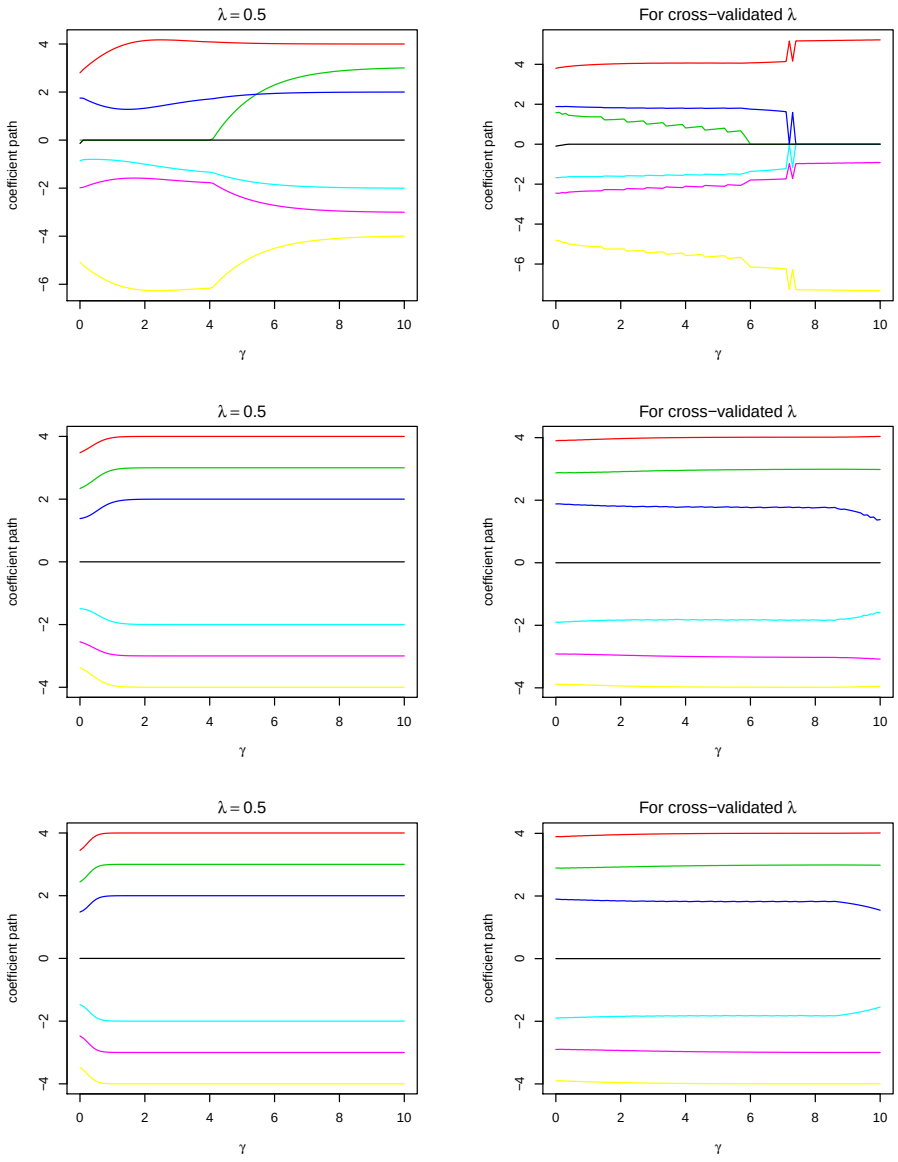

Figure 2: Coefficient paths of regression coefficients for fixed λ ($= 0.5$) and cross-validated λ for $p = 7$ and $n = 10$ (top), $n = 100$ (middle) and $n = 1000$ (bottom).

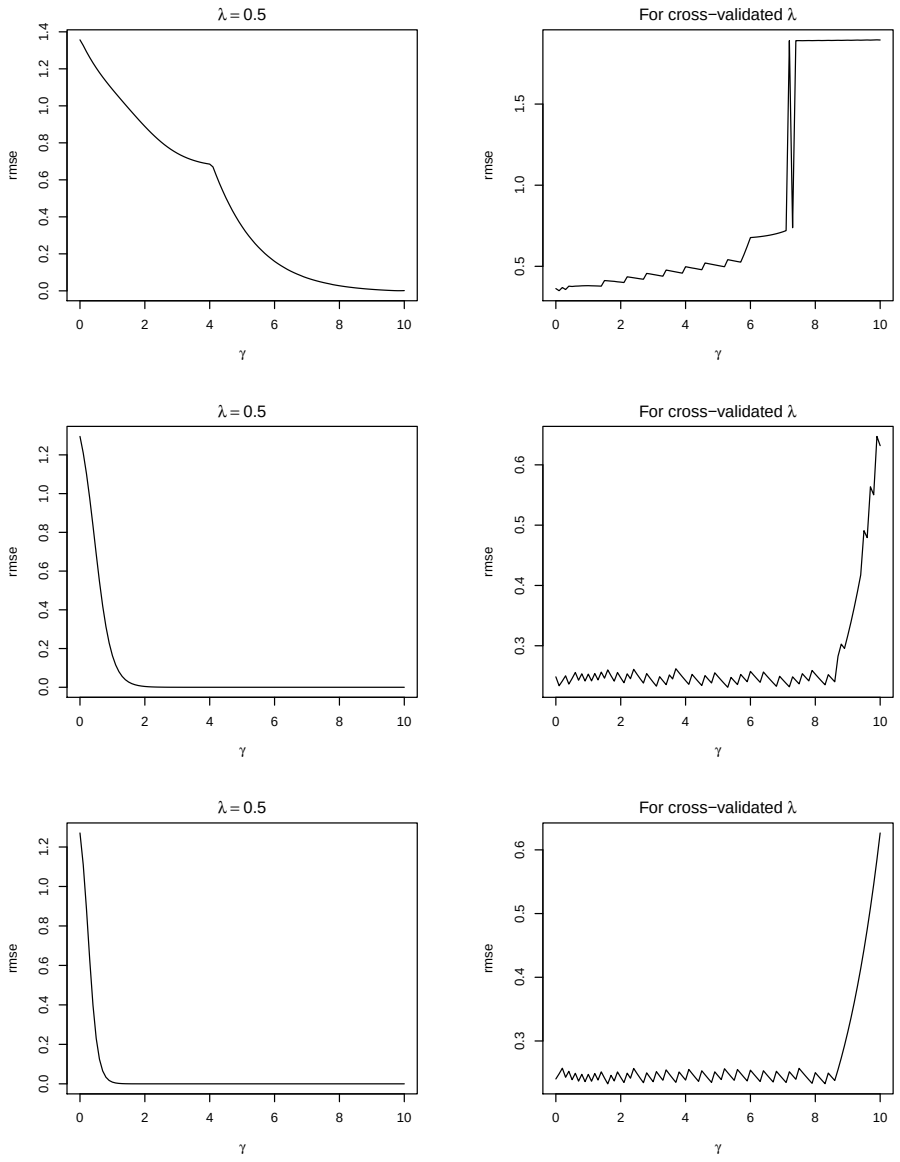

Figure 3: Comparison of root mean square error for fixed λ ($= 0.5$) and cross-validated λ for $p = 7$ and $n = 10$ (top), $n = 100$ (middle) and $n = 1000$ (bottom).

Methods	$n = 10$	$n = 100$	$n = 1000$
p = 7			
Fixed λ (= 0.5), $\gamma = 1$	1.088955	0.163489	0.008632
Fixed λ (= 0.5), $\gamma = 5$	0.346950	0.000001	0.000001
Fixed λ (= 0.5), $\gamma = 10$	0.001299	0.000001	0.000001
Lasso	0.362836	0.248448	0.240389
Adaptive lasso	0.349714	0.231137	0.232711
p = 20			
Fixed λ (= 73.83546), $\gamma = 1$	8.011325	0.006552	0.004495
Fixed λ (= 0.434882), $\gamma = 5$	0.003401	0.000763	0.000001
Fixed λ (= 0.428836), $\gamma = 10$	0.004101	0.000777	0.000001
Lasso	4.176416	0.248448	0.240389
Adaptive lasso	0.336443	0.232360	0.232481
p = 50			
Fixed λ (= 70.55713), $\gamma = 1$	7.387310	0.014160	0.002132
Fixed λ (= 0.490154), $\gamma = 5$	0.012112	0.005652	0.000012
Fixed λ (= 0.447235), $\gamma = 10$	0.000747	0.006424	0.000026
Lasso	0.731940	0.276000	0.246367
Adaptive lasso	0.264038	0.249035	0.237450

Table 1: Comparison of prediction accuracy (RMSE) between different methods.

6 Data Analysis

Gaia Dataset: We use the Gaia dataset for this example. This is a simulated dataset which as produced by the European Space Agency in preparation of the GAIA project,

which produces a 3-dimensional mapping of our home galaxy (Einbeck et al. 2008). In the Gaia dataset we have 16 predictors, which are 16 different photon bands, and stellar temperature as our response variable. The gaia dataset is a highly correlated dataset with the multicolinearity property. We show the scatterplot matrix of the predictors in Fig. 4.

Results: For the Gaia dataset, we take $0 \leq \gamma \leq 20$ to investigate the behaviour of the Adapative lasso under prior weights informed by ridge regression. We run our analysis for three different cases $n = 10, 100, 1000$. We observe that, because of the dataset's correlated nature (Fig. 4), regular Adaptive lasso tends to shrink all but one coefficient to zero (Fig. 5). This results in a large mean squared error (right side of Fig. 6). We use λ obtained from ridge regression to fix the value of λ to deal with scaling of the dataset. In our method, we see that for smaller values of γ it gives a sparse fit. However, for higher values of γ, the estimates become consistent. We observe that for $n = 10$, regular adaptive lasso is very sensitive with respect to γ. We illustrate a comparison between different methods in Table 2. We see the sensitivity of prediction accuracy in terms of γ. We also notice that, for $n = 1000$, the RMSE is greater for $\gamma = 20$ than that for $\gamma = 10$. This happens to be counter intuitive. We observed this behaviour in different runs and we conclude this to be an effect of high correlation between the covariates and ridge estimates (non-sparse) as weights.

7 Conclusion

We have presented a sensitivity analysis for adaptive lasso with respect to γ. We obtained to novel bounds on the lasso for large effects models. We have shown that the bias due to λ can be removed for larger values of γ.

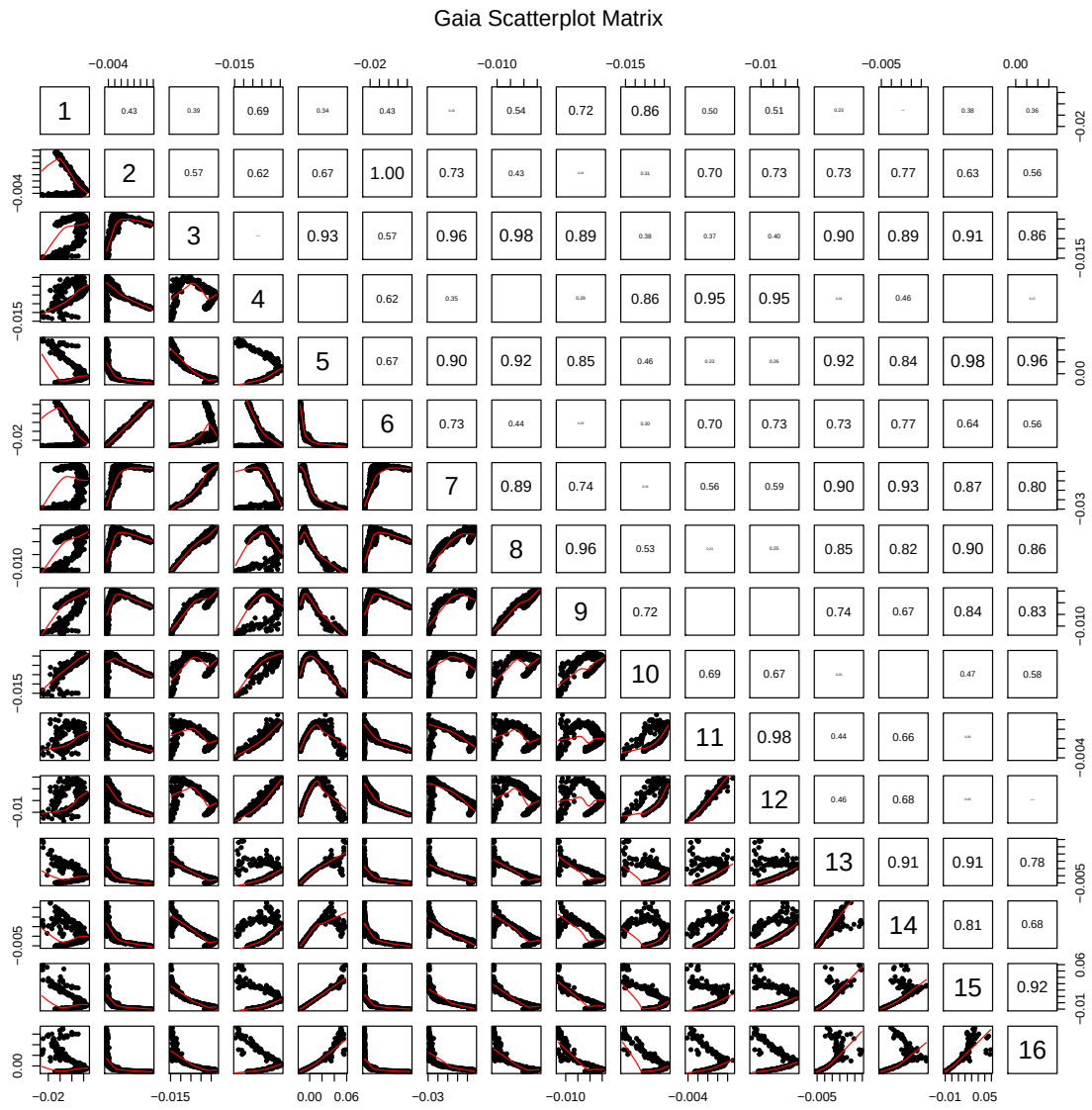

Figure 4: Scatterplot matrix of Gaia dataset.

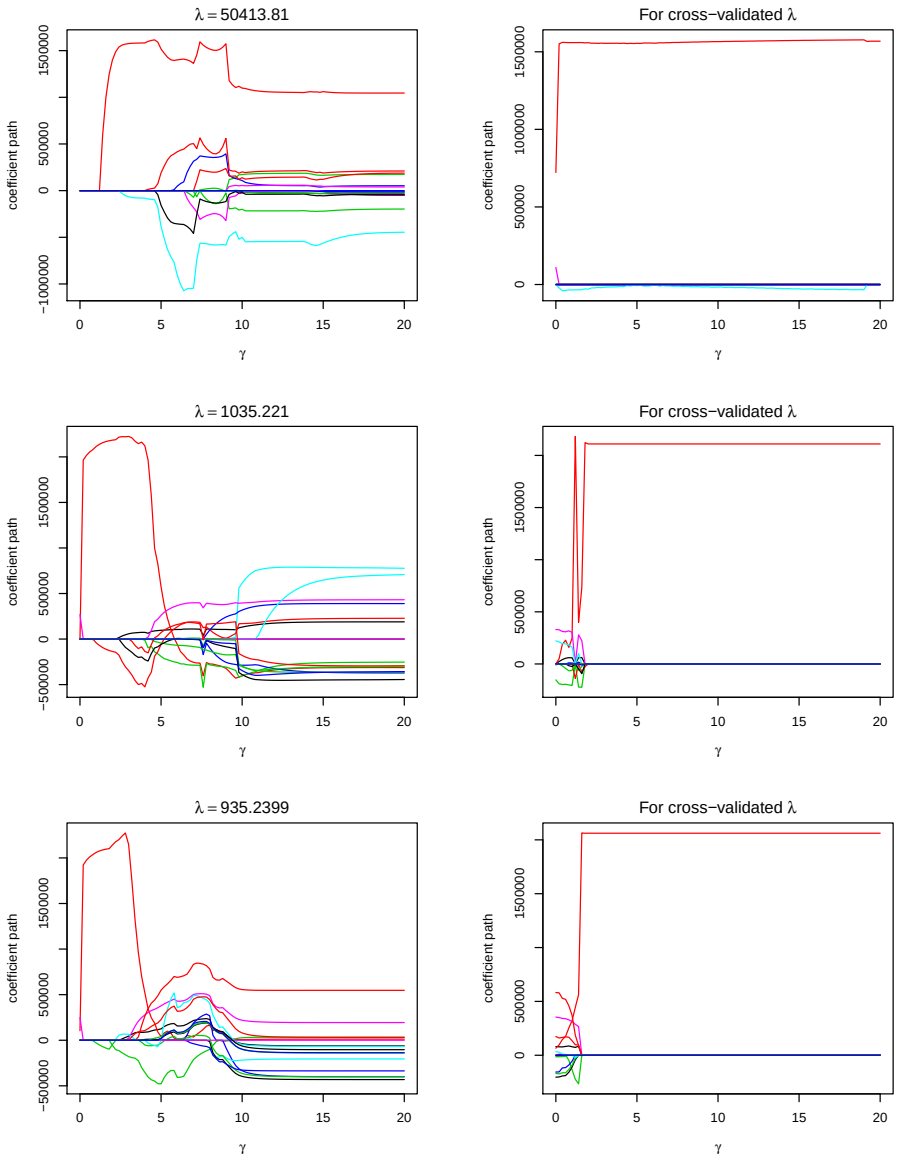

Figure 5: Coefficient paths of regression coefficients of Gaia predictors for fixed λ and cross-validated λ for $n = 10$ (top), $n = 100$ (middle) and $n = 1000$ (bottom).

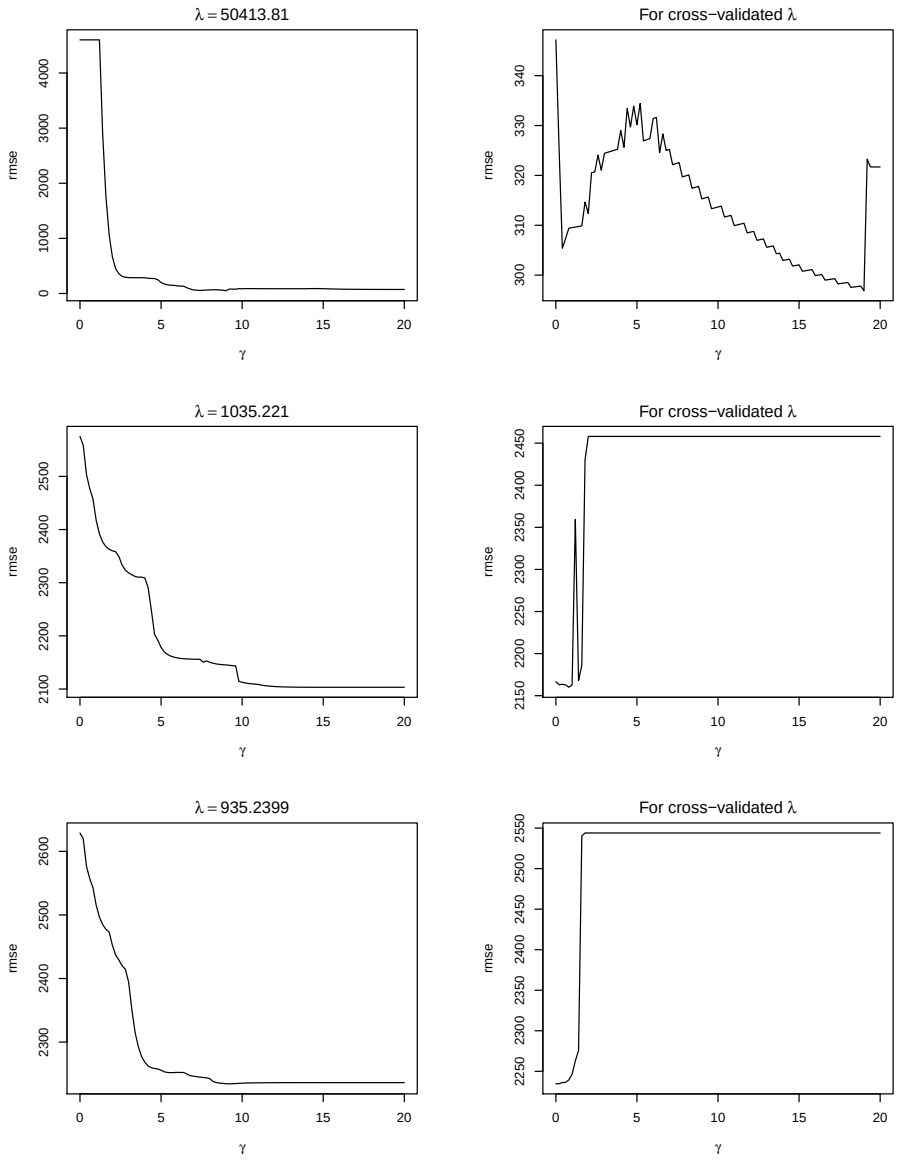

Figure 6: Mean square errors obtained from Gaia dataset for fixed λ and cross-validated λ for $n = 10$ (top), $n = 100$ (middle) and $n = 1000$ (bottom).

Methods	$n = 10$	$n = 100$	$n = 1000$
Fixed λ (= 50413.81),$\gamma = 1$	4601.578534	2417.014006	2515.063785
Fixed λ (= 1035.221),$\gamma = 10$	84.677687	2112.710441	2235.308301
Fixed λ (= 935.2399),$\gamma = 20$	73.016061	2103.379011	2236.224609
Lasso	347.180243	2166.554007	2234.551771
Adaptive lasso	296.878416	2160.011880	2234.731977

Table 2: Comparison of different methods for gaia dataset for $n = 10, 100, 1000$.

We have validated these results for simulated dataset with three different choice of predictors and sample size. We also have presented an illustration for a correlated dataset with three different sample size. For both cases, we found an agreement between the theoretical results and simulation. We noticed prediction accuracy can be increased by careful choice of γ.

Further our goal will be to generalise this result for mixed effect models with a weaker oracle condition and to obtain an estimate for the number of false positives and false negatives.

References

Einbeck, J., Evers, L. & Bailer-Jones, C. (2008), Representing complex data using localized principal components with application to astronomical data, *in* A. N. Gorban, B. Kégl, D. C. Wunsch & A. Y. Zinovyev, eds, 'Principal Manifolds for Data Visualization and Dimension Reduction', Springer, Berlin, Heidelberg, pp. 178–201.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its

oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.
URL: *http://www.jstor.org/stable/3085904*

Geer, S. A. V. D. & Bühlmann, P. (2009), 'On the conditions used to prove oracle results for the lasso', *Electron. J. Stat* .

Hastie, T., Tibshirani, R. & Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press.
URL: *https://books.google.co.uk/books?id=f-A_CQAAQBAJ*

Nesterov, Y. (2014), *Introductory Lectures on Convex Optimization: A Basic Course*, 1st edn, Springer Publishing Company, Incorporated.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288.
URL: *http://www.jstor.org/stable/2346178*

Tikhonov, A. N. (1963), 'On the solution of ill-posed problems and the method of regularization', *Dokl. Akad. Nauk SSSR* **151**(3), 501–504.
URL: *http://www.ams.org/mathscinet-getitem?mr=0162377*

Van de Geer, S., Bühlmann, P. & Zhou, S. (2011), 'The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso)', *Electron. J. Statist.* **5**, 688–749.
URL: *https://doi.org/10.1214/11-EJS624*

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2563.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.
URL: *https://doi.org/10.1198/016214506000000735*