Published in final edited form as:

J Bus Econ Stat. 2015 October 1; 33(4): 541–551. doi:10.1080/07350015.2014.965307.

Adaptive Modeling Procedure Selection by Data Perturbation*

Yongli Zhang¹ and Xiaotong Shen²

¹Assistant Professor, Lundquist College of Business, University of Oregon, 1208 University Ave, Eugene, OR 97403 (yongli@uoregon.edu)

²Professor, School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, MN 55455 (xshen@stat.umn.edu)

Summary

Many procedures have been developed to deal with the high-dimensional problem that is emerging in various business and economics areas. To evaluate and compare these procedures, modeling uncertainty caused by model selection and parameter estimation has to be assessed and integrated into a modeling process. To do this, a data perturbation method estimates the modeling uncertainty inherited in a selection process by perturbing the data. Critical to data perturbation is the size of perturbation, as the perturbed data should resemble the original dataset. To account for the modeling uncertainty, we derive the optimal size of perturbation, which adapts to the data, the model space, and other relevant factors in the context of linear regression. On this basis, we develop an adaptive data-perturbation method that, unlike its nonadaptive counterpart, performs well in different situations. This leads to a data-adaptive model selection method. Both theoretical and numerical analysis suggest that the data-adaptive model selection method adapts to distinct situations in that it yields consistent model selection and optimal prediction, without knowing which situation exists a priori. The proposed method is applied to real data from the commodity market and outperforms its competitors in terms of price forecasting accuracy.

Keywords

Adaptive model selection; high-dimensional data analysis; modeling uncertainty

1 Introduction

A modeling procedure typically proceeds in two steps, model selection and parameter estimation. Yet randomness inherited in a model selection process, which is referred to as model selection uncertainty (e.g., [7]), is often ignored in assessing its performance and statistical inference. For instance, in linear regression, a model is selected using an information criterion such as AIC (Akaike information criterion; [1]) and inference and risk evaluation are performed as if the model were pre-chosen [7]. This is viewed as “a quiet scandal in the statistical community” by [3]. Our primary goal here is to construct a general tool for assessing a modeling procedure’s performance that takes model selection

*The research is supported in part by National Science Foundation Grant DMS-0906616 and National Institute of Health Grant 1R01GM081535. The authors thank the editor, the associate editor and two referees for their helpful comments and suggestions.

uncertainty into account. On this ground, we will develop a procedure of adaptive model selection in linear regression.

The performance of a modeling procedure is usually evaluated by its risk, but risk estimation becomes nontrivial due to intractability of parameter estimators as a selected model is treated as a discretely distributed random variable [7]. For risk estimation, three types of resampling-based methods have emerged so far: cross-validation (CV), bootstrap, and data-perturbation. The CV is a nearly unbiased method for estimating the prediction risk [22, 11, 5, 26]. The bootstrap uses resampling to estimate covariance penalty, which is more precise than CV, as shown by [6], in that the covariance penalty has the same amount of bias with a smaller variance. Similarly, the data perturbation estimates the generalized degrees of freedom, or scaled covariance penalty, which again leads to a more precise estimate than CV [3, 27, 19].

In this article, we focus on data perturbation whose performance depends on the perturbation size. In particular, we develop a method to estimate the optimal perturbation size for evaluating a modeling procedure's risk to account for model selection uncertainty, and derive an adaptive model selection scheme. Several results are established here. First, we provide a theoretical justification for data perturbation as a tool for modeling procedure assessment, although some empirical rules have been available in [3, 27, 19, 28]. Second, on this ground, we construct our adaptive model selection routine (ADPC). Throughout this article the true model is assumed in the model space, but BIC (Bayesian information criterion; [17]) is not necessarily consistent as discussed in [4] and [30], and a larger penalty coefficient such as $\sqrt{n}$ or a smaller penalty such as $\log \log n$ may outperform $\log n$ of BIC. Therefore, a sensible strategy in model selection is to select the best penalty coefficient from a set of candidates and then use the best one for model selection. ADPC achieves this strategy. Empirically, we demonstrate that ADPC is more precise than its nonadaptive counterpart such as BIC. Theoretically, we show that ADPC is model selection consistent.

The rest of the article is organized as follows. Section 2 studies the properties of modeling procedure risk, develops an estimation method for modeling procedure risk based on data perturbation, and derives the optimal perturbation size. On this basis, an adaptive model selection criterion (ADPC) is proposed in Section 3. Section 4 is devoted to simulations to examine operating characteristics of ADPC and comparison against its competitors, in addition to an application to forecasting nodal hourly power price in the state of New York. Section 5 concludes this article, followed by technical proofs in Section 6.

2 Definition and Estimation of Modeling Procedure Risk

2.1 Modeling procedure and its risk

Consider linear models in which response Y depends on predictors $x_1, \dots, x_p$ as follows:

$$Y = \beta_0 + \sum_{x_j \in M_0} \beta_j x_j + \epsilon, \quad (1)$$

where $\beta_j, j = 1, \dots, p$, are regression parameters, M_0 denotes the true or correct model that is comprised of informative predictors, and p may greatly exceed the sample size n . Our goal is to estimate regression function $\mu = E(Y|x_1, \dots, x_p)$ as well as M_0 from a pool of candidate models M . Throughout this work we assume $M_0 \in \mathcal{M}$, so our focus is selection consistency. Let $|M|$ denote the cardinality of a model and let $|M_0| = q_0$, which is assumed $< n$. Throughout this article q_0 and p are allowed to grow with $n \rightarrow \infty$. Here M can be partitioned into $\mathcal{M} = \mathcal{M}^F \cup \mathcal{M}^T$, consisting of underfitting models and overfitting models (including M_0), respectively. Assume that $E\epsilon = 0$, $E\epsilon^2 = \sigma^2$, and $E\epsilon^{2k} < \infty$ for $k = 1, 2, \dots$.

Given n observations (Y_n, X) with X denoting a design matrix, many selection criteria have been proposed for model selection. Essentially all of them are in the form

$$IC_\lambda(M) = \|\hat{\mu}_n(M) - Y_n\|_2^2 + \lambda|M|\sigma^2, \quad (2)$$

where $M \subseteq \{x_1, \dots, x_p\}$, $\hat{\mu}_n(M) = P(M)Y_n$, and $P(M)$ is the projection matrix defined by design matrix $X(M)$ with predictors contained in M . The form (2) includes, but is not limited to, AIC with $\lambda = 2$, BIC with $\lambda = \log n$, risk inflation criterion (RIC, [9]) with $\lambda = 2 \log p$, the corrected risk inflation criterion (RIC_c, [30]) with $\lambda = 2 \log p + 2 \log \log p$, and the

covariance inflation criterion (CIC, [24]) with $\lambda = \frac{4}{|M|} \sum_{j=1}^{|M|} \log(p/j)$. In (2), different λ leads to dramatically different performances [18]. The need for data-dependent λ -values has been emphasized in [12, 19].

To take model selection into account, we consider a modeling procedure defined by the least squares estimate

$$\hat{\mu}_n(\hat{M}_\lambda) = P(\hat{M}_\lambda)Y_n \quad (3)$$

for a model $\hat{M}_{\mathcal{M}, \lambda}$, simplified as $\hat{M}_\lambda$, obtained by minimizing (2) over a pool of candidate models M . Consequently, the randomness in $\hat{\mu}_n(\hat{M}_\lambda)$ comprises of the estimation randomness due to $\hat{\mu}_n(M)$ and selection randomness due to $\hat{M}_\lambda$. For the sake of tuning parameter selection it is crucially important to develop a method of evaluating and comparing the performance of (3).

The performance of a modeling procedure (3) is assessed by its risk:

$$R_n(\hat{M}_\lambda) = E\|\hat{\mu}_n(\hat{M}_\lambda) - \mu_n\|_2^2, \quad (4)$$

where $\mu_n = E(Y_n|x_1, \dots, x_p)$ and the expectation is taken with respect to Y_n given $x_1, \dots, x_p$. Throughout this article we assume $\limsup_{n \rightarrow \infty} n^{-1} \|\mu_n\|_2^2 < \infty$. The following example illustrates the difference and connection between $R_n(\hat{M}_\lambda)$ and $R_n(M)$ in absence of model selection.

Example 1. Consider model (1) with two predictors x_1 and x_2 , where three models are considered: $M_1 = (1, x_1)$ is the true model, $M_2 = (1, x_1, x_2)$ is an over-fitting model, and $M_3 = (1, x_2)$ is an underfitting model. Assume $\epsilon \sim N(0, 1)$. Three cases are examined. In Case 1, $R_n(\hat{M}_\lambda) = 2\sigma^2$ when $\mathcal{M} = \{M_1\}$ and no model selection is involved. In Case 2, $R_n(\hat{M}_{\lambda=0}) = 3\sigma^2$ when $\mathcal{M} = \{M_1, M_2, M_3\}$ with $\lambda = 0$ in (2). This is because M_2 is always selected. In Case 3, $R_n(\hat{M}_\lambda) = \sigma^2 \left(2 + P(\chi_3^2 > 1)\right)$ when $\mathcal{M} = \{M_1, M_2\}$, where χ_3^2 denotes a chi-square random variable on three degrees of freedom (the technique details are deferred to Section 6.1).

In view of Cases 1–3, it appears that the risk of a modeling procedure depends on M and λ , and a small λ results in high $R_n(\hat{M}_\lambda)$ for a given M . Next we investigate the properties of $R_n(\hat{M}_\lambda)$ in a more general setting. The modeling procedure risk can be decomposed as

$$R_n(\hat{M}_\lambda) = E\|Q(\hat{M}_\lambda)\mu_n\|_2^2 + E\|P(\hat{M}_\lambda)\epsilon_n\|_2^2, \quad (5)$$

where $Q(M) = I - P(M)$ (I is an $n \times n$ identity matrix), and ϵ_n are n iid realizations of ϵ in (1). In (5), $E\|Q(\hat{M}_\lambda)\mu_n\|_2^2$ may be regarded as the approximation error of $\hat{M}_\lambda$, while $E\|P(\hat{M}_\lambda)\epsilon_n\|_2^2$ standardized by σ^2 ,

$$U_{\mathcal{M}}(\lambda) = \sigma^{-2} E\|P(\hat{M}_\lambda)\epsilon_n\|_2^2 = \sigma^{-2} \sum_{M \in \mathcal{M}} E\left(\epsilon_n' P(M) \epsilon_n \mathbb{I}_{\{\hat{M}_\lambda = M\}}\right), \quad (6)$$

can be interpreted as the complexity of a modeling procedure. Throughout this article we assume that $(\sigma^2 q_0 \log p) / \min_{M \in \mathcal{M}^F} \|Q(M)\mu_n\|_2^2 \rightarrow 0$ as $n \rightarrow \infty$ to guard against underfitting.

The modeling procedure complexity $U_{\mathcal{M}}(\lambda)$ defined in (6) is closely related to the generalized degree of freedom (GDF, [27]), the covariance penalty [6], the effective number of parameters (ENP, [24]), and the perturbation instability in estimation (PIE, [28]). Although these measures aim to capture model selection uncertainty, the covariance penalty and the effective degrees of freedom could become negative, with the covariance penalty unit-dependent. In contrast, $U_{\mathcal{M}}(\lambda)$ is always nonnegative and unit-free.

Depending on if $\hat{M}_\lambda$ belongs to $\mathcal{M}^T$ or $\mathcal{M}^F$, the modeling procedure complexity (6) can be

$$\begin{aligned}\sigma^2 U_{\mathcal{M}}(\lambda) &= E \left\| P \left(\hat{M}_\lambda \right) \epsilon_n \right\|_2^2 I_{\{\hat{M}_\lambda \in \mathcal{M}^F\}} \\ &\quad + E \left\| P \left(\hat{M}_\lambda \right) \epsilon_n \right\|_2^2 I_{\{\hat{M}_\lambda \in \mathcal{M}^T\}} \\ &= E \left(\left\| P \left(\hat{M}_\lambda \right) \epsilon_n \right\|_2^2 - \left\| P \left(\hat{M}_{\mathcal{M}^T, \lambda} \right) \epsilon_n \right\|_2^2 \right) I_{\{\hat{M}_\lambda \in \mathcal{M}^F\}} \\ &\quad + \sigma^2 U_{\mathcal{M}^T}(\lambda),\end{aligned}$$

decomposed further as $\sigma^2 U_{\mathcal{M}^T}(\lambda)$, where

$$\sigma^2 U_{\mathcal{M}^T}(\lambda) = E \left\| P \left(\hat{M}_{\mathcal{M}^T, \lambda} \right) \epsilon_n \right\|_2^2. \text{ Therefore,}$$

$$R_n(\hat{M}_\lambda) = E \left(\left\| Q \left(\hat{M}_{\mathcal{M}^F, \lambda} \right) \mu_n \right\|_2^2 + \left\| P \left(\hat{M}_{\mathcal{M}^F, \lambda} \right) \epsilon_n \right\|_2^2 - \left\| P \left(\hat{M}_{\mathcal{M}^T, \lambda} \right) \epsilon_n \right\|_2^2 \right) I_{\{\hat{M}_\lambda \in \mathcal{M}^F\}} + \sigma^2 U_{\mathcal{M}^T}(\lambda) \quad (7)$$

where $\sigma^2 U_{\mathcal{M}^T}(\lambda)$ dominates the risk $R_n(\hat{M}_\lambda)$ as underfitting does not occur, that is, the following condition is satisfied.

- *Condition A:* For some $k \geq 4$,

$$\sum_{M_f \in \mathcal{M}^F} \frac{n((\lambda - 1)(q_0 + 1)\sigma^2)^{k/4}}{\left(\left\| Q(M_f) \mu_n \right\|_2^2 + (\lambda - 1) |M_f| \sigma^2 \right)^{k/4}} \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (8)$$

Condition A requires that the underfitting error resulted from an overly large λ is ignorable as model selection is performed by (2) for (X, Y_n) . Sometime, it suffices to verify a stronger but simpler assumption, Condition B, which states:

- *Condition B:* For some $k \geq 4$,

$$|\mathcal{M}^F| \frac{n((\lambda - 1)(q_0 + 1)\sigma^2)^{k/4}}{\left(\min_{M_f \in \mathcal{M}^F} \left\| Q(M_f) \mu_n \right\|_2^2 \right)^{k/4}} \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (9)$$

Condition B requires that the number of underfitting models $|\mathcal{M}^F|$ is ignorable relative to the least favorable approximation error of the wrong candidate models $\min_{M_f \in \mathcal{M}^F} \left\| Q(M_f) \mu_n \right\|_2^2$. On one hand, Conditions A and B can accommodate more underfitted models when k or the approximation error becomes large. On the other hand, they guard against underfitting as λ or q_0 increases.

LEMMA 1 Assume that Condition A holds. Then for a modeling procedure $\hat{\mu}_n(\hat{M}_\lambda)$, $R_n(\hat{M}_\lambda) - \sigma^2 U_{\mathcal{M}^T}(\lambda) \rightarrow 0$ as $n \rightarrow \infty$. Moreover, $U_{\mathcal{M}^T}(\lambda)$ is decreasing in λ .

Lemma 1 confirms the phenomenon in Example 1, that is, the modeling procedure risk (4) is a decreasing function of λ . The closeness between $R_n(\hat{M}_\lambda)$ and $\sigma^2 U_{\mathcal{M}^T}(\lambda)$ provides a strategy of estimating $R_n(\hat{M}_\lambda)$ based on data perturbation.

2.2 Estimation of modeling procedure risk by data perturbation

The optimal modeling procedure or λ can be selected by minimizing $R_n(\hat{M}_\lambda)$ over Γ , a set of candidate λ values. Thus we need an estimate of $R_n(\hat{M}_\lambda)$ for a given λ . However, $R_n(\hat{M}_\lambda)$ may depend on model selection routines in use, which is difficult to derive an analytic form. We therefore employ the data perturbation scheme [19]. As the first step we derive an estimate of $U_{\mathcal{M}^T}(\lambda)$ due to its dominant role in (4) as indicated by Lemma 1.

Our proposed method proceeds as follows. Generate m samples $Y_{n,j}^\tau (j=1, \dots, m)$ according to $Y_{n,j}^\tau = Y_n + \tau \epsilon_{n,j}^*$, where $\epsilon_{n,j}^*$ comprises of n iid random entries distributed as $N(0, \sigma^2)$ and is independent of ϵ_n , and τ is the perturbation size to be chosen to achieve our goal of accurate selection of tuning parameters. The pseudo data Y_n^τ has the following characteristics:

$E(Y_{n,j}^\tau | Y_n) = Y_n$ and $\text{var}(Y_{n,j}^\tau | Y_n) = \tau^2 \sigma^2 I$. Now define a model selection criterion for the pseudo data $Y_{n,j}^\tau$ based on (2):

$$IC_{\lambda,j}^\tau(M) = \|Q(M) Y_{n,j}^\tau\|_2^2 + \lambda |M| (1 + \tau^2) \sigma^2, \quad (10)$$

where $\text{var}(Y_{n,j}^\tau) = (1 + \tau^2) I \sigma^2$ as opposed to $\text{var}(Y_n) = I \sigma^2$. Let

$\hat{M}_{\lambda,j}^\tau = \argmin_{M \in \mathcal{M}} IC_{\lambda,j}^\tau(M)$ and $\hat{M}_{\lambda,j,t}^\tau = \argmin_{M \in \mathcal{M}^T} IC_{\lambda,j}^\tau(M)$ are selected models based on $Y_{n,j}^\tau$ over M and $\mathcal{M}^T$, respectively. Note that the m samples are dependent but approximately independent as $\tau \rightarrow \infty$. In other words, the inter-sample dependency effect due to Y_n is eliminated as $\tau \rightarrow \infty$.

This yields our estimation scheme for $U_{\mathcal{M}}(\lambda)$.

Step 1. Generate m random vectors $Y_{n,1}^\tau, \dots, Y_{n,m}^\tau$ as described.

Step 2. Compute

$$\hat{U}_{\mathcal{M}}^\tau(\lambda) = \frac{\sum_{j=1}^m \|P(\hat{M}_{\lambda,j}^\tau) Y_{n,j}^\tau - P(\hat{M}_{\lambda,j}^\tau) Y_n\|_2^2}{m \tau^2 \sigma^2}. \quad (11)$$

As $m \rightarrow \infty$, (11) approximates $\sigma^{-2} E \left(\|P(\hat{M}_{\lambda,j}^\tau) \epsilon_n^*\|_2^2 | Y_n \right)$ by the law of large numbers.

Clearly, the performance of (11) as an estimate of $U_{\mathcal{M}}(\lambda)$ is dependent on the perturbation size τ . As discussed above a large τ is desired to eliminate the inter-sample dependency across $Y_{n,j}^\tau (j=1, \dots, m)$, but an overly large τ may result in underfitting, that is, the probability of selecting wrong models by (10) becomes nonignorable. Next we examine the relationship between the perturbation size τ and the accuracy of (11) with regard to estimation of $U_{\mathcal{M}}(\lambda)$ in Case 3 of Example 1, where $U_{\mathcal{M}^T}(\lambda) = (2 + P(\chi_3^2 > \lambda))$, so

$U_{\mathcal{M}^T}(\lambda=2)=2.57$, $U_{\mathcal{M}^T}(\lambda=\log 50)=2.27$ and $U_{\mathcal{M}^T}(\lambda=\sqrt{50})=2.07$ for $n=50$. Suppose each row in the 50×2 design matrix X is generated from $N(0, \Sigma_{2 \times 2})$ independently, where the diagonals of $\Sigma_{2 \times 2}$ are 1 and the off-diagonals equal to $\rho = -0.5$. Moreover, the response vector Y_n is obtained through $Y_n = 2X[1] + \epsilon_n$, where $X[1]$ is the first column of X , and ϵ_n are n iid random variables following $N(0, 1)$. Let $m = 50$.

The average MSE of (11) is used to measure the estimation accuracy of (11), which is

computed by averaging $\left(\hat{U}_{\mathcal{M}}^{\tau}(\lambda) - U_{\mathcal{M}}(\lambda)\right)^2$ of 200 replications. Consider two cases: (1) $\mathcal{M}^F$ is empty and (2) $\mathcal{M}^F$ is nonempty and the relationship between τ and average MSE of (11) is shown in Figure 1.

As suggested in the left plot of Figure 1, as $\mathcal{M}^F$ is empty the estimates of $U_{\mathcal{M}}(\lambda)$ approach to its true value with an increasing τ , while as $\mathcal{M}^F$ is nonempty the average MSE of (11) displays a decreasing-then-increasing pattern as τ increases. At the stage of decreasing, estimation accuracy improves because an increased perturbation size eliminates the inter-sample dependency due to Y_n . At the second stage, the wrong model is selected more often than the first stage resulting in an increase in estimation error. Another finding is that for a large λ , a small perturbation size is adequate for accurate estimation of $U_{\mathcal{M}}(\lambda)$, while a small λ implies large model selection uncertainty, and thus requires a large τ . In conclusion, an ideal perturbation size should be inversely proportional to λ and as large as possible without causing underfitting.

2.3 The optimal perturbation size

As depicted above, according to whether $\hat{M}_{\lambda}^{\tau}$ belongs to $\mathcal{M}^T$ or $\mathcal{M}^F$ the bias of $\hat{U}_{\mathcal{M}}^{\tau}(\lambda)$ (11), namely, $Err = E\left(\hat{U}_{\mathcal{M}}^{\tau}(\lambda) - U_{\mathcal{M}^T}(\lambda)\right) = E\hat{U}_{\mathcal{M}}^{\tau}(\lambda) - U_{\mathcal{M}^T}(\lambda)$, is in turn decomposed into

$$Err = \frac{1}{\sigma^2} E\left\|P\left(\hat{M}_{\lambda}^{\tau}\right) \epsilon_n^*\right\|_2^2 \left(\mathbf{I}_{\{\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^F\}} + \mathbf{I}_{\{\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^T\}}\right) - U_{\mathcal{M}^T}(\lambda) = Err_t + Err_f \quad (12)$$

$$Err_t = \frac{1}{\sigma^2} E\left\{\left\|P\left(\hat{M}_{\lambda,t}^{\tau}\right) \epsilon_n^*\right\|_2^2 - \left\|P\left(\hat{M}_{\lambda,t}^{\diamond}\right) \epsilon_n^*\right\|_2^2\right\} \mathbf{I}_{\{\hat{M}_{\lambda,t}^{\tau} \neq \hat{M}_{\lambda,t}^{\diamond}\}} \quad (13)$$

$$Err_f = \frac{1}{\sigma^2} E\left\{\left\|P\left(\hat{M}_{\lambda,f}^{\tau}\right) \epsilon_n^*\right\|_2^2 - \left\|P\left(\hat{M}_{\lambda,t}^{\tau}\right) \epsilon_n^*\right\|_2^2\right\} \mathbf{I}_{\{\hat{M}_{\lambda,t}^{\tau} \in \mathcal{M}^F\}} \quad (14)$$

where $\hat{M}_{\lambda,t}^{\diamond} = \operatorname{argmin}_{M \in \mathcal{M}^T} \|Q(M) \epsilon_n^*\|_2^2 + \lambda|M|\sigma^2$,

$\hat{M}_{\lambda,t}^{\tau} = \operatorname{argmin}_{M \in \mathcal{M}^T} \|Q(M) (\epsilon_n + \tau \epsilon_n^*)\|_2^2 + \lambda|M|(1+\tau^2)\sigma^2$ and

$\hat{M}_{\lambda,f}^{\tau} = \operatorname{argmin}_{M \in \mathcal{M}^F} \|Q(M) (\epsilon_n + \tau \epsilon_n^*)\|_2^2 + \lambda|M|(1+\tau^2)\sigma^2$.

Obviously, $\mathbf{I}_{\{\hat{M}_{\lambda,t}^{\tau} \neq \hat{M}_{\lambda,t}^{\diamond}\}} \rightarrow 0$ in probability as $\tau \rightarrow \infty$ for a given λ , $\mathcal{M}^T$ and sample as indicated by (13). Hence, Err_t is the estimation error when $\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^T$ in which a large τ is

desired to eliminate the effect of ϵ_n , while Err_f is the error resulted from underfitting introduced by $\tau\epsilon_n^*$, because (10) may select a wrong model with a nonignorable probability as τ becomes overly large. In other words, the choice of τ needs to balance the two aspects, which is one key focus.

In what follows, we derive the upper bound of these two types of errors because Err_t and Err_f are analytically intractable,

$$|\text{Err}_t| \leq \sqrt{E\|\epsilon_n^*/\sigma\|_4^2} \sqrt{P(\hat{M}_{\lambda,t}^\tau \neq \hat{M}_{\lambda,t}^\diamond)}, \quad |\text{Err}_f| \leq \sqrt{E\|\epsilon_n^*/\sigma\|_2^4} \sqrt{P(\hat{M}_\lambda^\tau \in \mathcal{M}^F)}, \quad (15)$$

where $P(\hat{M}_{\lambda,t}^\tau \neq \hat{M}_{\lambda,t}^\diamond)$ is a decreasing function of τ . Contrarily, $P(\hat{M}_\lambda^\tau \in \mathcal{M}^F)$ is an increasing function of τ .

Concerning the upper bound of $|\text{Err}|$, we have the following result.

LEMMA 2 For a given $\lambda \geq 2$, some $k \geq 4$ and a large enough τ

$$|\text{Err}| \leq d(4k)n \left(\left(\frac{16}{(1+\tau^2)(\lambda-1)} \right)^{k/4} + \left(\frac{(1+\tau^2)(\lambda-1)(q_0+1)}{\sigma^{-2}\|Q(M_f^m)\mu_n\|_2^2} \right)^{k/4} \right) \sum_M \frac{(q_0)^{k/2}}{|M|^{k/4}}, \quad (16)$$

where $d(4k)$ is a generic constant depending on $E(\epsilon/\sigma)^{4k}$ and

$$M_f^m = \arg \min_{M \in \mathcal{M}^F} \|Q(M)\mu_n\|_2^2.$$

As a consequence of Lemma 2, the choice of τ depends on λ , q_0 and $\|Q(M_f)\mu_n\|_2^2$. In more detail, underfitting is unlikely to happen as for large $\|Q(M_f)\mu_n\|_2^2$, small λ and q_0 , so a large perturbation size τ is allowed.

Next we choose τ to strike a balance to give an optimal τ -value by minimizing (16) with respect to τ . This leads to an optimal value of τ , τ_λ^o satisfying

$$1 + (\tau_\lambda^o)^2 = \frac{16\|Q(M_f^m)\mu_n\|_2^2}{\sigma(\lambda-1)\sqrt{q_0+1}}. \quad (17)$$

As displayed in (17), τ_λ^o decreases in (λ, σ^2, q_0) and increases in the approximation error of wrong models. Evidently, τ_λ^o depends on unknown parameters q_0 and μ_n . According to the assumption we made in Section 2.1, namely, $\|Q(M_f^m)\mu_n\|_2^2/(\sigma q_0 \log p) \rightarrow \infty$, we suggest a more practical perturbation size

$$\tau_\lambda = \frac{4(\log p)^{1/4}}{(\lambda-1)^{1/2}}. \quad (18)$$

As a consequence of Lemma 2, the following result is established if the following condition holds.

Condition C. for some $k \geq 4$

$$\sum_{M \in \mathcal{M}^F} \frac{n(q_0+1)^{2k} (\log p)^{k/4}}{(\sigma^{-2} \|Q(M_f) \mu_n\|_2^2 |M_f|)^{1/2}} \rightarrow 0, \quad \sum_{M \in \mathcal{M}^T} \frac{n(q_0+1)^{k/8}}{(|M_t| \log p)^{k/8}} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (19)$$

LEMMA 3 Assume Condition C holds. Then $\hat{U}_{\mathcal{M}}^{\tau_\lambda}(\lambda) \rightarrow U_{\mathcal{M}}(\lambda)$ in probability as $n^{-2}m$ and $n \rightarrow \infty$.

After the λ -dependent τ_λ is computed in (18), the perturbed data $\{Y_{n,j}^{\tau_\lambda}; j=1, \dots, m\}$ with τ replaced by τ_λ in $Y_{n,j}^\tau$ is employed to assess model uncertainty in model selection, that is, to compute estimated models based on $\{Y_{n,j}^{\tau_\lambda}; j=1, \dots, m\}; j=1, \dots, m\}$:

$$\hat{M}_{\lambda,j}^{\tau_\lambda} = \argmin_{M \in \mathcal{M}} \left(\|Q(M) Y_{n,j}^{\tau_\lambda}\|_2^2 + \lambda |M| (1 + \tau_\lambda^2) \sigma^2 \right), \quad (20)$$

as in (11). This results in our estimate of $U_{\mathcal{M}}(\lambda)$

$$\hat{U}_{\mathcal{M}}^{\tau_\lambda}(\lambda) = \frac{\sum_{j=1}^m \|P(\hat{M}_{\lambda,j}^{\tau_\lambda}) Y_{n,j}^{\tau_\lambda} - P(\hat{M}_{\lambda,j}^{\tau_\lambda}) Y_n\|_2^2}{m \tau_\lambda^2 \sigma^2}, \quad (21)$$

which consistently estimates $U_{\mathcal{M}^T}(\lambda)$ as suggested by Lemma 3.

3 Adaptive model selection through data-driven regularization

This section investigates modeling procedure or tuning parameter selection by the data perturbation method. Based on an approximately unbiased estimate of (4), we propose our adaptive model selection criterion, denoted by ADPC,

$$ADPC_{\mathcal{M}}(\lambda) = \frac{1}{m} \sum_{j=1}^m \|Q(\hat{M}_{\lambda,j}^{\tau_\lambda}) Y_n\|_2^2 + 2\sigma^2 \hat{U}_{\mathcal{M}}^{\tau_\lambda}(\lambda), \quad (22)$$

which is a function of λ instead of others. In (22), $\hat{\sigma}^2$ replaces σ^2 in the case of unknown σ^2 and more details are provided in Section 4. If $\hat{M}_{\lambda,j}^{\tau_\lambda}$ were (approximately) independent of ϵ_n^* and $\hat{U}_{\mathcal{M}}^{\tau_\lambda}$ is a consistent estimate of $U_{\mathcal{M}}(\lambda)$, then $E(ADPC_{\mathcal{M}}(\lambda)) = R_n(\hat{M}_\lambda) + n\sigma^2$ holds approximately.

As a result, model selection by ADPC proceeds in two steps.

1. Obtain $\hat{\lambda}$ by minimizing ADPC (22) over a set of tuning parameters, Λ .
2. Plug $\hat{\lambda}$ into (2) and then obtain $\hat{M}_{\hat{\lambda}}$ by minimizing (2).

This leads to higher accuracy of model selection as showed in the following theorem.

THEOREM 1 Let $\lambda_0 = \argmin_{\lambda \in \Lambda} R_n(\hat{M}_\lambda)$ and suppose $\alpha(\lambda) = R_n(\hat{M}_\lambda) - R_n(\hat{M}_{\lambda_0}) \geq \delta > 0$ for $\lambda \neq \lambda_0$, $|\Lambda| < \infty$ and $IC_\lambda(M) - IC_\lambda(M_0) \leq E(\min_M IC_\lambda(M) - IC_\lambda(M_0))$ almost surely for $\lambda \neq \lambda_0$. Let $\hat{\lambda} = \argmin_{\lambda \in \Lambda} ADPC_{\mathcal{M}}(\lambda)$. Then tuning parameter selection consistency is established for $ADPC_{\mathcal{M}}(\lambda) : P(\hat{\lambda} = \lambda_0) \rightarrow 1$ provided that $n^2 m^{-1} \rightarrow 0$ and Condition C holds as $n \rightarrow \infty$.

According to Theorems 1, ADPC approximates the optimal performance of a class of model selection routines defined by (2), indexed by $\lambda \in \Lambda$. This implies that ADPC is expected to out-perform any nonadaptive model selection (2), including AIC, BIC, RIC, among others. It is worth pointing out that Condition A is not required for Theorem 1 because an overly large λ -value is not selected by ADPC (22).

4 Numerical results

4.1 Simulations

This subsection studies operating characteristics of ADPC in classical linear regression ($n = 400$ and $p = 20$) and high-dimensional situation ($n = 200$ and $p = 1000$) with a sparse true model through simulations, respectively, and compares it against strong competitors regarding selective and predictive accuracy.

For accuracy of selection, we use false and negative positives, which are the average number of irrelevant predictors falsely selected as true predictors (FP) and that of true predictors falsely excluded as irrelevant predictors (FN). In addition, the proportion of correctly, underfitted or overfitted models are reported based on 100 simulation replications. For accuracy of prediction, we employ the usual metric the mean squares error (MSE) defined as

$$\|\hat{\mu}_n - \mu_n\|_2^2$$

Simulations are performed in (1). Simulated data are generated as follows. Each row of the $n \times p$ design matrix X is independently generated from $N(0, \Sigma_{p \times p})$, where 0 denotes the p -dimensional vector of 0, and the ij th entry of covariance matrix $\Sigma_{p \times p}$ is 1 if $i = j$ and $\rho^{|i-j|}$ if otherwise. Two values of $\rho = \pm 0.5$ are examined in addition to different sizes of the true model and different signal-to-noise ratios (SNR). The SNR ratio is defined as $\sigma^{-1} n^{-1/2} \|\mu_n\|_2$ and two values of SNR, 2 and 0.5 are examined. The true model is obtained by $\mu_n = X\eta$, where η is a p -dimension vector of regression coefficients, and only the first $q_0 = |M_0| = 1, 3, 5$ components of η , $\eta_1 = \dots = \eta_{q_0}$, are nonzeros and the rest are zeros. Moreover, ϵ in (1) follows $N(0, 1)$.

4.1.1 Standard regression—In a low-dimensional situation, we will demonstrate that ADPC is capable of picking up the best model selection criterion from a class of information criteria in the form of (2). We assume that σ^2 is unknown and it is estimated by the full model.

Consider a case of $p = 20$ for AIC, BIC, and BIC_c , where exhaustive search is performed. In simulations, ADPC is employed to select among three candidate criteria: AIC ($\lambda = 2$), BIC ($\lambda = \log n$) and BIC_c ($\lambda = \sqrt{n}$), followed by model selection using the criterion chosen by ADPC.

As indicated by Table 1, ADPC identifies the best performer BIC_c among the three criteria, which is evidenced by its identical performance to BIC_c , when the signal-to-noise ratio is high or the true model size is small. Interestingly, BIC does not seem to achieve selection consistency in this situation and tends to select an overfitting model due to a not-large-enough penalty. On the other hand, as the size of the true model gets large while the signal gets weak, BIC's performance gets better and ultimately dominates others in the case of $q_0 = 3, 5$ and $SNR = 0.5$ while BIC_c tends to underfit due to an overly large penalty and AIC tends to overfit due to an overly small penalty. Nevertheless, in these situations, ADPC is identical to BIC. In other words, ADPC approximates the best performance among AIC, BIC, and BIC_c adaptively, regardless of the size of the true model, signal-to-noise ratio and the design matrix structure (some results ($\rho = 0.5$) are not presented here but have the similar pattern). In summary, the simulation confirms Theorem 1 and demonstrates that ADPC enables to select the best one from a class of candidate model selection criteria or tuning parameters.

4.1.2 High-dimensional regression—In the high-dimensional setting we let $n = 200$ and $p = 1000$. The competitors include LASSO [23] with CV, MCP [29] with CV, SCAD [8] with CV, and TLP with CV [20], denoted as LASSO-CV, MCP-CV, SCAD-CV, and TLP-CV, where the first three are calculated through the R-package *ncvreg* in CRAN (<http://cran.r-project.org>); see [2]. Here a 10-fold CV, the default set-up, given in the R-routines, is used to select tuning parameters over default 100 grid points for these methods. We still assume that σ^2 is unknown and use stepwise regression to select a model and estimate σ^2 by fitting this model using the least squares method.

The estimated tuning parameter $\hat{\lambda}$ for ADPC and TLP+CV is chosen from 100 equally spaced λ points over $[2, \max |X'Y|]$, where our limited experience indicates that 100 points suffice to deliver satisfactory performance.

With respect to selective and predictive accuracy, ADPC outperforms the other four competitors across all the situations, as indicated by Table 2. As a matter of fact, the MSE improvements over TLP-CV, MCP-CV, SCAD-CV, and LASSO-CV range from 41% to 65%, 20% to 55%, 20% to 47%, and 86% to 91% for $\rho = -0.5$ and $SNR = 2$. Interestingly, the improvement appears larger for $\rho = 0.5$ (not shown here) than for $\rho = -0.5$. Moreover, ADPC identifies the true model with the proportion above 88% while all other methods select the true model with the proportion below 70%.

As the signal-to-noise ratio is small ($SNR = 0.5$), or equivalently, the magnitude of nonzero coefficient decreases, the performance of all five methods deteriorates. This is expected in view of the necessary condition for feature selection consistency demonstrated in [20] that indicates that underfitting tends to happen as nonzero coefficients are below some threshold. At the same time, all methods tend to select some irrelevant variables (overfitting) and miss

some predictors in the true model (underfitting) in all situations due to difficulty to distinguish zero coefficients with close-to-zero coefficients.

With respect to predictive stability, as suggested by Table 2, ADPC has the smallest standard error of MSE in every situation. This says that the proposed method has the smallest amount of model selection uncertainty. As suggested by [6], CV tends to have high variability for model selection. Moreover, a comparison of TLP-CV against ADPC suggests that our risk estimation has higher accuracy than CV does.

In conclusion, ADPC yields the smallest MSE with the largest proportion of correct selection across all the situations. This simulation result is in accordance with the theoretical result in Theorem 1.

4.2 A real data example

In this subsection the method ADPC is used to construct a regression model of forecasting nodal hourly power price in the state of New York and its performance is compared with the other four competitors LASSO-CV, MCP-CV, SCAD-CV, and TLP-CV. The electricity is priced locally at a specific location such as a big generator or consumer, which is called a *node*, in a regional wholesale power market. If two nodes are geographically close to each other, they tend to have similar prices. The dataset represents 150 hourly prices on each of the 423 *nodes* of the state of New York, one of which works as the response. For our data $p = 422$ is greater than $n = 150$, but only a small fraction of nodes are expected useful in forecasting the price. Therefore it is highly desirable to make use of a high dimensional data learning approach to develop a predictive model.

The cross-validation method is used to compare predictive performance of all five procedures. The 150 observations are divided into a training set including the first 75 observations and a validation set including the remaining 75 observations in a timely order. Let Y_i and $\hat{Y}_i^p$; $i = 1, \dots, 75$ denote the observed and predicted prices for the observations in the validation set, respectively. The prediction error (PE) is defined as $|Y_i - \hat{Y}_i^p|$; $i = 1, \dots, 75$. The five procedures are employed to the training set developing five models, which are used to make predictions on the validation set. The error variance σ^2 is estimated by the least squared based on a model selected by stepwise regression over the whole dataset.

The mean absolute value and standard deviation of PE's of all five procedures are displayed in Table 3.

As indicated in Table 3, the model selected by ADPC yields the best prediction accuracy among all five competitors. The out-performance of ADPC over LASSO-CV, MCP-CV, SCAD-CV, and TLP-CV confirms the advantage of ADPC in high-dimensional data mining.

5 Discussion

It is of practical importance to evaluate and account for modeling uncertainty involving model selection, especially in the age of big data. Big data are generated every day at financial institutions. As a result, there is an increasing need for mining out pattern from

large-scale data, hence statistical analysis is critical for high frequency transaction data, particularly in behavior modeling, telecommunications, and cyber-security. Unlike in classical statistics, modeling uncertainty resulting from model exploration has become a dominant part and needs to be taken into account in statistical modeling.

This article presents a method of measuring and estimating modeling uncertainty based on data perturbation, which has desirable statistical properties. On this basis, we derive an adaptive model selection routine that achieves adaptivity by selecting the optimal procedure or tuning parameter from a pool of candidates. Simulation results demonstrate that the proposed method compares favorably against its competitors in terms of the accuracy of selection and prediction. Further investigation is necessary to generalize the proposed method to other nonlinear models.

Appendix: Proofs

Proof of some probability bounds and conclusions in Example 1

The upper bound of $P(\hat{M}_\lambda^\tau = M_t)$ and $P(\hat{M}_\lambda^\tau = M_f)$: Supposing that $E\epsilon_j^{4k} < \infty$ ($k \geq 4$) ($k \geq 4$), an application of Whittle's inequality [25] yields that

$$E\left(\|P(M)\epsilon_n\|_2^2 - |M|\right)^{2k} \leq 2^{2k} c(4k) |M|^k \sigma^{4k}; \quad E\left(\epsilon_n' Q(M) \mu_n\right)^{4k} \leq c(4k) \|Q(M) \mu_n\|_2^{4k} \sigma^{4k} \quad (23)$$

where $c(4k)$ is a constant depending on $E(\epsilon_j/\sigma)^{4k}$.

The probability of selecting a wrong model M_f by (10) is bounded by

$$\begin{aligned} P(\hat{M}_\lambda^\tau = M_f) &\leq P(IC_\lambda^\tau(M_f) \leq IC_\lambda^\tau(M_0)) \leq P\left(\|P(M_f)(\epsilon_n + \tau\epsilon_n^*)\|_2^2 - \|P(M_0)(\epsilon_n + \tau\epsilon_n^*)\|_2^2 - 2(\epsilon_n + \tau\epsilon_n^*)' Q(M_f) \mu_n \geq \|Q(M_f) \mu_n\|_2^2\right. \\ &\quad \left.+ \lambda |M_f| (1 + \tau^2) \sigma^2 - \lambda q_0 (1 + \tau^2) \sigma^2\right) \leq p_1 + p_2 \end{aligned}$$

, where $p_1 = P\left(\|P(M_f)(\epsilon_n + \tau\epsilon_n^*)\|_2^2 - |M_f| (1 + \tau^2) \sigma^2 \geq \alpha/2\right)$,

$p_2 = P\left(2(\epsilon_n + \tau\epsilon_n^*)' Q(M_f) \mu_n \geq \alpha/2\right)$ and

$\alpha = \|Q(M_f) \mu_n\|_2^2 + (\lambda - 1) |M_f| (1 + \tau^2) \sigma^2 - \lambda q_0 (1 + \tau^2) \sigma^2$.

An application of (23) yields that

$$p_1 \leq \frac{2^{4k} c(4k) |M_f|^k ((1 + \tau^2) \sigma^2)^{2k}}{\alpha^{2k}}; \quad p_2 \leq \frac{2^{4k} c(4k) \|Q(M_f) \mu_n\|_2^{2k} ((1 + \tau^2) \sigma^2)^k}{\alpha^{2k}}.$$

As $\|Q(M_f) \mu_n\|_2^2 \geq 2(\lambda - 1)(q_0 + 1)(1 + \tau^2) \sigma^2$,

$$\begin{aligned}
p_1 &\leq \frac{2^{4k} c(4k) |M_f|^k ((1+\tau^2)\sigma^2)^{2k}}{\alpha^{2k}} \leq \frac{2^{4k} c(4k) |M_f|^k ((\lambda-1)(q_o+1)(1+\tau^2)\sigma^2)^{2k}}{\left(\|Q(M_f)\mu_n\|_2^2 + (\lambda-1)|M_f|(1+\tau^2)\sigma^2\right)^{2k}} \\
&\leq \frac{2^{4k} c(4k) ((\lambda-1)(1+\tau^2)\sigma^2)^k (q_o+1)^{2k}}{\left(\|Q(M_f)\mu_n\|_2^2 + (\lambda-1)|M_f|(1+\tau^2)\sigma^2\right)^k} \leq \frac{2^{3k} c(4k) (q_o+1)^{2k} ((\lambda-1)(1+\tau^2)\sigma^2)^{k/2}}{\left(\|Q(M_f)\mu_n\|_2^2 |M_f|\right)^{k/2}}; \\
p_2 &\leq \frac{2^{4k} c(4k) \|Q(M_f)\mu_n\|_2^{2k} ((1+\gamma^2)\sigma^2)^k}{\alpha^{2k}} \leq \frac{2^{4k} c(4k) (q_o+1)^{2k} ((1+\gamma^2)\sigma^2)^k}{\left(\|Q(M_f)\mu_n\|_2^2 + (\lambda-1)|M_f|(1+\tau^2)\sigma^2\right)^k} \\
&\leq \frac{2^{4k} c(4k) ((\lambda-1)(1+\tau^2)\sigma^2)^k (q_o+1)^{2k}}{\left(\|Q(M_f)\mu_n\|_2^2 + (\lambda-1)|M_f|(1+\tau^2)\sigma^2\right)^k} \leq \frac{2^{3k} c(4k) q_o+1)^{2k} ((\lambda-1)(1+\tau^2)\sigma^2)^{k/1}}{\left(\|Q(M_f)\mu_n\|_2^2 |M_f|\right)^{k/2}}
\end{aligned}$$

Then, the probability of selecting an underfitting model M_f by (10) is bounded above by

$$P\left(\hat{M}_{\lambda}^{\tau} = M_f\right) \leq \frac{2^{3k} c(4k) (q_o+1)^{2k} ((\lambda-1)(1+\tau^2)\sigma^2)^{k/2}}{\left(\|Q(M_f)\mu_n\|_2^2 |M_f|\right)^{k/2}}. \quad (24)$$

Similarly, supposing $M_{\lambda,f}^m = \argmin_{M \in \mathcal{M}} \|Q(M)\mu_n\|_2^2 + (\lambda-1)|M|\sigma^2$ and $M_{\lambda,f}^m \in \mathcal{M}^F$, we have the following inequality

$$P\left(\hat{M}_{\lambda}^{\tau} = M_F, M_f \neq M_{\lambda,f}^m\right) \leq \frac{2^{3k} c(4k) (q_o+1)^{2k} ((\lambda-1)(1+\tau^2)\sigma^2)^{k/2}}{\left(\|Q(M_f)\mu_n\|_2^2 |M_f|\right)^{k/2}}. \quad (25)$$

Proof of Conclusions in Example 1: Since M_1 is nested within M_2 , we get

$$\begin{aligned}
R_n\left(\hat{M}_{\lambda}\right) &= E\|P(M_1)\epsilon_n\|_2^2 \mathbf{I}_{\{IC_{\lambda}(M_2) \geq IC_{\lambda}(M_1)\}} + E\left(\|P(M_2)\epsilon_n\|_2^2 \mathbf{I}_{\{IC_{\lambda}(M_2) \geq IC_{\lambda}(M_1)\}}\right) \\
&= E\|P(M_1)\epsilon_n\|_2^2 \mathbf{I}_{\{\|P(M_2)\epsilon_n\|_2^2 - \|P(M_1)\epsilon_n\|_2^2 < \lambda\sigma^2\}} + E\|P(M_2)\epsilon_n\|_2^2 \mathbf{I}_{\{\|P(M_2)\epsilon_n\|_2^2 - \|P(M_1)\epsilon_n\|_2^2 > \lambda\sigma^2\}} \\
&= E\|P(M_1)\epsilon_n\|_2^2 + E\left(\|P(M_2)\epsilon_n\|_2^2 - \|P(M_1)\epsilon_n\|_2^2\right) \mathbf{I}_{\{\|P(M_2)\epsilon_n\|_2^2 - \|P(M_1)\epsilon_n\|_2^2 > \lambda\sigma^2\}} \\
&= 2\sigma^2 + \sigma^2 \int_{\lambda}^{\infty} \frac{x}{2^{1/2}\Gamma(1/2)} x^{\frac{1}{2}-1} \exp\left(-\frac{x}{2}\right) dx = \sigma^2 (2 + P(\chi_3^2 > \lambda)).
\end{aligned}$$

Proof of Lemmas 1-3

Proof of Lemma 1: First, it follows from the definition of $\hat{M}_{\lambda}$ that as $\lambda_1 < \lambda_2$

$$\begin{aligned}
\|Q(\hat{M}_{\lambda_1})Y_n\|_2^2 + \lambda_1 |\hat{M}_{\lambda_1}| \sigma^2 &\leq \|Q(\hat{M}_{\lambda_2})Y_n\|_2^2 + \lambda_1 |\hat{M}_{\lambda_2}| \sigma^2, \\
\|Q(\hat{M}_{\lambda_2})Y_n\|_2^2 + \lambda_2 |\hat{M}_{\lambda_2}| \sigma^2 &\leq \|Q(\hat{M}_{\lambda_1})Y_n\|_2^2 + \lambda_2 |\hat{M}_{\lambda_1}| \sigma^2.
\end{aligned}$$

Hence $(\lambda_2 - \lambda_1) (|\hat{M}_{\lambda_2}| - |\hat{M}_{\lambda_1}|) \leq 0$ and

$$\|Q(\hat{M}_{\lambda_1})Y_n\|_2^2 - \|Q(\hat{M}_{\lambda_2})Y_n\|_2^2 \leq \lambda_1 (|\hat{M}_{\lambda_2}| - |\hat{M}_{\lambda_1}|) \sigma^2.$$

As a consequence,

$$\|Q(\hat{M}_{\lambda_1})Y_n\|_2^2 - \|Q(\hat{M}_{\lambda_2})Y_n\|_2^2 \leq 0 \text{ for } \lambda_2 \lambda_1. \quad (26)$$

Therefore, if $\mathcal{M}^F$ is empty then $\|P(\hat{M}_{\lambda_2})\epsilon_n\|_2^2 \leq \|P(\hat{M}_{\lambda_1})\epsilon_n\|_2^2$ and

$U_{\mathcal{M}^T}(\lambda_2) \leq U_{\mathcal{M}^T}(\lambda_1)$. Now consider the case that $\mathcal{M}^F$ is nonempty, in which we need to prove in (7)

$$E\left(\left(\|Q(\hat{M}_{\lambda,f})\mu_n\|_2^2 + \|P(\hat{M}_{\lambda,f})\epsilon_n\|_2^2 - \|P(\hat{M}_{\lambda,t})\epsilon_n\|_2^2\right)I_{\{\hat{M}_{\lambda} \in \mathcal{M}^F\}}\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} E\left(\|Q(\hat{M}_{\lambda,f})\mu_n\|_2^2 I_{\{\hat{M}_{\lambda} \in \mathcal{M}^F\}}\right) &\leq \|\mu_n\|_2^2 P(\hat{M}_{\lambda} \in \mathcal{M}^F), \\ E\left(\|P(\hat{M}_{\lambda,f})\epsilon_n\|_2^2 I_{\{\hat{M}_{\lambda} \in \mathcal{M}^F\}}\right) &\leq \sqrt{E\left(\|P(\hat{M}_{\lambda,f})\epsilon_n\|_2^2\right)^2 P(\hat{M}_{\lambda} \in \mathcal{M}^F)}, \\ E\left(\|P(\hat{M}_{\lambda,t})\epsilon_n\|_2^2 I_{\{\hat{M}_{\lambda} \in \mathcal{M}^F\}}\right) &\leq \sqrt{E\left(\|P(\hat{M}_{\lambda,t})\epsilon_n\|_2^4\right)^2 P(\hat{M}_{\lambda} \in \mathcal{M}^F)}, \end{aligned}$$

all of which converge to 0 as $n \rightarrow \infty$ by Condition A. This completes the proof.

Proof of Lemma 2: Let

$$\Delta IC_{\lambda}^{\diamond}(M, \hat{M}_{\lambda,t}^{\diamond}) = \left(-\|P(M)\epsilon_n^*\|_2^2 + \lambda|M|\sigma^2\right) - \left(-\|P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n^*\|_2^2 + \lambda|\hat{M}_{\lambda,t}^{\diamond}|\sigma^2\right) \text{ and}$$

obviously $\Delta IC_{\lambda}^{\diamond}(M, \hat{M}_{\lambda,t}^{\diamond}) > 0$ for $\hat{M}_{\lambda,t}^{\diamond} \neq M$. Consider

$$P(\hat{M}_{\lambda,t}^{\diamond} \neq \hat{M}_{\lambda,t}^{\tau}) = E\left(P(\hat{M}_{\lambda,t}^{\diamond} \neq \hat{M}_{\lambda,t}^{\tau} | \epsilon_n^*)\right) \text{ and we establish}$$

$$\begin{aligned} &P(\hat{M}_{\lambda,t}^{\diamond} \neq \hat{M}_{\lambda,t}^{\tau} | \epsilon_n^*) \\ &\leq P\left(\frac{-1}{1+\tau^2} \|P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n + \tau\epsilon_n^*\|_2^2 + \lambda|\hat{M}_{\lambda,t}^{\diamond}|\sigma^2 > \frac{-1}{1+\tau^2} \|P(\hat{M}_{\lambda,t}^{\tau})\epsilon_n + \tau\epsilon_n^*\|_2^2 + \lambda|\hat{M}_{\lambda,t}^{\tau}|\sigma^2; \hat{M}_{\lambda,t}^{\tau} \neq \hat{M}_{\lambda,t}^{\diamond} | \epsilon_n^*\right) \\ &\leq \sum_{M \in \mathcal{M}^T} P\left(\frac{-1}{1+\tau^2} \|P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n + \tau\epsilon_n^*\|_2^2 + \lambda|\hat{M}_{\lambda,t}^{\diamond}|\sigma^2 > \frac{-1}{1+\tau^2} \|P(M)\epsilon_n + \tau\epsilon_n^*\|_2^2 + \lambda|M|\sigma^2; \right. \\ &\quad \left. \hat{M}_{\lambda,t}^{\diamond} \neq M; \hat{M}_{\lambda,t}^{\tau} = M | \epsilon_n^*\right) \\ &\leq \sum P\left(-2\tau\epsilon_n' P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n^* + 2\tau\epsilon_n' P(M)\epsilon_n^* - \|P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n\|_2^2 + \|P(M)\epsilon_n\|_2^2 \right. \\ &\quad \left. + \|P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n^*\|_2^2 - \|P(M)\epsilon_n^*\|_2^2 > (1+\tau^2) \Delta IC_{\lambda}^{\diamond}(M, \hat{M}_{\lambda,t}^{\diamond}); \hat{M}_{\lambda,t}^{\tau} = M | \epsilon_n^*\right). \end{aligned}$$

Then by Markov Inequality,

$$P\left(\hat{M}_{\lambda,t}^{\diamond} \neq \hat{M}_{\lambda,t}^{\tau} | \epsilon_n^*\right) \leq \sum_{M \in \mathcal{M}^T} \frac{E\left(-2\tau\epsilon_n' P\left(\hat{M}_{\lambda,t}^{\diamond}\right) \epsilon_n^* + 2\tau\epsilon_n' P(M) \epsilon_n^* + \|P(M/M_0) \epsilon_n\|_2^2 + \left\|P\left(\hat{M}_{\lambda,t}^{\diamond}/M_0\right) \epsilon_n^*\right\|_2^2\right)^{2k}}{(1+\gamma^2)^{2k} \left(\Delta IC_{\lambda}^{\diamond}\left(M, \hat{M}_{\lambda,t}^{\diamond}\right)\right)^{2k}}$$

As τ is large enough such that

$$\begin{aligned} (1+\tau^2)^{k/2} &\geq E\left(\Delta IC_{\lambda}^{\diamond}\left(M, \hat{M}_{\lambda,t}^{\diamond}\right)\right)^{2k} E\left(1/\Delta IC_{\lambda}^{\diamond}\left(M, \hat{M}_{\lambda,t}^{\diamond}\right)\right)^{2k} \text{ and} \\ (1+\tau^2)^{k/2} &\geq \|P\left(\hat{M}_{\lambda,t}^{\diamond}/M_0\right) \epsilon_n^*\|_2^{2k}, \text{ then} \end{aligned}$$

$$\begin{aligned} &P\left(\hat{M}_{\lambda,t}^{\diamond} \neq \hat{M}_{\lambda,t}^{\tau} | \epsilon_n^*\right) \\ &\leq \sum_{M \in \mathcal{M}^T} \frac{2^{4k} c(4k) \left(\tau^{2k} \|P(\hat{M}_{\lambda,t}^{\diamond}/M_0) \epsilon_n^*\|_2^{2k} + \tau^{2k} \|P(M/M_0) \epsilon_n^*\|_2^{2k} + \|P(\hat{M}_{\lambda,t}^{\diamond}/M_0) \epsilon_n^*\|_2^{4k} + E\|P(M/M_0) \epsilon_n^*\|_2^{4k}\right)}{(1+\tau^2)^{2k} \left(\Delta IC_{\lambda}^{\diamond}\left(M, \hat{M}_{\lambda,t}^{\diamond}\right)\right)^{2k}} \\ &\leq \sum_{M \in \mathcal{M}^T} \frac{2^{4k} c(4k) \left(\|P(\hat{M}_{\lambda,t}^{\diamond}/M_0) \epsilon_n^*\|_2^{2k} + \|P(M/M_0) \epsilon_n^*\|_2^{2k} + \|P(\hat{M}_{\lambda,t}^{\diamond}/M_0) \epsilon_n^*\|_2^{2k} + E\|P(M/M_0) \epsilon_n^*\|_2^{2k}\right)}{(1+\tau^2)^k \left(\Delta IC_{\lambda}^{\diamond}\left(M, \hat{M}_{\lambda,t}^{\diamond}\right)\right)^{2k}} \end{aligned}$$

In conclusion,

$$P\left(\hat{M}_{\lambda,t}^{\diamond} \neq \hat{M}_{\lambda,t}^{\tau}\right) \leq \sum_{M \in \mathcal{M}^T} \frac{2^{5k} c(4k) (q_0+1)^k}{((1+\tau^2)(\lambda-1)|M|)^{k/2}} \quad (27)$$

For Err_t in (13),

$$|\text{Err}_t| \leq \sqrt{E\|P\left(\hat{M}_{\lambda,t}^{\tau}\right) \epsilon_n^*\|_4^2} \sqrt{P\left(\hat{M}_{\lambda,t}^{\tau} \neq \hat{M}_{\lambda,t}^{\diamond}\right)} + \sqrt{E\|P\left(\hat{M}_{\lambda,t}^{\diamond}\right) \epsilon_n^*\|_2^4} \sqrt{P\left(\hat{M}_{\lambda,t}^{\tau} \neq \hat{M}_{\lambda,t}^{\diamond}\right)}. \quad (28)$$

For Err_f , $|\text{Err}_f| \leq E\|P\left(\hat{M}_{\lambda}^{\tau}\right) \epsilon_n^*\|_2^2 I_{\{\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^F\}} + E\|P\left(\hat{M}_{\lambda,t}^{\tau}\right) \epsilon_n^*\|_2^2 I_{\{\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^F\}}$ which indicates

$$|\text{Err}_f| \leq \sqrt{E\|P\left(\hat{M}_{\lambda}^{\tau}\right) \epsilon_n^*\|_2^4} \sqrt{P\left(\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^F\right)} + \sqrt{E\|P\left(\hat{M}_{\lambda,t}^{\tau}\right) \epsilon_n^*\|_2^4} \sqrt{P\left(\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^F\right)}. \quad (29)$$

A combination of (28) and (29) leads to

$$|\text{Err}| \leq 2\sqrt{E\|\epsilon_n^*/\sigma\|_2^4} \left(\sqrt{P\left(\hat{M}_{\lambda,t}^{\tau} \neq \hat{M}_{\lambda,t}^{\diamond}\right)} + \sqrt{P\left(\hat{M}_{\lambda}^{\tau} \in \mathcal{M}^F\right)}\right), \quad (30)$$

which in turn yields that

$$|Err| \leq d(4k)n \left(\left(\frac{16}{(1+\tau^2)(\lambda-1)} \right)^{k/4} + \left(\frac{(1+\gamma^2)(\lambda-1)(q_0+1)^2}{(\sigma^{-2}\|Q(M_f^m)\mu_n\|_2^2)} \right)^{k/4} \right) \sum_M \frac{(q_0+1)^{k/2}}{|M|^{k/4}}.$$

This completes the proof.

Proof of Lemma 3: Note that

$var(\hat{U}_{\mathcal{M}}^{\tau}(\lambda)) = var(E(\hat{U}_{\mathcal{M}}^{\tau}(\lambda)|Y_n)) + E(var(\hat{U}_{\mathcal{M}}^{\tau}(\lambda)|Y_n))$. Then

$$var(\hat{U}_{\mathcal{M}}^{\tau}(\lambda)) = \frac{var(E(\|P(\hat{M}_{\lambda}^{\tau})\epsilon_n^*\|_2^2|Y_n))}{\sigma^4\tau^4} + \frac{E(var(\|P(\hat{M}_{\lambda}^{\tau})\epsilon_n^*\|_2^2|Y_n))}{m\sigma^4\tau^4}, \quad (31)$$

where the second term goes to 0 as $m \rightarrow \infty$.

Moreover,

$$var(E(\|P(\hat{M}_{\lambda}^{\tau})\epsilon_n^*\|_2^2|Y_n)) = var(E(\|P(\hat{M}_{\lambda}^{\tau})\epsilon_n^*\|_2^2 - \|P(\hat{M}_{\lambda}^{\diamond})\epsilon_n^*\|_2^2|Y_n)). \quad (32)$$

Thus, we need to prove $E(\|P(\hat{M}_{\lambda,t}^{\tau})\epsilon_n^*\|_2^2 - \|P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n^*\|_2^2|Y_n) \rightarrow 0$ in probability as $\tau \rightarrow \infty$. Moreover,

$$E(\|P(\hat{M}_{\lambda,t}^{\tau})\epsilon_n^*\|_2^2 - \|P(\hat{M}_{\lambda,t}^{\diamond})\epsilon_n^*\|_2^2|Y_n) \leq 2\sqrt{E\|Q(M_0)\epsilon_n^*\|_2^4}\sqrt{P(\hat{M}_{\lambda,t}^{\tau} \neq \hat{M}_{\lambda,t}^{\diamond}|Y_n)}$$

. Consequently, $var(\hat{U}_{\mathcal{M}}^{\tau}(\lambda)) \rightarrow 0$ as Condition C holds. This completes the proof.

Proof of Theorem 1

First, we derive the upper bound of $P(ADPC_{\mathcal{M}}(\lambda_0) \geq ADPC_{\mathcal{M}}(\lambda))$ for a given $\lambda \geq \lambda_0$

with $R(\hat{M}_{\lambda}) - R(\hat{M}_{\lambda_0}) = \alpha(\lambda)$. Since

$$E(ADPC_{\mathcal{M}}(\lambda_0)|Y_n) = E\left(\|Q(\hat{M}_{\lambda_0}^{\tau\lambda_0})\mu_n\|_2^2 + 2\mu_n'Q(\hat{M}_{\lambda_0}^{\tau\lambda_0})\epsilon_n + \|Q(\hat{M}_{\lambda_0}^{\tau\lambda_0})\epsilon_n\|_2^2|Y_n\right) + 2\delta^2E(\hat{U}_{\mathcal{M}}^{\tau\lambda_0}(\lambda_0)|Y_n);$$

$$E(ADPC_{\mathcal{M}}(\lambda)|Y_n) = E\left(\|Q(\hat{M}_{\lambda}^{\tau\lambda})\mu_n\|_2^2 + 2\mu_n'Q(\hat{M}_{\lambda}^{\tau\lambda})\epsilon_n + \|Q(\hat{M}_{\lambda}^{\tau\lambda})\epsilon_n\|_2^2|Y_n\right) + 2\delta^2E(\hat{U}_{\mathcal{M}}^{\tau\lambda}(\lambda)|Y_n),$$

$$P(ADPC_{\mathcal{M}}(\lambda_0) \geq ADPC_{\mathcal{M}}(\lambda)) \leq p_{11} + p_{12}, \quad (33)$$

$$p_{11} = P\left(E(ADPC_{\mathcal{M}}(\lambda_0) - ADPC_{\mathcal{M}}(\lambda)|Y_n) + \alpha(\lambda) \geq \frac{\alpha(\lambda)}{2}\right), \quad (34)$$

$$p_{12} = P \left(ADPC_{\mathcal{M}}(\lambda_0) - E(ADPC_{\mathcal{M}}(\lambda_0) | Y_n) - ADPC_{\mathcal{M}}(\lambda) + E(ADPC_{\mathcal{M}}(\lambda) | Y_n) \geq \frac{\alpha(\lambda)}{2} \right). \quad (35)$$

Towards this end, we decompose p_{11} and p_{12} as

$$\begin{aligned} p_{11} &\leq P \left(E(ADPC(\lambda_0) | Y_n) - \|\epsilon_n\|_2^2 - R(\hat{M}_{\lambda_0}) \geq \frac{\alpha(\lambda)}{4} \right) + P \left(R(\hat{M}_{\lambda}) - E(ADPC(\lambda) | Y_n) + \|\epsilon_n\|_2^2 \geq \frac{\alpha(\lambda)}{4} \right) \\ &\equiv p_{11,\lambda_0} + p_{11,\lambda}; \\ p_{12} &\leq P \left(ADPC_{\mathcal{M}}(\lambda_0) - E(ADPC_{\mathcal{M}}(\lambda_0) | Y_n) \geq \frac{\alpha(\lambda)}{4} \right) + P \left(E(ADPC_{\mathcal{M}}(\lambda) | Y_n) - ADPC_{\mathcal{M}}(\lambda) \geq \frac{\alpha(\lambda)}{4} \right) \\ &\equiv p_{12,\lambda_0} + p_{12,\lambda}. \end{aligned}$$

First, we establish the upper bound of $p_{12,\lambda}$ and p_{12,λ_0} . The probability $p_{12,\lambda}$ is abounded above by

$$p_{12,\lambda} \leq P \left(E(ADPC_{\mathcal{M}}(\lambda) | Y_n) - ADPC_{\mathcal{M}}(\lambda) \geq \frac{\alpha(\lambda)}{4} \right) \leq \frac{4}{\sqrt{m\alpha(\lambda)}} \sqrt{E(\text{var}(ADPC_{\mathcal{M}}(\lambda) | Y_n))},$$

where

$$E(\text{var}(ADPC_{\mathcal{M}}(\lambda) | Y_n)) \leq \text{var} \left(\|Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \mu_n\|_2^2 + 2\mu_n' Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \epsilon_n \right) + \|Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \epsilon_n\|_2^2 + \|Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \epsilon_n^*\|_2^2$$

$$\text{Moreover, } \text{var} \left(\|Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \mu_n\|_2^2 \right) \leq \|\mu_n\|_2^4, \text{var} \left(\mu_n' Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \epsilon_n \right) \leq n \|\mu_n\|_2^2 \sigma^2,$$

$$\text{var} \left(\|Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \epsilon_n\|_2^2 \right) \leq n^2 \sigma^4 \text{ and } \text{var} \left(\|Q(\hat{M}_{\lambda,j}^{\tau\lambda}) \epsilon_n^*\|_2^2 \right) \leq n^2 \sigma^4. \text{ In conclusion,}$$

$$p_{12} \leq \frac{d_{12} n \sigma^2}{\alpha(\lambda) \sqrt{m}} \text{ where } d_{12} > 0 \text{ is a generic constant.}$$

Second, let $\hat{M}_{\lambda}^{\diamond} = \argmin \left(\|Q(M)(\mu_n + \epsilon_n^*)\|_2^2 + \lambda |M| \sigma^2 \right)$, and concerning $p_{11,\lambda}$ we establish

$$\begin{aligned} p_{11,\lambda} &= P \left(R(\hat{M}_{\lambda}) - E(ADPC(\lambda) | Y_n) + \|\epsilon_n\|_2^2 \geq \frac{\alpha(\lambda)}{4} \right) \\ &\leq P \left(R(\hat{M}_{\lambda}) - E \left(\|Q(\hat{M}_{\lambda}^{\tau\lambda}) \mu_n\|_2^2 + 2\epsilon_n' Q(\hat{M}_{\lambda}^{\tau\lambda}) \mu_n - \|P(\hat{M}_{\lambda}^{\tau\lambda}) \epsilon_n\|_2^2 + 2\|P(\hat{M}_{\lambda}^{\tau\lambda}) \epsilon_n^*\|_2^2 \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{4} \\ &\leq p_{11,\lambda,1} + p_{11,\lambda,2}, \end{aligned}$$

where $2\epsilon_n' Q(\hat{M}_{\lambda}^{\tau\lambda}) \mu_n$ is uniformly ignorable for $\|Q(\hat{M}_{\lambda}^{\tau\lambda}) \mu_n\|_2^2$ in $p_{11,\lambda,1}$ and

$$p_{11,\lambda,1} = P \left(R(\hat{M}_{\lambda}) - E \left(\|Q(\hat{M}_{\lambda}^{\tau\lambda}) \mu_n\|_2^2 + \|P(\hat{M}_{\lambda}^{\tau\lambda}) \epsilon_n^*\|_2^2 \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{8}$$

$$p_{11,\lambda,2} = P \left(E \left(\|P(\hat{M}_{\lambda}^{\tau\lambda}) \epsilon_n\|_2^2 - \|P(\hat{M}_{\lambda}^{\tau\lambda}) \epsilon_n^*\|_2^2 \right) | Y_n \geq \frac{\alpha(\lambda)}{8} \right).$$

As for $p_{11,\lambda,1}$, we establish

$$\begin{aligned}
p_{11,\lambda,1} &\leq P \left(R(\hat{M}_\lambda) - E \left(\left(\|Q(\hat{M}_\lambda^{\tau_\lambda}) \mu_n\|_2^2 + \|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n^*\|_2^2 \right) \left(I_{\{\hat{M}_\lambda^\tau = \hat{M}_\lambda^\diamond\}} + I_{\{\hat{M}_\lambda^\tau \neq \hat{M}_\lambda^\diamond\}} \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{8} \right) \\
&\leq P \left(E \left(\left(\|Q(\hat{M}_\lambda^{\tau_\lambda}) \mu_n\|_2^2 + \|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n^*\|_2^2 \right) \left(I_{\{\hat{M}_\lambda^\tau = \hat{M}_\lambda^\diamond\}} \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{16} \right) \\
&\quad + P \left(E \left(\left(\|Q(\hat{M}_\lambda^\diamond) \mu_n\|_2^2 + \|P(\hat{M}_\lambda^\diamond) \epsilon_n^*\|_2^2 \right) \left(I_{\{\hat{M}_\lambda^\tau \neq \hat{M}_\lambda^\diamond\}} \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{16} \right) \\
&\leq \frac{16 \sqrt{(n^2 + \|\mu_n\|_4^2) P(\hat{M}_\lambda^\tau \neq \hat{M}_\lambda^\diamond)}}{\alpha(\lambda)},
\end{aligned}$$

where $P(\hat{M}_\lambda^\tau \neq \hat{M}_\lambda^\diamond) \leq P(\hat{M}_\lambda^\tau \neq \hat{M}_\lambda^\diamond; \hat{M}_\lambda^\tau \in \mathcal{M}^T) + P(\hat{M}_\lambda^\tau \neq \hat{M}_\lambda^\diamond; \hat{M}_\lambda^\tau \in \mathcal{M}^F)$. Thus, $p_{11,\lambda,1} \rightarrow \infty$ as $n \rightarrow \infty$ by making use of the probability upper bound (25).

Next we establish

$$\begin{aligned}
p_{11,\lambda,2} &= P \left(E \left(\left(\|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n\|_2^2 - \|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n^*\|_2^2 \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{8} \right) \\
&\leq P \left(E \left(\left(\|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n\|_2^2 - \lambda |\hat{M}_\lambda^{\tau_\lambda}| \sigma^2 + \lambda |\hat{M}_\lambda^{\tau_\lambda}| \sigma^2 - \|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n^*\|_2^2 \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{8} \right) \\
&\leq P \left(E \left(\left(\|P(\hat{M}_\lambda^\diamond) \epsilon_n\|_2^2 - \lambda |\hat{M}_\lambda^\diamond| \sigma^2 Y_n \right) + E \left(\lambda |\hat{M}_\lambda^\diamond| \sigma^2 - \|P(\hat{M}_\lambda^\diamond) \epsilon_n^*\|_2^2 \right) \geq \frac{\alpha(\lambda)}{16} \right) \right) \\
&\quad + P \left(E \left(-\lambda |\hat{M}_\lambda^\diamond| \sigma^2 + \|P(\hat{M}_\lambda^\diamond) \epsilon_n^*\|_2^2 - E \left(\|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n^*\|_2^2 - \lambda |\hat{M}_\lambda^{\tau_\lambda}| \sigma^2 \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{32} \right) \\
&\quad + P \left(E \left(-\lambda |\hat{M}_\lambda^\diamond| \sigma^2 + \|P(\hat{M}_\lambda^\diamond) \epsilon_n^*\|_2^2 + E \left(\|P(\hat{M}_\lambda^{\tau_\lambda}) \epsilon_n\|_2^2 - \lambda |\hat{M}_\lambda^{\tau_\lambda}| \sigma^2 \right) | Y_n \right) \geq \frac{\alpha(\lambda)}{32} \right)
\end{aligned}$$

which converges to 0 as $n \rightarrow \infty$. This completes the proof.

References

- [1]. Akaike, H. In: Petrov, B.; Csaki, F., editors. Information theory and an extension of the maximum likelihood principle; Second International Symposium on Information Theory; Budapest. 1973; p. 267-281. Akademiai Kiado
- [2]. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*. 2011; 5:232–253. [PubMed: 22081779]
- [3]. Breiman L. The little bootstrap and other methods for dimensionality selection in regression: X-fitted prediction error. *Journal of the American Statistical Association*. 1992; 87:738–754.
- [4]. Chen J, Chen Z. Extended Bayesian information criterion for model selection with large model space. *Biometrika*. 2008; 95:759–771.
- [5]. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*. 1983; 78:316–331.
- [6]. Efron B. The estimation of prediction error: covariance penalties and cross-validation (with discussion). *Journal of the American Statistical Association*. 2004; 99:691–632.
- [7]. Efron B. Estimation and accuracy after model selection (working report). 2013
- [8]. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1361.
- [9]. Foster DP, George EI. The risk inflation criterion for multiple regression. *The Annals of Statistics*. 1994; 22:1947–1975.
- [10]. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*. 2008; 33(1):1–22. [PubMed: 20808728]

- [11]. Geisser S. The predictive sample reuse method with applications. *Journal of the American Statistical Association*. 1975; 70:320–328.
- [12]. George EI, Foster DP. Calibration and empirical Bayes variable selection. *Biometrika*. 2000; 87:731–747.
- [13]. Lopez JA, Walter CA. Evaluating covariance matrix forecasts in a value-at-risk framework. *Journal of Risk*. 2001; 3:69–98.
- [14]. Mallows CL. Some comments on C_p . *Technometrics*. 1973; 15:661–675.
- [15]. Michaud RO. The Markowitz optimization enigma: Is ‘optimized’ optimal? *Financial Analyst Journal*. 1989; 45(1):31–42.
- [16]. Pourahmadi M. Covariance estimation: The GLM and regularization perspectives. *Statistical Science*. 2011; 26:369–387.
- [17]. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–464.
- [18]. Shao J. An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*. 1997; 7:221–264.
- [19]. Shen X, Ye J. Adaptive model selection. *Journal of the American Statistical Association*. 2002; 97:210–221.
- [20]. Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*. 2012; 107:223–232. [PubMed: 22736876]
- [21]. Shibata R. An optimal selection of regression variables. *Biometrika*. 1981; 68:45–54.
- [22]. Stone M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*. 1974; 36:111–147.
- [23]. Tibshirani R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, B*. 1996; 58:267–288.
- [24]. Tibshirani R, Knight K. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B*. 1999; 61:529–546.
- [25]. Whittle P. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Application*. 1960; 5:302–305.
- [26]. Yang Y. Consistency of cross validation for comparing regression procedures. *Annals of Statistics*. 2007; 35:2450–2473.
- [27]. Ye J. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*. 1998; 93:120–131.
- [28]. Yuan Z, Yang Y. Combining linear regression models: When and how? *Journal of the American Statistical Association*. 2005; 100:1202–1214.
- [29]. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*. 2010; 38:894–942.
- [30]. Zhang Y, Shen X. Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining*. 2010; 3:350–358. [PubMed: 21116443]

Figure 1.

Comparison of averaging MSE of $\hat{U}_{\mathcal{M}}(\lambda)$ as $\mathcal{M}^F$ is empty and nonempty for $\lambda=2$ (—), $\log(n)$ (---) and $\sqrt{n}$ (- - -) with $n = 50$, $p = 2$, $q_0 = 1$ and $\sigma = 1$.

Table 1

MSE (with standard error shown in the bracket), the average number of true predictors falsely excluded (FN), the average number of falsely included redundant variables (FP), and the proportions of underfitted, correctly or overfitted models, based on 100 simulation replications, for competing methods with $\rho = -0.5$, $n = 400$, $p = 20$ and $\sigma = 1$.

q_0	Method	MSE	# Non-zeros		% of fits		
			FN	FP	Corr	Under	Over
SNR=2							
1	AIC	12.627 (0.711)	0	3.100	0.070	0	0.930
	BIC	3.779 (0.421)	0	0.260	0.780	0	0.220
	BIC _c	1.878 (0.178)	0	0	1	0	0
	ADPC	1.878 (0.178)	0	0	1	0	0
3	AIC	12.940 (0.710)	0	2.500	0.080	0	0.920
	BIC	5.480 (0.435)	0	0.190	0.820	0	0.180
	BIC _c	4.123 (0.321)	0	0	1	0	0
	ADPC	4.123 (0.321)	0	0	1	0	0
5	AIC	14.779 (0.743)	0	2.500	0.080	0	0.920
	BIC	7.379 (0.493)	0	0.240	0.790	0	0.210
	BIC _c	5.514 (0.290)	0	0	1	0	0
	ADPC	5.514 (0.290)	0	0	1	0	0
SNR=0.5							
1	AIC	12.987 (0.700)	0	3.050	0.060	0	0.940
	BIC	4.021 (0.458)	0	0.250	0.780	0	0.220
	BIC _c	2.020 (0.185)	0	0	1	0	0
	ADPC	2.020(0.185)	0	0	1	0	0
3	AIC	12.721 (0.793)	0	2.592	0.031	0	0.969
	BIC	5.164 (0.519)	0.000	0.204	0.837	0.010	0.163
	BIC _c	5.437 (1.099)	0.061	0.000	0.667	0.969	0.031
	ADPC	5.164 (0.519)	0	0.204	0.837	0.010	0.163
5	AIC	14.229 (0.732)	0	2.414	0.141	0	0.859
	BIC	8.061 (0.601)	0.010	0.293	0.768	0.010	0.232
	BIC _c	52.470 (3.066)	2.475	0	0.212	0.788	0
	ADPC	8.061 (0.601)	0.010	0.293	0.768	0.010	0.232

Table 2

MSE (with standard error shown in the bracket), the average number of true predictors falsely excluded (FN), the average number of falsely included redundant variables (FP), and the proportions of underfitted, correctly or overfitted models, based on 100 simulation replications, for competing methods with $\rho = -0.5$, $n = 200$, $p = 1000$ and $\sigma = 1$.

q_0	Method	MSE	# Non-zeros		% of fits		
			FN	FP	Corr	Under	Over
SNR=2							
1	TLP+CV	5.637 (0.837)	0	7.570	0.550	0	0.450
	MCP+CV	4.382 (0.510)	0	2.330	0.620	0	0.380
	SCAD+CV	3.733 (0.400)	0	4.120	0.580	0	0.420
	LASSO+CV	14.727 (0.764)	0	11.77	0.150	0	0.850
	ADPC	1.985 (0.191)	0	0.170	0.890	0	0.110
3	TLP+CV	6.881 (0.744)	0	4.920	0.650	0	0.350
	MCP+CV	6.087 (0.604)	0	2.050	0.680	0	0.320
	SCAD+CV	5.299 (0.408)	0	3.110	0.690	0	0.310
	LASSO+CV	45.193(1.594)	0	38.180	0	0	1.000
	ADPC	3.995 (0.254)	0	0.120	0.910	0	0.0900
5	TLP+CV	11.390(2.875)	0.011	3.989	0.600	0.011	0.389
	MCP+CV	8.244 (0.762)	0	1.911	0.700	0	0.300
	SCAD+CV	8.144 (0.592)	0	4.644	0.556	0	0.444
	LASSO+CV	71.461 (1.636)	0	63.400	0	0	1.000
	ADPC	6.554(0.385)	0	0.144	0.878	0	0.122
SNR=0.5							
1	TLP+CV	4.873 (0.787)	0	7.230	0.370	0	0.630
	MCP+CV	6.614 (0.569)	0	3.410	0.300	0	0.700
	SCAD+CV	8.372 (0.710)	0	11.340	0.070	0	0.930
	LASSO+CV	12.989 (0.758)	0	10.520	0.120	0	0.880
	ADPC+CV	1.749 (0.156)	0	1.290	0.520	0	0.480
3	TLP+CV	37.473 (1.316)	0.917	27.444	0	0.736	1.00
	MCP+CV	35.952(0.947)	1.514	5.403	0	0.931	0.764
	SCAD+CV	36.215 (0.916)	1.250	13.958	0	0.847	0.861
	LASSO+CV	36.877 (0.923)	1.278	15.111	0	0.931	0.917
	ADPC	33.927 (0.777)	1.153	16.819	0	0.903	1
5	TLP+CV	43.561 (1.035)	2.827	32.612	0	0.990	1
	MCP+CV	45.332 (0.796)	3.816	5.245	0	0.990	0.673
	SCAD+CV	44.984(0.749)	3.510	13.796	0	0.990	0.755
	LASSO+CV	45.454 (0.859)	3.551	15.296	0	1	0.776
	ADPC+CV	41.326(0.649)	2.929	26.735	0	1	1

Table 3

Mean absolute value (MAPE) and standard deviation (SD) of prediction error.

Procedure	LASSO+CV	SCAD+CV	MAP+CV	TLP+CV	ADPC
MAPE	23.69	1.51	1.46	4.00	1.20
SD	78.13	2.88	2.88	24.70	1.69