TWO FORMS OF AIC BASED ON MODIFIED LASSO

Yujie Xue
Department of Pure and Applied Mathematics, Waseda University
3 Chome-4-1, Okubo, Shinjuku-ku, Tokyo, Japan
ORCID: 0000-0002-4607-7498
yujiexue23@asagi.waseda.jp

Received August 30,2020 ; revised September 11,2020

Abstract. The least absolute shrinkage and selection operator (LASSO) is a popular technique for variable selection and estimation in linear regression models. Introduction of information criteria for LASSO can decrease the computational cost efficiently. So far the forms of some classic information criteria for LASSO are derived. In fact, there exists some regression matrix such that the ordinary LASSO may not select the correct model efficiently even by information criteria. In such situation, [9] introduced modified LASSO approach. In this paper, we introduce two forms of Akaike information criterion (AIC) based on modified LASSO estimation to help find the optimal tuning parameters for prediction and variable selection purposes respectively. The properties of those two forms are shown and a simulation study comparing these two forms is conducted.

1 Introduction The least absolute shrinkage and selection operator (LASSO) is proposed by [7], and is a popular technique for variable selection and estimation in linear regression models. As we know, the performance of the LASSO relies heavily on the choice of tuning parameter λ to select the optimal model. For prediction purpose, the prediction error is estimated by using cross-validation (CV) or by information criteria ([2]). A drawback of using information criteria is that the degrees of freedom must be known. [8] showed that the number of nonzero coefficients is an unbiased estimate for the degree of freedom of the LASSO, and the unbiased estimator is shown to be asymptotically consistent. For variable selection purpose, choosing the optimal tuning parameter is more difficult since the prediction optimal value is inconsistent in the sense of correct selection. [4] shows that for certain high dimensional cases, generalized information criterion (GIC) on sub-models decided by LASSO is consistent in the sense of correct selection. In the paper, we consider a more general linear model, where the ordinary LASSO estimation may not work well. We consider the following linear model:

$$Y_i = \boldsymbol{x}_i' \boldsymbol{\beta}^* + \varepsilon_i,$$

where $1 \leq i \leq n$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$, and $\{\varepsilon_i\}$ is independent and identically distributed process with $\varepsilon_i \sim N(0, \sigma^2)$. Let $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ be a known nonrandom function of i. By $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, we discuss the estimation of $\boldsymbol{\beta}^*$ based on an observed stretch $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$. Let $a_{jk}^n = \sum_{t=1}^n x_{tj} x_{tk}$, and we assume the following conditions on $\{\boldsymbol{x}_i\}$.

Assumption 1 *1.* $a_{jj}^n \to \infty$ $(n \to \infty)$, $(j = 1, \ldots, p)$.

2. $\lim_{n\to\infty} \frac{x_{n+1,j}^2}{a_{jj}^n} = 0$, $(j = 1, \ldots, p)$.

3. *The limit*
$$\lim_{n\to\infty}\frac{a^n_{jk}}{\sqrt{a^n_{jj}a^n_{kk}}}=\rho_{jk}$$
exists for $j,k=1,\ldots,p,\ h\in\mathbb{Z}$.

4. *Letting* $\Phi\equiv\{\rho_{jk}:j,k=1,\ldots,p\}$, Φ *is regular.*

The point of item 2 of Assumption 1 is to prevent that the last $x^2_{n+1,j}$ from being an appreciable part of the sum of squares for large n. Item 3 shows that the relations between regressors for all sufficiently large n are approximately fixed values. Item 4 is for avoidance of multicollinearity of the model. Obviously, the model includes the case that the norm of different column in regression matrix may have different order of sequence length n. For example, letting $x_{ij}=i^{j-1}$, $O(\sum_{i=1}^n x^2_{ij})$ is greater than $O(n)$ when $j\geq 2$. In such condition, the ordinary LASSO estimation, where the estimators for the coefficients $\boldsymbol{\beta}^*$ are obtained by
$$\tilde{\boldsymbol{\beta}}(\lambda_n)=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^n(Y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2+\sum_{j=1}^p\lambda_n|\beta_j|,$$
might not work well in the sense of variable selection, where λ_n is a given tuning parameter. Correspondingly, it requires the modified LASSO estimation to match the different order of each column which was introduced by [9] as follows:
$$\hat{\boldsymbol{\beta}}(\lambda_n)=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^n(Y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2+\sum_{j=1}^p\lambda_n\sqrt{a^n_{jj}}|\beta_j|.$$
In the numerical results ([9]), it was shown that the estimation of modified LASSO has higher probability of correct selection of true model than that of the ordinary LASSO even by selecting an optimal λ_n with Akaike information criterion (AIC). In this paper, we construct two forms of AIC based on modified LASSO for prediction and variable selection purposes respectively in Section 2. In Section 3, the numerical analysis part, the selection and prediction performance of the modified LASSO when using the above two forms of AIC are analysised.

2 Main results We first define some notations. Let $\hat{\boldsymbol{\mu}}_{\lambda_n}$ be the modified LASSO fit. $\hat{\mu}_i$ is the ith component of $\hat{\boldsymbol{\mu}}$. For convenience, we let $df(\lambda_n)$ stands for $df(\hat{\boldsymbol{\mu}}_{\lambda_n})$, the degrees of freedom of the modified LASSO. Suppose $\boldsymbol{W}$ is a matrix with p column. Let $\mathcal{S}$ be a subset of the indices set $\{1,2,\ldots,p\}$. Denote $\boldsymbol{W}_{\mathcal{S}}=[\cdots W_j\cdots]_{j\in\mathcal{S}}$, where W_j is the jth column of $\boldsymbol{W}$. Similarly, define $\boldsymbol{\beta}_{\mathcal{S}}=(\cdots\beta_j\cdots)_{j\in\mathcal{S}}$ for any vector $\boldsymbol{\beta}$ of length p. Let $\mathrm{sgn}(\cdot)$ be the sign function: $\mathrm{sgn}(x)=1$ if $x>0$; $\mathrm{sgn}(x)=0$ if $x=0$; $\mathrm{sgn}(x)=-1$, if $x<0$. Let $\mathcal{S}_0=\{j:\mathrm{sgn}(\boldsymbol{\beta}^*)_j\neq 0\}$ be the active set of $\boldsymbol{\beta}^*$, where $\mathrm{sgn}(\boldsymbol{\beta})$ is the sign vector of $\boldsymbol{\beta}$ given by $\mathrm{sgn}(\boldsymbol{\beta})_j=\mathrm{sgn}(\beta_j)$. We denote the active set of $\hat{\boldsymbol{\beta}}(\lambda_n)$ as $\mathcal{S}_0(\lambda_n)$ and the corresponding sign vector $\mathrm{sgn}(\hat{\boldsymbol{\beta}}(\lambda_n))$ as $\mathrm{sgn}(\lambda_n)$.

2.1 Prediction purpose Prediction accuracy of a model can be assessed by calculating its prediction error, that is, the error when the model is used to predict a new sample of observations. Let $\hat{\boldsymbol{\mu}}$ be a model fit decided by $\boldsymbol{Y}$. The estimation of prediction error, covariance penalties (C_p) which was first introduced by [5], can be treated as a criterion to show how well $\hat{\mu}$ will predict a future dataset independently generated by the same linear regression model. Mallows shows that if $\hat{\boldsymbol{\mu}}=\boldsymbol{M}\boldsymbol{Y}$, where $\boldsymbol{M}$ is an $n\times n$ matrix not depending on $\boldsymbol{Y}$, $C_p(\hat{\boldsymbol{\mu}}) := \frac{\|\boldsymbol{Y}-\hat{\boldsymbol{\mu}}\|^2}{n} + \frac{2\,\mathrm{trace}(\boldsymbol{M})}{n}\sigma^2$ is an unbiased estimation for the expectation of prediction error. For a general estimate $\hat{\boldsymbol{\mu}} = m(\boldsymbol{Y})$, as in [10], C_p can be extended to

$$C_p(\hat{\boldsymbol{\mu}}) := \frac{\|\boldsymbol{Y}-\hat{\boldsymbol{\mu}}\|^2}{n} + \frac{2df(\hat{\boldsymbol{\mu}})}{n}\sigma^2,$$

where $df(\hat{\boldsymbol{\mu}}) := \frac{\sum_{i=1}^n \mathrm{cov}(\hat{\mu}_i, Y_i)}{\sigma^2}$. By the connection between Mallows's C_p ([5]) and AIC ([1]), we know

$$AIC(\hat{\boldsymbol{\mu}}) = \frac{C_p(\hat{\boldsymbol{\mu}})}{\sigma^2}.$$

In the following, we introduce the form of AIC for modified LASSO by following the line of [8].
From the properties of modified LASSO solution, for a given $\boldsymbol{Y}$, there is a finite sequence,

$$\lambda_{n0} > \lambda_{n1} > \lambda_{n2} > \cdots > \lambda_{nK} = 0,$$

such that for all $\lambda_n > \lambda_{n0}$, $\hat{\boldsymbol{\beta}}(\lambda_n) = \boldsymbol{0}$, and that for all $\lambda_n \in (\lambda_{n,m+1}, \lambda_{nm})$, the active set $\mathcal{S}_0(\lambda_n)$ and sign vector $\mathrm{sgn}(\lambda_n)$ are invariant with respect to λ_n. Thus we write them as $\mathcal{S}_m$ and sgn_m for simplicity. Noticing that for any $m = 0, \ldots, K-1$, when λ_n decreases from the right hand side of λ_{nm}, some predictors with zero coefficient at λ_{nm} are about to have nonzero coefficients, we call λ_{nm} as a transition point. Correspondingly, for any $\lambda_n \in [0,\infty) - \{\lambda_{nm}, m = 0, \ldots, K-1\}$, it is called as a non-transition point.

Theorem 1 *For any $\lambda_0 \geq 0$, the modified LASSO fit $\hat{\boldsymbol{\mu}}_{\lambda_n}(\boldsymbol{Y})$ is uniformly Lipschitz. Furthermore, under the condition that $\boldsymbol{X}$ is full rank, the degree of freedom of $\hat{\boldsymbol{\mu}}_{\lambda_n}(\boldsymbol{Y})$ equals the expectation of the cardinality of the active set $\mathcal{S}_0(\lambda_n)$, that is,*

$$df(\lambda_n) = \mathrm{E}|\,\mathcal{S}_0(\lambda_n)|.$$

Theorem 1 shows that $\widehat{df}(\lambda_n) \equiv |\,\mathcal{S}_0(\lambda_n)|$ is an unbiased estimate for $df(\lambda_n)$. In the following, we show that $\widehat{df}(\lambda_n)$ is also consistent.

Assumption 2 *There exists $\gamma > 0$ so that*

$$\min_i a_{ii}^n = O(n^\gamma), \quad for\ n \to \infty.$$

Lemma 1 *Assume that Assumptions 1 and 2 hold, and that $\lambda_n = O(n^\zeta)$ where $0 < \zeta < \gamma/2$, then,*

$$P(\,\mathcal{S}_0(\lambda_n) = \mathcal{S}_0) = 1, \quad for\ n \to \infty.$$

Theorem 2 *If $\frac{\lambda_n}{n^\zeta} \to \lambda^*$, then $\widehat{df}(\lambda_n) \to df(\lambda_n)$ in probability.*

Proof. From Lemma 1, $P(\,\mathcal{S}_0(\lambda_n) = \mathcal{S}_0) \to 1$. Immediately we see $\widehat{df}(\lambda_n) \to \mathcal{S}_0$ in probability. Then by the dominated convergence theorem, we have

$$df(\lambda_n) = E[\widehat{df}(\lambda_n)] \to |\,\mathcal{S}_0|.$$

Thus the theorem holds.

Based on the above discussion, the unbiased estimator for $\widehat{df}(\lambda_n)$ suffices to provide an unbiased estimate to the true prediction error of $\hat{\boldsymbol{\mu}}_{\lambda_n}$, as

$$C_p(\hat{\boldsymbol{\mu}}_{\lambda_n}) = \frac{\|\boldsymbol{Y}-\hat{\boldsymbol{\mu}}_{\lambda_n}\|^2}{n} + \frac{2}{n}|\,\mathcal{S}_0(\lambda_n)|\sigma^2.$$

Correspondingly, AIC for the modified LASSO is defined as follows:

$$AIC(\hat{\boldsymbol{\mu}}_{\lambda_n}) = \frac{\|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_{\lambda_n}\|^2}{n\sigma^2} + \frac{2}{n}|\,\mathcal{S}_0(\lambda_n)|.$$

Using AIC to find the optimal modified LASSO model, we introduce the following theorem to find the optimal λ_n where $AIC(\hat{\boldsymbol{\mu}}_{\lambda_n})$ get its minimum.

Theorem 3 *To find optimal $\lambda_n(optimal)$, we only need to solve*

$$m^* = \arg \min_{m \in \{0,1,\ldots,K\}} AIC(\hat{\boldsymbol{\mu}}_{\lambda_{nm}});$$

then $\lambda_n(optimal) = \lambda_{nm^*}$.

2.2 Variable selection porpuse In the least squares fit $\hat{\boldsymbol{\mu}}(\pi)$ for a given subset $\pi \subset \{1, \ldots, p\}$,

$$AIC(\hat{\boldsymbol{\mu}}(\pi)) = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\pi)\|^2}{n\sigma^2} + \frac{2}{n}|\pi|,$$

where $\hat{\boldsymbol{\beta}}(\pi)$ is the lest squares estimate of $\boldsymbol{\beta}^*$ where $\hat{\boldsymbol{\beta}}(\pi)_{\{1,\ldots,p\}-\pi} \equiv \mathbf{0}$, that is

$$\hat{\boldsymbol{\beta}}(\pi) = \arg \min_{\boldsymbol{\beta}:\beta_j = 0 for j \notin \pi} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2.$$

Then we define $\hat{\pi}$ as

$$\hat{\pi} = \arg \min_{\pi \subset \{1,\ldots,p\}} AIC(\hat{\boldsymbol{\mu}}(\pi)).$$

We say AIC is consistent if $P(\hat{\pi} = \mathcal{S}_0) \to 1$, as $n \to \infty$. Then by setting $\boldsymbol{x}_i^* = \boldsymbol{D}_n^{-1}\sqrt{n}\boldsymbol{x}_i$, where $\boldsymbol{D}_n = \text{diag}\{\sqrt{a_{11}^n}, \ldots, \sqrt{a_{pp}^n}\}$, for the original model

$$Y_i = \boldsymbol{x}_i'\boldsymbol{\beta}^* + \varepsilon_i,$$

it is transferred into

$$Y_i = (\boldsymbol{x}_i^*)'\boldsymbol{D}_n\boldsymbol{\beta}^*/\sqrt{n} + \varepsilon_i.$$

After the transformation, from Theorem 2 in [4], we can get that $AIC(\hat{\boldsymbol{\mu}}(\pi))$ is consistent. However, to find the $\hat{\pi}$, $O(2^n)$ times computational cost of a single least squares fit is needed. Define $\hat{m}$ as

$$\hat{m} = \arg\min_{m} AIC(\hat{\boldsymbol{\mu}}(\,\mathcal{S}_m)),$$

where m is the index of the transition point λ_{nm} of modified LASSO. Then the following theorem holds.

Theorem 4 *$P(\,\mathcal{S}_{\hat{m}} = \,\mathcal{S}_0) = 1$, as $n \to \infty$.*

Proof. From Lemma 1, among the transition points, the probability that there exists m such that $\mathcal{S}_0 = \,\mathcal{S}_m$ converges to 1. Noticing that $P(\hat{\pi} = \,\mathcal{S}_0) \to 1$, we get $P(\,\mathcal{S}_{\hat{m}} = \,\mathcal{S}_0) \to 1$.

From Theorem 4, the consistency of AIC on sub-models decided by modified LASSO approach is shown. Considering the computational cost can be reduced, it is reasonable to use $\mathcal{S}_{\hat{m}}$ to estimate $\mathcal{S}_0$.

3 Numerical results In this section, the simulation study analyses the selection and prediction performance of the modified LASSO when using the above two forms of AIC. We set $p = 8$ and $\boldsymbol{\beta} = \{1, 0, 1, 1, 0, 0, 0, 0\}$ i.e.

$$Y_i = x_{i1} + x_{i3} + x_{i4} + \varepsilon_i,$$

where the sequence $x_{i1} = 1$ for all $i \in \mathcal{N}$, $x_{i2} = i$, $x_{ij} = \cos\frac{\pi ij}{9}$ for $j = 3, \ldots, 8, i \in \mathcal{N}$ and $\{\varepsilon_i\}$ is generated by identically distributed Gaussian disturbances with length n going from 50 to 500 and variance $\sigma^2 = 0.1, 0.5, 1$ respectively. Here we use C_1 to stand for the $AIC(\hat{\boldsymbol{\mu}}_{\lambda_n})$, and use C_2 to stand for $AIC(\hat{\boldsymbol{\mu}}(\pi))$ for brevity. 100 replications are performed for each situation.
From table 1, we compare C_1 and C_2 by the bias and and mean squared error (MSE) of their estimators in the sense of parameter estimates. Here the bias and MSE are defined as follows:

$$Bias(\hat{\boldsymbol{\beta}}) = \frac{1}{s}\sum_{t=1}^{s}\sum_{j=1}^{8}(\hat{\beta}_{tj} - \beta_j^*);$$

$$MSE(\hat{\boldsymbol{\beta}}) = \frac{1}{s}\sum_{t=1}^{s}\sum_{j=1}^{8}(\hat{\beta}_{tj} - \beta_j^*)^2,$$

where s is the amount of replications. It is shown that the prediction performance of modified LASSO with C_1 is better than that of modified LASSO with C_2, noticing that both of absolute value of bias and MSE of C_1 are smaller than those of C_2. Besides, we can notice that, with sequence length n increases, the performance of C_2 gets worse. It agrees to the condition of consistency, that the optimal λ_n increases as n increases. From Table 2, the results from five aspects in the sense of variable selection are shown, which are the probability of correct selection, the probability of relevant variables included, the probability of irrelevant variables excluded, average number of included variables and average number of included irrelevant variables. It is shown that the results by C_2 are better than those by C_1 overall. From the probability of true model included, the probabilities by C_2 are greater than those by C_1. Besides, by comparing the values as n increases, it is shown that the the probability of correct selection of the true model increases, which keep consist with the consistency properties shown in Section 2. From the probability of relevant variables included, almost all the probabilities by C_1 and C_2 are 1, which means that by both C_1 and C_2, the probabilities that relevant variables are excluded are low.

Table 1: Parameter estimates

	C_1				C_2			
n	50	100	300	500	50	100	300	500
Bias								
N(0.1)	-0.1107	-0.0687	-0.0511	-0.0434	-1.4880	-1.4839	-1.5848	-1.6751
N(0.5)	-0.2390	-0.1712	-0.1010	-0.0903	-1.4646	-1.5184	-1.5618	-1.5790
N(1)	-0.3724	-0.2915	-0.1792	-0.1323	-1.5117	-1.5352	-1.5453	-1.5854
MSE								
N(0.1)	0.0221	0.0115	0.0042	0.0024	1.3915	1.3944	1.5361	1.6444
N(0.5)	0.1090	0.0522	0.0180	0.0120	1.2837	1.3877	1.4639	1.5000
N(1)	0.2365	0.1135	0.0412	0.0231	1.3124	1.3421	1.4166	1.4922

Table 2: Veriable selection

	C_1				C_2			
n	50	100	300	500	50	100	300	500
Probability of correct selection								
N(0.1)	0.53	0.55	0.55	0.57	0.74	0.73	0.79	0.84
N(0.5)	0.47	0.52	0.48	0.55	0.70	0.74	0.77	0.78
N(1)	0.35	0.41	0.46	0.49	0.68	0.71	0.75	0.78
Probability of relevant variables included								
N(0.1)	1	1	1	1	1	1	1	1
N(0.5)	1	1	1	1	1	1	1	1
N(1)	1	1	1	1	0.99	1	1	1
Probability of irrelevant excluded								
N(0.1)	0.53	0.55	0.55	0.57	0.74	0.73	0.79	0.84
N(0.5)	0.47	0.52	0.48	0.55	0.70	0.74	0.77	0.78
N(1)	0.35	0.41	0.46	0.49	0.68	0.71	0.75	0.78
Average number of included variables								
N(0.1)	3.81	3.80	3.66	3.56	3.33	3.34	3.23	3.16
N(0.5)	3.93	3.83	3.94	3.64	3.35	3.36	3.26	3.25
N(1)	4.11	3.96	3.79	3.75	3.41	3.33	3.25	3.24
Average number of included irrelevant variables								
N(0.1)	0.81	0.8	0.66	0.56	0.33	0.34	0.23	0.16
N(0.5)	0.93	0.83	0.94	0.64	0.35	0.36	0.26	0.25
N(1)	1.11	0.96	0.79	0.75	0.41	0.33	0.25	0.24

4 Conclusion In Sections 2 and 3, the prediction performance of C_1 is better than C_2. Whereas, the selection performance of the later is better than the former. Since these two forms of AIC are derived by the ideas of [8] and [4], in fact a more general form of the criteria can be derived. Besides, noticing that the consistency depends on the consistency in the sense that the probability of correct selection of the true model converges to 1 as the sequence length n goes to infinity, the consistency on high dimensions can be discussed furthermore.

Acknowledgements I would like to express my sincere gratitude to my supervisor Prof. Masanobu Taniguchi for the continuous support of my study and research, for his insightful and detailed comments to help improve the paper. Besides, I would like to thank the support by JSPS funding: Kiban(S)(18H05290) as well.

References

[1] Akaike, H. (1998) Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike, Springer, New York, NY, 199-213.

[2] Hastie, T., Tibshirani, R. & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[3] Hurvich, C.M. & Tsai, C.L. (1989). Regression and time series model selection in small samples. Biometrika, 76(2), 297-307.

[4] Kim, Y., Kwon, S. & Choi, H. (2012). Consistent model selection criteria on high dimensions. J Mach Learn Res, 13, 1037-1057.

[5] Mallows, C.L. (2000). Some comments on Cp. Technometrics, 42(1), 87-94.

[6] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. Ann Stat, 9, 1135-1151.

[7] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J R Stat Soc Series B, 58, 267-288.

[8] Zou, H., Hastie, T., &Tibshirani, R. (2007). On the degrees of freedom of the lasso. Ann Stat, 35(5), 2173-2192.

[9] Xue, Y. & Taniguchi, M. (2020). Modified LASSO estimators for time series regression models with dependent disturbances. Stat Methods Appt, DOI: 10.1007/s10260-020-00506-w.

[10] Ye, J. (1998). On Measuring and Correcting the Effects of Data Mining and Model Selection. J Am Stat Assoc, 93, 120131.

Appendix A. Proof of Theorem 1 To prove Theorem 1, we introduce the following lemma. For given $\boldsymbol{Y}$, there exists a set of transition point, $\{\lambda_{nm} : m = 0, \dots, K\}$. Recall the definition of notations, $\mathcal{S}_m = \mathcal{S}_0(\lambda_{nm})$, and $\mathrm{sgn}_m = \mathrm{sgn}(\lambda_{nm})$.

Lemma 2 *Suppose $\lambda_n \in (\lambda_{n,m+1}, \lambda_{nm})$. Then we have*

$$\hat{\boldsymbol{\beta}}(\lambda_n)_{\mathcal{S}_m} = (\boldsymbol{X'}_{\mathcal{S}_m} \boldsymbol{X}_{\mathcal{S}_m})^{-1} (\boldsymbol{X'}_{\mathcal{S}_m} \boldsymbol{Y} - \frac{\lambda_n}{2} \boldsymbol{D}_n \mathrm{sgn}_m),$$

where $\boldsymbol{D}_n = \mathrm{diag}\{\sqrt{a^n_{11}(0)}, \dots, \sqrt{a^n_{pp}(0)}\}$.

Lemma 3 *Consider the transition point λ_{nm}, when λ_n decreases from the right hand side of λ_{nm} to λ^-_{nm}, i_{add} is an index added into $\mathcal{S}_m$, and the order index of i_{add} is i^*, that is, $i_{add} = (\mathcal{S}_m)_{i^*}$. Denote the kth element of any vector $\boldsymbol{a}$ by $(\boldsymbol{a})_k$. We can express the transition point λ_{nm} as*

$$\lambda_{nm} = \frac{2((\boldsymbol{X'}_{\mathcal{S}_m} \boldsymbol{X}_{\mathcal{S}_m})^{-1} \boldsymbol{X'}_{\mathcal{S}_m} \boldsymbol{Y})_{i^*}}{((\boldsymbol{X'}_{\mathcal{S}_m} \boldsymbol{X}_{\mathcal{S}_m})^{-1} \boldsymbol{D}_n^{-1} \mathrm{sgn}_m)_{i^*}}.$$

Lemma 4 *For any $\lambda_n > 0$, there exists a null set $\mathcal{N}_{\lambda_n}$ which is a finite collection of hyperplanes in $\mathbb{R}^n$. Let $\mathrm{G}_{\lambda_n} = \mathbb{R}^n - \mathcal{N}_{\lambda_n}$. Then $\forall \boldsymbol{Y} \in \mathrm{G}_{\lambda_n}$, λ_n is not any of the transition points for $\boldsymbol{Y}$.*

Lemma 5 *$\forall \lambda_n > 0$, $\hat{\boldsymbol{\beta}}(\lambda_n)$ is a continuous function with respect to $\boldsymbol{Y}$.*

Lemma 6 *Fix any $\lambda_n > 0$ and consider $\boldsymbol{Y} \in \mathrm{G}_{\lambda_n}$ as defined in Lemma 4. The active set $\mathcal{S}_0(\lambda_n)$ and the sign vector $\mathrm{sgn}(\lambda_n)$ are locally constant with respect to $\boldsymbol{Y}$.*

Lemma 7 *Let $\mathrm{G}_0 = \mathbb{R}^n$. For any $\lambda_n \geq 0$, on the set G_{λ_n} as defined in Lemma 4, the modified LASSO fit $\hat{\boldsymbol{\mu}}_{\lambda_n}(\boldsymbol{Y})$ is uniformly Lipschitz. Precisely,*

$$\|\hat{\boldsymbol{\mu}}_{\lambda_n}(\boldsymbol{Y} + \Delta \boldsymbol{Y}) - \hat{\boldsymbol{\mu}}_{\lambda_n}(\boldsymbol{Y})\| \leq \|\Delta \boldsymbol{Y}\|,$$

for sufficiently small $\Delta \boldsymbol{Y}$. Moreover, we have the divergence formula

$$\nabla \cdot \hat{\mu}_{\lambda_n}(\boldsymbol{Y}) = |\mathcal{S}_0(\lambda_0)|,$$

where $|\mathcal{S}_0(\lambda_n)|$ stands for the cardinality of $\mathcal{S}_0(\lambda_n)$.

The proofs of Lemma 2 to 7 are similar to those in [8], here we omit the proofs.
Proof of Theorem 1: By Lemma 4-7, $\hat{\mu}_{\lambda_n}(\boldsymbol{Y})$ is differentiable almost every where. Then by the Stein's unbiased risk estimation theory ([6]) and Lemma 7,

$$df(\lambda_n) = \mathrm{E} \nabla \cdot \hat{\mu}_{\lambda_n}(\boldsymbol{Y}) = \mathrm{E}|\mathcal{S}_0(\lambda_n)|.$$

Thus Theorem 1 holds.

Appendix B. Proof of Lemma 1
Noticing that $\{\varepsilon_i\}$ is independent and identically distributed process with $\varepsilon_i \sim N(0, \sigma^2)$, then $\sum_{i=1}^n b_{nij} \varepsilon_i$ satisfies the Bernstein inequality, where $b_{nij} = \frac{x_i^j}{\sqrt{a^n_{ii}(0)}}$ for $j = 1, \dots, p$. Thus, Lemma 1 is a special case of Theorems 4 and 5 in [9], which implies it holds.

Appendix C. Proof of Theorem 3

From the forms of AIC, we know that

$$\lambda_n(optimal) = \arg\min_{\lambda_n} AIC(\hat{\boldsymbol{\mu}}_{\lambda_n}) = \arg\min_{\lambda_n} \frac{\|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_{\lambda_n}\|^2}{n\sigma^2} + \frac{2}{n}|\,\mathcal{S}_0(\lambda_n)|.$$

From Lemma 2, for $\lambda_n \in (\lambda_{n,m+1}, \lambda_{nm})$, we have

$$\|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_{\lambda_n}\|^2 = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{X}_{\mathcal{S}_m}(\boldsymbol{X}'_{\mathcal{S}_m}\boldsymbol{X}_{\mathcal{S}_m})^{-1}\boldsymbol{X}'_{\mathcal{S}_m})\boldsymbol{Y} + \frac{\lambda_n^2}{4}\mathrm{sgn}'_m\boldsymbol{D}_n(\boldsymbol{X}'_{\mathcal{S}_m}\boldsymbol{X}_{\mathcal{S}_m})^{-1}\boldsymbol{D}_n\mathrm{sgn}_m,$$

where I is the $n \times n$ identity matrix. Thus we conclude that in the interval $(\lambda_{n,m+1}, \lambda_{nm})$, $\|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}_{\lambda_n}\|^2$ is strictly increasing with respect to λ_n. On the other hand, note that $|\,\mathcal{S}_0(\lambda_{nm})| \geq |\,\mathcal{S}_0(\lambda_{n,m+1})|$. Therefore, the optimal choice of λ_n in $[\lambda_{n,m+1}, \lambda_{nm})$ is $\lambda_{n,m+1}$, which means $\lambda_n(optimal) \in \{\lambda_{nm} : m + 0, \ldots, K\}$. Thus Theorem 3 holds.

Communicated by *Masanobu Taniguchi*