TR 681

Submitted to the Annals of Statistics
arXiv: math.PR/0000000

PARAMETRIC OR NONPARAMETRIC? A PARAMETRICNESS INDEX FOR MODEL SELECTION

BY WEI LIU* AND YUHONG YANG*

University of Minnesota

In model selection literature two classes of criteria perform well asymptotically in different situations: Bayesian information criterion (BIC) (as a representative) is consistent in selection when the true model is finite dimensional (parametric scenario); Akaike's information criterion (AIC) performs well when the true model is infinite dimensional (nonparametric scenario). But there is little work that addresses if it is possible and how to detect the situation that a specific model selection problem is in. In this work, we differentiate the two scenarios theoretically. We develop a measure, parametricness index (PI), to assess whether a model selected by a consistent procedure can be practically treated as the true model, which also hints on AIC or BIC is better suited for the data. A consequence is that by switching between AIC and BIC based on the PI, the resulting regression estimator is simultaneously asymptotically efficient for both parametric and nonparametric scenarios. In addition, we systematically investigate the behaviors of PI in simulation and real data and show its usefulness.

1. Introduction. When considering parametric models for data analysis, model selection methods have been commonly used for various purposes. If one candidate model describes the data really well (e.g., a physical law), it is obviously desirable to identify it. Consistent model selection rules such as BIC [50] are proposed for this purpose. In contrast, when the candidate models are constructed to progressively approximate an infinite-dimensional truth with a decreasing approximation error, the main interest is usually on estimation and one hopes that the selected model performs optimally in terms of a risk of estimating a target function (e.g., the regression function). AIC [2] has been shown to be the right criterion from an asymptotic efficiency and also a minimax-rate optimality views.

The question if we can statistically distinguish between parametric and nonparametric scenarios motivated our research. In this paper, for regres-

*Supported by NSF grant DMS-0706850.

AMS 2000 subject classifications: Primary 62J05, 62F12; secondary 62J20

Keywords and phrases: Model selection, parametricness index (PI), Model selection diagnostics

sion based on finite-dimensional models, we develop a simple parametricness index (PI) that has the following properties.

1. With probability going to 1, PI separates typical parametric and nonparametric scenarios.
2. It advises on whether identifying the true or best candidate model is feasible at the given sample size or not by assessing if one of the models stands out as a stable parametric description of the data.
3. It informs us whether interpretation and statistical inference based on the selected model are reasonably reliable or not due to model selection uncertainty.
4. It tells us whether AIC is likely better than BIC or not for the data at hand.
5. It can be used to approximately achieve the better estimation performance of AIC and BIC for both parametric and nonparametric scenarios.

A tremendous amount of exciting research has been done on theory, computation and application of various model selection methods. However, comparisons of different model selection criteria are still mostly limited to scattered and selective numerical studies and asymptotic investigations that do not yet provide clear guidelines for real data analysis. In our view, model selection diagnostic measures that address reliability and comparison of different model selection methods are fundamentally important for a sound statistical practice.

In the rest of the introduction, we provide a relevant background of model selection and present views on some fundamental issues.

1.1. *Model selection criteria and their possibly conflicting properties.* To assess performance of model selection criteria, pointwise asymptotic results (e.g., [18, 26, 39, 45, 46, 47, 48, 51, 54, 58, 60, 64, 68, 71, 72]) have been established mostly in terms of either selection consistency or an asymptotic optimality. It is well-known that AIC [2], C_p [44], and FPE [1, 55] have an asymptotic optimality property which says the accuracy of the estimator based on the selected model is asymptotically the same as the best candidate model when the true model is infinite dimensional. In contrast, BIC and the like are consistent when the true model is finite-dimensional and is among the candidate models.

Another direction of model selection theory focuses on oracle risk bounds (also called index of resolvability bounds) that immediately lead to minimax type of results. That is, a risk bound with minimal assumptions is derived for a model selection method, which link the risk of the selected model to

the smallest risk among all the candidate models. Given a class of functions to which the target function belong, by maximizing the risk bound over the function class, a bound on the worst-case risk of the model-selection-based estimator is readily obtained. When the candidate models are constructed to work well for target function classes, this yields minimax-rate or near minimax-rate optimality results. Publications of work in this direction include [3, 4, 5, 6, 10, 14, 15, 22, 23, 41, 66], to name a few. In particular, AIC type of model selection methods are minimax-rate optimal for both parametric and nonparametric scenarios (see [5, 63]). A remarkable feature of the works inspired by [6] is that with a complexity penalty (other than one in terms of model dimension) added to deal with a large number of (e.g., exponentially many) models, the resulting risk or loss of the selected model automatically achieves the best trade-off between approximation error, estimation error and the model complexity, which provides tremendous theoretical flexibility to deal with a fixed countable list of models (e.g., for series expansion based modeling) or a list of models chosen to depend on the sample size (see, e.g., [5, 66, 61]).

While pointwise asymptotic results are certainly interesting, since they describe limiting behaviors when the sample size approaches infinity, similarly to the issue of asymptotic normality of the sample mean from a possibly highly non-Gaussian density, it is not surprising that the limiting behaviors can be very different from the finite-sample reality, especially when model selection is involved. For finite-sample results and effects of model selection on post-model-selection estimators, see e.g., [21, 42]. In particular, it is argued there that the use of a consistent model selection procedure does not necessarily allows one to act as if the true model were known in advance.

The general forms of AIC and BIC make it very clear that they and similar criteria (such as GIC in [49]) cannot simultaneously enjoy the properties of consistency in a parametric scenario and asymptotic optimality in a nonparametric scenario. Efforts have been put on using penalties that are data-dependent and adaptive (see, e.g., [7, 30, 33, 38, 52, 53, 65]). Shen and Ye [53] proposed an adaptive model selection procedure to approximate the best performance of a class of procedures across a variety of situations. Yang [65] showed that the asymptotic optimality of BIC for a parametric scenario (which follows directly from consistency of BIC) and asymptotic optimality of AIC for a nonparametric scenario can be shared by an adaptive model selection criterion. A similar two-stage adaptive model selection rule for time series autoregression has been proposed by Ing [38]. However, Yang [63, 65] proved that no model selection procedure can be both consistent (or pointwise adaptive) and minimax-rate optimal at the same time. More recently,

Erven, Grünwald, de Rooij [25] found that if a cumulative risk (i.e., the sum of risks from the sample size 1 to n) is considered instead of the usual risk at sample size n, then the conflict between consistency in selection and minimax-rate optimality can be resolved by a Bayesian strategy that allows switching between models. As will be seen, if we can properly distinguish between parametric and nonparametric scenarios, a consequent choice of AIC or BIC simultaneously achieves asymptotic efficiency for both parametric and nonparametric situations.

1.2. *Model selection: A gap between theory and practice.* It is well-known that for a typical regression problem with a number of predictors, AIC and BIC tend to choose models of significantly different sizes, which may have serious practical consequences. Therefore, it is important to decide which criterion to apply for a data set at hand. Unfortunately, the current theories on model selection have little to offer to address this issue. Consequently, it is rather common that statisticians/statistical users resort to the "faith" that the true model certainly cannot be finite-dimensional for the choice of AIC, or to the strong preference of parsimony or goal of model identification to defend his/her use of BIC.

To us, this disconnectedness between theory and practice of model selection needs not to continue. From various angles, the question whether or not AIC is more appropriate than BIC for the data at hand should and can be addressed statistically rather than based on one's preferred assumption. This is the major motivation for us to try to go beyond presenting a few theorems in this work.

Model selection is fundamentally important for statistics or even sciences. Indeed, the conflict between AIC and BIC has received a lot of attention not only in the statistics literature but also in fields such as psychology and biology [56]. There has been a lot of debate in the literature regarding model selection procedures that is not only statistical but also philosophical, especially about the existence of a true model and the ultimate goal of statistical modeling (see, e.g., [8, 13, 16, 29, 70], and references therein). Some researchers have no problem with assuming existence of a true finite-dimensional model while others regard it as a fiction.

We would like to quote a leading statistician here:

"*It does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. The idea that complex physical, biological, and sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a*

vital part of general scientific analysis and statistical models, especially substantive ones (Cox, 1990), do not seem essentially different from other kinds of model. " (Cox [20])

Fisher in his pathbreaking 1922 paper [28], provided thoughts on the foundations of statistics, including model specification. He stated: "More or less elaborate forms will be suitable according to the volume of the data". Cook [19] discussed Fisher's insights in details.

We certainly agree with the statements by Fisher and Cox. What we are interested in this and future work on model selection is to address the general question that in what ways and to what degrees a selected model is useful. In this paper, after answering the theoretical question if we can construct a measure that consistently tells us if we are in an AIC scenario or BIC scenario, we propose a practically relevant use of the measure based on our views on model selection.

We have mixed feelings towards the concept of consistency (i.e., the property that the true model, assumed to be among the candidates, will be selected with probability going to 1 as the sample size goes to infinity). One indeed should not read too much into it either for interpretation or for estimation because the asymptotic view is often overly optimistic. First, the property of consistency often comes with restrictive assumptions among which is the existence of a finite-dimensional true model, which is strongly objected by many. Second, even if the finite-dimensional true model assumption is justifiable, a consistent model selection method does not necessarily perform well due to the often high model selection uncertainty and the necessary lack of protection in the worst case of every consistent model selection method as a price paid to aggressively pursue a lower-dimensional alternative to explain the data.

On the other hand, finding a stable finite-dimensional model to describe the nature of the data as well as to predict the future is very appealing. Following up in the spirit of Cox mentioned above, if a model stably stands out among the competitors, whether it is the true model or not, from a practical perspective, why should not we extend the essence of consistency to mean the ability to find it? In our view, if we are to accept any statistical model (say infinite-dimensional) as a useful vehicle to analyze data, it is difficult to philosophically reject the more restrictive assumption of a finite-dimensional model, because both are convenient and certainly simplified descriptions of the reality, their difference being that between 50 paces and 100 paces as in the 2000 year old Chinese idiom *One who retreats fifty paces mocks one who retreats a hundred.*

The above considerations lead to our second question: Can we construct a

practical measure that gives us a proper indication on whether the selected model deserves to be crowned as the best model *at the time being*? We emphasize *at the time being* to make it clear that we are not going after the best limiting model (no matter how that is defined), but instead we seek a model that stands out for sample sizes around what we have now.

Of course, when we do not assume the true model is finite-dimensional, unavoidably we have to specify a performance measure in order to define that a model stably stands out.

While there are many different performance measures that we can use to assess if one model stands out, following our results on distinguishing between parametric and nonparametric scenarios, we focus on an estimation/prediction accuracy measure. We call it *parametricness index* (PI), which is relative to the list of candidate models and the sample size. Our theoretical results show that this index converges to infinity for a parametric scenario and converges to 1 for a typical nonparametric scenario. Our suggestion is that when the index is significantly larger than 1, we can treat the selected model as a stably standing out model from the estimation perspective. Otherwise, the selected model is just among a few or more equally well-performing candidates. We call the former case practically parametric and the latter practically nonparametric.

As will be demonstrated in our numerical work, PI can be close to 1 for a truly parametric scenario and large for a nonparametric scenario. In our view, this is not a problem. For instance, for a truly parametric scenario with many small coefficients of various magnitudes, for a small or moderate sample size, the selected model will most likely be different from the true model and it is also among multiple models that perform similarly in estimation of the regression function. We would view this as "practically nonparametric" in the sense that with the information available we are not able to find a single standing-out model and the model selected provides a good trade-off between approximation capability and model dimension. In contrast, even if the true model is infinite-dimensional, at a given sample size, it is quite possible that a number of terms are significant and others are too small to be relevant at the given sample size. Then we are willing to call it "practically parametric" in the sense that as long as the sample size is not substantially increased, the same model is expected to perform better than the other candidates. For example, in properly designed experimental studies, when a working model clearly stands out and is very stable, then it is desirable to treat it as a parametric scenario even though we know surely it is an approximating model. This is often the case in physical sciences when a law-like relationship is evident under controlled experimental conditions.

Note that given an infinite-dimensional true model and a list of candidate models, we may declare the selected models to be practically parametric for some sample sizes and to be practically nonparametric for others.

For the numerical work to investigate the performance of our methodology, instead of giving two or three favorable examples, we intend to study various representative scenarios so as to get informative and fair numerical results.

The rest of the paper is organized as follows. In Section 2, we set up the regression framework and give some notations. We then in Section 3 develop the measure PI and show that theoretically it differentiates a parametric scenario from a nonparametric one under some conditions for both known and unknown σ^2 respectively. Consequently, the pointwise asymptotic efficiency properties of AIC and BIC can be combined for parametric and nonparametric scenarios. In Section 4, we propose a proper use of PI for applications. Simulation studies and real data examples are reported in Sections 5 and 6, respectively. Concluding remarks are given in Section 7 and the proofs are in an appendix.

2. Setup of the regression problem. Consider the regression model

$$Y_i = f(x_i) + \epsilon_i \quad i = 1, 2, \cdots, n,$$

where $x_i = (x_{i1}, \cdots, x_{id})$ is the value of a d-dimensional fixed design variable at the ith observation, Y_i is the response, f is the true regression function, and the random errors ϵ_i are assumed to be independent and normally distributed with mean zero and variance σ^2.

To estimate the regression function, a list of linear models are being considered, from which one is to be selected:

$$Y = f_k(x, \theta_k) + \epsilon,$$

where, for each k, $\mathcal{F}_k = \{f_k(x, \theta_k), \theta_k \in \Theta_k\}$ is a linear family of regression functions with θ_k being the parameter of finite dimension m_k. Let Γ be the collection of the model indices k. Γ can be fixed or change with the sample size.

The above framework includes the usual subset-selection and order-selection problems in linear regression. It also includes nonparametric regression based on series expansion, where the true function is approximated by linear combinations of appropriate basis functions, such as polynomials, splines or wavelets.

Parametric modeling typically intends to capture the essence of the data by a finite-dimensional model, and nonparametric modeling tries to achieve

the best trade-off between approximation error and estimation error for a target infinite-dimensional function. See, e.g., [67] for general relationship between rate of convergence for function estimation and full or sparse approximation based on a linear approximating system.

Theoretically speaking, the essential difference between parametric and nonparametric scenarios is that the best model has no approximation error for the former and all the candidate models have non-zero approximation errors for the latter.

In this paper we consider the least squares estimators when defining the parametricness index, although the model being examined can be based any consistent model selection method that may or may not involve least squares estimation.

Notation and definitions. Let $\mathbf{Y}_n = (Y_1, \cdots, Y_n)^T$ be the response vector and M_k be the projection matrix for model k. Denote $\hat{\mathbf{Y}}_k = M_k \mathbf{Y}_n$. Let $f_n = (f(x_1), \cdots, f(x_n))^T$, $e_n = (\epsilon_1, \cdots, \epsilon_n)^T$, and I_n be the identity matrix. Let $\|\cdot\|$ denote the Euclidean distance in the R^n space, and let $TSE(k) = \|f_n - \hat{\mathbf{Y}}_k\|^2$ be the total square error of the LS estimator from model k.

Let the rank of M_k be r_k. In this work, we do not assume that all the candidate models have the rank of the design matrix equal the model dimension m_k, which may not hold when a large number of models are considered. Let N_j denote the number of models with $r_k = j$ for $k \in \Gamma$. For a given model k, let $S_1(k)$ be the set of all sub-models k' of k in Γ such that $r_{k'} = r_k - 1$. Throughout the paper, for technical convenience, we assume $S_1(k)$ is not empty for all k with $r_k > 1$.

For a sequence $\lambda_n > 0$ and a constant $d \geq 0$, let

$$IC_{\lambda_n, d}(k) = \|\mathbf{Y}_n - \hat{\mathbf{Y}}_k\|^2 + \lambda_n \log(n) r_k \sigma^2 - n\sigma^2 + d n^{1/2} \log(n) \sigma^2$$

when σ is known, and

$$IC_{\lambda_n, d}(k, \hat{\sigma}^2) = \|\mathbf{Y}_n - \hat{\mathbf{Y}}_k\|^2 + \lambda_n \log(n) r_k \hat{\sigma}^2 - n\hat{\sigma}^2 + d n^{1/2} \log(n) \hat{\sigma}^2$$

when σ is estimated by $\hat{\sigma}$. A discussion on choice of λ_n and d will be given later in Section 3.5. We emphasize that our use of $IC_{\lambda_n, d}(k)$ or $IC_{\lambda_n, d}(k, \hat{\sigma}^2)$ is for defining the parametricness index as below and it may not be the one used for model selection.

3. Main Theorems. Consider a consistent model selection method. Let $\hat{k}_n$ be the selected model at sample size n. We define the *parametricness index* (PI) as follows:

1. When σ is known, $PI_n = \begin{cases} \inf_{k \in S_1(\hat{k}_n)} \frac{IC_{\lambda_n, d}(k)}{IC_{\lambda_n, d}(\hat{k}_n)} & \text{if } r_{\hat{k}_n} > 1 \\ n & \text{if } r_{\hat{k}_n} = 1 \end{cases}$;
2. When σ is estimated by $\hat{\sigma}$,

$$PI_n = \begin{cases} \inf_{k \in S_1(\hat{k}_n)} \frac{IC_{\lambda_n, d}(k, \hat{\sigma}^2)}{IC_{\lambda_n, d}(\hat{k}_n, \hat{\sigma}^2)} & \text{if } r_{\hat{k}_n} > 1 \\ n & \text{if } r_{\hat{k}_n} = 1 \end{cases}.$$

The reason behind the definition is that a correctly specified parametric model must be very different from any sub-model (bias of a sub-model is dominatingly large asymptotically speaking), but for a nonparametric scenario, the model selected is only slightly affected in terms of estimation accuracy when one or a few least important terms are dropped. When $r_{\hat{k}_n} = 1$, the value of PI is arbitrarily defined as long as it goes to infinity as n increases.

3.1. *Parametric Scenarios.* Now consider a *parametric scenario* where the true finite dimensional model at sample size n is indexed by $k_n^* \in \Gamma$ with $r_{k_n^*} > 1$. Let $A_n = \inf_{k \in S_1(k_n^*)} \|(I_n - M_k) f_n\|^2 / \sigma^2$. Note that A_n/n is the best approximation error (squared bias) of models in $S_1(k_n^*)$.

Conditions:

(P1). There exists $0 < \tau \leq \frac{1}{2}$ such that A_n is of order $n^{\frac{1}{2}+\tau}$ or higher.
(P2). The dimension of the true model does not grow too fast with sample size n in the sense that $r_{k_n^*} \lambda_n \log(n) = o(n^{\frac{1}{2}+\tau})$.
(P3). The selection procedure is consistent: $P(\hat{k}_n = k_n^*) \to 1$ as $n \to \infty$.

THEOREM 1. *Assume Conditions (P1)-(P3) are satisfied for the parametric scenario.*

(i). With σ^2 known, we have

$$PI_n \xrightarrow{p} \infty \quad \text{as } n \to \infty.$$

(ii). When σ is unknown, let $\hat{\sigma}_n^2$ be the unbiased estimator of σ^2 from the selected model. We also have

$$PI_n \xrightarrow{p} \infty \quad \text{as } n \to \infty.$$

Remarks:

1. If the number of models of each dimension is of a polynomial order in n, then Condition (P1) can be relaxed. For example, it is sufficient to require A_n of an order higher than $n^{\frac{1}{2}}(\log(n))^{\lambda}$ for some $\lambda > 1$.

Note that in a conventional case with Γ fixed, A_n is typically of order n and (P1) certainly holds. The conditions basically eliminates the case that the true model and a sub-model with one fewer term are not distinguishable with the information available in the sample.

2. In our formulation above, we considered comparison of two immediately nested models. One can consider comparing two nested models with size difference m ($m > 1$) and similar results hold.
3. The case $\lambda_n = 1$ corresponds to using BIC in defining the PI. And $\lambda_n = 2/\log(n)$ corresponds to using AIC.
4. When the number of predictors increases with n, Chen and Chen [17] showed that a higher penalty than BIC leads to consistency in all subset selection under some conditions.

3.2. *Nonparametric Scenarios.* Now the true model at each sample size n is not in the list Γ and may change with sample size, which we call a *nonparametric scenario.* For $j < n$, denote

$$B_{j,n} = \inf_{k\in\Gamma}\{(\lambda_n \log(n) - 1)j + \|(I_n - M_k)f_n\|^2/\sigma^2 + dn^{1/2}\log(n) \ : r_k = j\},$$

where the infimum is taken over all the candidate models with $r_k = j$. For $1 < j < n$, let $L_j = \max_{k\in\Gamma}\{card(S_1(k)) : r_k = j\}$. Let $P_{k^{(s)},k} = M_k - M_{k^{(s)}}$ be the difference between the projection matrices of the two nested models. Clearly, $P_{k^{(s)},k}$ is the projection matrix onto the orthogonal complement of the column space of model $k^{(s)}$ with respect to that of the larger model k.

Conditions: There exist two sequences of integers $1 \le a_n < b_n < n$ (not necessarily known) with $a_n \to \infty$ such that the following holds.

(N1). $P(a_n \le r_{\hat{k}_n} \le b_n) \to 1$ and $\sup_{a_n\le j\le b_n} \frac{B_{j,n}}{n-j} \to 0$ as $n \to \infty$.

(N2). There exist a positive sequence $\zeta_n \to 0$ and constants $c_0 > 0$ such that for $a_n \le j \le b_n$,

$$N_j \cdot L_j \le c_0 e^{\zeta_n B_{j,n}}, \ N_j \le c_0 e^{\frac{B_{j,n}^2}{10(n-j)}}, \ \text{and} \ \limsup_{n\to\infty} \sum_{j=a_n}^{b_n} e^{-\frac{B_{j,n}^2}{10(n-j)}} = 0.$$

(N3). $\limsup_{n\to\infty}\left[\sup_{\{k:\ a_n\le r_k\le b_n\}} \frac{\inf_{k^{(s)}\in S_1(k)}\|P_{k^{(s)},k}f_n\|^2}{(\lambda_n\log(n)-1)r_k+\|(I_n-M_k)f_n\|^2/\sigma^2+dn^{1/2}\log(n)}\right] = 0.$

THEOREM 2. *Assuming Conditions (N1)-(N3) are satisfied for a nonparametric scenario and σ^2 is known, then we have*

$$PI_n \xrightarrow{p} 1 \quad as \ \ n \to \infty.$$

Remarks:

1. The first part of Condition (N1) says that the dimension of the selected model lies in a range $[a_n, b_n]$ with probability going to 1, where $a_n \to \infty$. For nonparametric regression, for familiar model selection methods, the order of r_{k_n} can be identified (e.g., [38, 67]), sometimes loosing a logarithmic factor. The requirement $\frac{B_{j,n}}{n-j} \to 0$ is easily satisfied in a typical nonparametri scenario.
2. Condition (N2) basically ensures that the number of subset models of each dimension does not grow too fast relative to $B_{j,n}$. When the best model has a slower rate of convergence in regression estimation, more candidate models can be allowed without detrimental selection bias.
3. Roughly speaking, Condition (N3) says that when the model dimension is in a range that contains the selected model with probability approaching 1, the least significant term in the regression function projection is negligible compared to the sum of approximation error, the dimension of the model times $\lambda_n \log(n)$, and the term $dn^{1/2}\log(n)$. This condition is mild.
4. A choice of $d > 0$ can handle situations where the approximation error decays fast, e.g., exponentially fast (see Section 3.4), in which case the stochastic fluctuation of $IC_{\lambda_n,d}$ with $d = 0$ is relatively too large for PI to converge to 1 in probability. In applications, for separating reasonably distinct parametric and nonparametric scenarios, we recommend the choice of $d = 0$.
5. Except for (N1), no specific properties of the model selection rule are assumed for this result.

When σ^2 is unknown but estimated from the selected model, PI_n is correspondingly defined. For $j < n$, let

$$E_{j,n} = \inf_{k\in\Gamma,\ r_k=j} \left\{ \left[(\lambda_n \log(n) - 1)j + dn^{1/2}\log(n)\right] \left[1 + \|(I_n - M_k)f_n\|^2/((n-j)\sigma^2)\right] \right\}$$

Conditions: There exist two sequences of integers $1 \le a_n < b_n < n$ with $a_n \to \infty$ such that the following holds.

(N2′). There exist a positive sequence $\rho_n \to 0$ and constant $c_0 > 0$ such that for $a_n \le j \le b_n$,

$$N_j \cdot L_j \le c_0 e^{\rho_n E_{j,n}}, \text{ and } \limsup_{n\to\infty} \sum_{j=a_n}^{b_n} e^{-\rho_n E_{j,n}} = 0.$$

(N3′). $\limsup_{n\to\infty} \left[\sup_{\{k:\ a_n \le r_k \le b_n\}} \frac{\inf_{k(s)} \|P_{k(s),k} f_n\|^2}{[(\lambda_n \log(n)-1)r_k + dn^{1/2}\log(n)][1+\|(I_n - M_k)f_n\|^2/(\sigma^2(n-r_k))]}\right] = 0.$

THEOREM 3. *Assuming Conditions* $(N1)$, $(N2')$, *and* $(N3')$ *hold for a nonparametric scenario, then we have*

$$PI_n \xrightarrow{p} 1 \quad as \quad n \to \infty.$$

Remarks:

1. Conditions $(N2')$ and $(N3')$ have similar interpretations as $(N2)-(N3)$ except that $B_{j,n}$ is replaced by $E_{j,n}$. Conditions $(N2')$ and $(N3')$ are stronger than $(N2)$ and $(N3)$ due to estimation of σ^2.
2. For the σ^2 unknown case, we used $\hat{\sigma}_n^2$ from the selected model and hence condition $(N3')$. The conditions $(N2')$ and $(N3')$ may be relaxed if more is known about $\hat{k}_n$, which is possible under smoothness assumptions on the regression function.

3.3. *PI separates parametric and nonparametric scenarios.* The results in Sections 3.1 and 3.2 say that with a consistent model selection procedure, the PI goes to ∞ and 1 in probability in parametric and nonparametric scenarios, respectively.

COROLLARY 1. *Consider a model selection setting where* Γ_n *includes models of sizes approaching* ∞ *as* $n \to \infty$. *Assume the true model is either parametric or nonparametric satisfying (P1)-(P2) or (N1)-(N3), respectively. Then* PI_n *has distinct limits in probability for the two scenarios.*

3.4. *Examples.* We now take a closer look at the Conditions (P1)-(P3) and (N1)-(N3) for two settings: all subset selection and order selection (i.e., the candidate models are nested).

(1). All subset selection

Let p_n be the number of terms to be considered.

(i). Parametric with true model k_n^* fixed.

In this case, A_n is typically of order n for a reasonable design and then Condition (P1) is met. Condition (P2) is obviously satisfied when $\lambda_n = o(n^{\frac{1}{2}})$.

(ii). Parametric with k_n^* changing with sample size n, say $r_{k_n^*}$ increases with n.

In this case, both $r_{k_n^*}$ and p_n go to infinity with the sample size n. Since there are more and more terms in the true model, in order for A_n not to be too small, the terms should not be too highly correlated. Otherwise, the true model may not be distinguishable

from a sub-model based on the data. An extreme case is that one term in the true model is almost linearly dependent on the others. Then $A_n \approx 0$. To understand Condition (P1) in terms of the coefficients in the true model, under an orthogonal design, Condition (P1) is more or less equivalent to that the square of the smallest coefficient in the true model is of order $n^{\tau-1/2}$ or higher. Since τ can be arbitrarily close to 0, the smallest coefficient should basically be larger than $n^{-\frac{1}{4}}$.

Condition (P2) is also satisfied when $\lambda_n r_{k_n^*} \log(n)$ is not too large.

(iii). Nonparametric.

Condition (N1) holds for any model selection method that yields a consistent regression estimator. Note that $N_j = \begin{pmatrix} p_n \\ j \end{pmatrix} < \frac{p_n^j}{j!} < (\frac{p_n \cdot e}{j})^j$ and basically $L_j \leq j$. Then $N_j \leq c_0 e^{\frac{B_{j,n}^2}{10(n-j)}}$ is roughly equivalent to $j \log(p_n/j) \leq [dn^{1/2}\log(n) + \lambda_n \log(n) j + \|(I_n - M_k) f_n\|^2/\sigma^2]^2/10(n-j)$ for $a_n \leq j \leq b_n$. A sufficient condition then is $p_n \leq b_n e^{B_{j,n}^2/(10(n-j)b_n)}$ for $a_n \leq j \leq b_n$. As to the condition that $N_j \cdot L_j \leq c_0 e^{\zeta_n' B_{j,n}}$, as long as $\sup_{a_n \leq j \leq b_n} \frac{B_{j,n}}{n-j} \to 0$, then it is implied by the above one. For the condition that $\sum_{j=a_n}^{b_n} e^{-\frac{B_{j,n}^2}{10(n-j)}} \to 0$, it is automatically satisfied for any $d > 0$ and also satisfied for $d = 0$ when the approximation error does not decay too fast.

For Condition (N3), under an orthonormal design, the requirement is similar to the order selection case (see below).

(2). Order selection in series expansion

We only need to discuss the nonparametric scenario. (The parametric scenarios are similar to the above.)

In this setting, there is only one model of each dimension. So Condition (N2) reduces to: $\sum_{j=a_n}^{b_n} e^{-\frac{B_{j,n}^2}{10(n-j)}} \to 0$. Note that $\sum_{j=a_n}^{b_n} e^{-\frac{B_{j,n}^2}{10(n-j)}} < (b_n - a_n) \cdot e^{-(\log(n))^2/10} < n \cdot e^{-(\log(n))^2/10} \to 0$.

To check Condition (N3), for a demonstration, consider orthogonal designs. Let $\Phi = \{\phi_1(x), \cdots, \phi_k(x), \cdots\}$ be a collection of orthonormal basis functions and the true regression function is $f(x) = \sum_{i=1}^{\infty} \beta_i \phi_i(x)$. For model k, the model with the first k terms, $\inf_{k^{(s)} \in S_1(k)} \|P_{k^{(s)},k} f_n\|^2$ is roughly $\beta_k^2 \|\phi_k(\mathbf{X})\|^2$ and $\|(I_n - M_k) f_n\|^2$ is roughly $\sum_{i=k+1}^{\infty} \beta_i^2 \|\phi_i(\mathbf{X})\|^2$, where $\phi_i(\mathbf{X}) = (\phi_i(x_1), \cdots, \phi_i(x_n))^T$. Since $\|\phi_i(\mathbf{X})\|^2$ is of order n, Con-

dition (N3) is roughly equivalent to the following:

$$\limsup_{n\to\infty}\left[\sup_{a_n\le k\le b_n}\frac{n\beta_k^2}{(\lambda_n\log(n)-1)k+n\sum_{i=k+1}^{\infty}\beta_i^2/\sigma^2+dn^{1/2}\log(n)}\right]=0.$$

Then a sufficient condition for Condition (N3) is that $d=0$ and $\lim_{k\to\infty}\frac{\beta_k^2}{\sum_{i=k+1}^{\infty}\beta_i^2}=0$, which is true if $\beta_k=k^{-\delta}$ for some $\delta>0$ but not true if $\beta_k=e^{-ck}$ for some $c>0$. When β_k decays faster so that $\frac{\beta_k^2}{\sum_{i=k+1}^{\infty}\beta_i^2}$ is bounded away from zero and $\sup_{a_n\le k\le b_n}|\beta_k|=o\left(\frac{\sqrt{\log(n)}}{n^{1/4}}\right)$, any choice of $d>0$ makes Condition (N3) satisfied. An example is the exponential-decay case, i.e., $\beta_k=e^{-ck}$ for some $c>0$. According to [38], when $\hat{k}_n$ is selected by BIC for order selection, we have that $r_{\hat{k}_n}$ basically falls within a constant from $\frac{1}{2c}\log(n/\log(n))$ in probability. In this case, $\beta_k\approx\frac{\sqrt{\log(n)}}{n^{1/2}}$ for $k\approx\frac{1}{2c}\log(n/\log(n))$. Thus Condition (N3) is satisfied.

3.5. *On the choice of λ_n and d.* A natural choice of (λ_n,d) is $\lambda_n=1$ and $d=0$, which is expected to work well to distinguish parametric and nonparametric scenarios that are not too close to each other for order selection or all subset selection with p_n increasing not fast in n. Other choices can handle more difficult situations, mostly entailing the satisfaction of (N2) and (N3). With a larger λ_n or d, PI tends to be closer to 1 for a nonparametric case, but at the same time, it makes a parametric case less obvious. When there are many models being considered, λ_n should not be too small so as to avoid severe selection bias. The choice of $d>0$ handles fast decay of the approximation error in nonparametric scenarios, as mentioned already.

3.6. *Combining strengths of AIC and BIC.* From above, for any given cutoff point bigger than 1, the PI in a parametric scenario will eventually exceed it while the PI in a nonparametric scenario will eventually drops below it when the sample size gets large enough.

It is well-known that AIC is asymptotically loss (or risk) efficient for nonparametric scenarios and BIC is consistent when there are finite-dimensional correct models (see [17] for a recent result), which implies that BIC is asymptotically loss efficient [51].

COROLLARY 2. *For a given number $c>1$, let δ be the model selection*

procedure that chooses either the model selected by AIC or BIC as follows:

$$\delta = \begin{cases} AIC & \text{if } PI < c \\ BIC & \text{if } PI \geq c. \end{cases}$$

Under Conditions P1-P3, N1-N3, δ is asymptotically loss efficient in both parametric and nonparametric scenarios as long as AIC and BIC are loss efficient for the respective scenarios.

Remarks:

1. In application, for a finite sample, a good cutoff point c needs to be chosen. We will explore this in numerical examples.
2. Previous work on sharing the strengths of AIC and BIC utilized minimum description length criterion in an adaptive fashion ([7, 33]), or flexible priors in a Bayesian framework ([30, 25]). To our knowledge, only Ing [38] and Yang [65] established (independently) simultaneous asymptotic efficiency for both parametric and nonparametric scenarios. Differently from our work here, their main idea to separate parametric and nonparametric scenarios is to compare the models selected by BIC at different sample sizes. In contrast, our method in this paper compares neighboring models at the full sample size.

4. PI as a model selection diagnostic measure, i.e., *P*ractical *I*dentifiability of the best model. Based on the theory presented in the previous section, it is natural to use the simple rule for answering the question if we are in a parametric or non-parametric scenario: call it parametric if PI is larger than c for some $c > 1$ and otherwise nonparametric. Theoretically speaking, we will be right with probability going to one.

Keeping in mind that the concepts such as parametric, nonparametric, consistency and asymptotic efficiency are all mathematical abstractions that hopefully characterize the nature of the data and the behaviors of estimators, our intended use of PI is not a rigid one so as to be practically relevant and informative, as we explain below.

Both parametric and nonparametric methods have been widely used in statistical applications. One specific approach to nonparametric estimation is to use parametric models as approximations to an infinite-dimensional function, which is backed up by approximation theories. However, it is in this case that the boundary between parametric and nonparametric estimations becomes blurred, and our work tries to address the issue.

From a theoretical perspective, the difference between parametric and nonparametric modeling is quite clear in this context. Indeed, when one is

willing to assume that the data come from a member in a parametric family, the focus is then naturally on the estimation of the parameters, and finite-sample and large sample properties (such as UMVUE, BLUE, minimax, Bayes, and asymptotic efficiency) are well understood. For nonparametric estimation, given infinite-dimensional smooth function classes, various approximation systems (such as polynomial, trigonometric and wavelets) have been shown to lead to minimax-rate optimal estimators via various statistical methods (e.g., [9, 23, 37, 57]). In addition, given a function class defined in terms of approximation error decay behavior by an approximating system (or smoothness of the function), rates of convergence of minimax risks have been established (see, e.g., [67]). As is expected, the optimal model size (in rate) based on linear approximation depends on the sample size (and other things) for a nonparametric scenario. In particular, for full and sparse approximation sets of functions, the minimax theory shows that for a typical nonparametric scenario, the optimal model size makes the approximation error (squared bias) roughly equal to estimation error (model dimension over the sample size) [67]. Furthermore, adaptive estimators that are simultaneously optimal for multiple function classes can be obtained by model selection or model combining (see, e.g, [5, 62] for many references).

From a practical perspective, unfortunately, things are much less clear. Consider, for example, the simple case of polynomial regression. In linear regression textbooks, one often finds data that show obvious linear or quadratic behavior, in which case perhaps most statisticians would be unequivocally happy with a linear or quadratic model (think of Hooke's law for describing elasticity). When the underlying regression function is much more complicated so as to require 4th or 5th power, it becomes difficult to classify the situation as parametric or nonparametric. While few (if any) statisticians would challenge the notion that in both cases, the model is only an approximation to reality, what makes the difference in calling one case parametric quite comfortably but not the other? Perhaps simplicity and stability of the model play key roles as mentioned in Cox [20]. Roughly speaking, when a model is simple and fits the data excellently (e.g, with R^2 close to 1) so that there is little room to significantly improve the fit, the model obviously stands out. In contrast, if we have to use a 10th order polynomial to be able to fit the data with 100 observations, perhaps few would call it a parametric scenario. Most of the situations may be in between.

Differently from the order selection problem, the case of subset selection in regression is substantially more complicated due to the much increased complexity of the list of models. It seems to us that when all subset regression is performed, it is usually automatically treated as a parametric problem in

the literature. While this is not surprising, our view is different. When the number of variables is not very small relative to the sample size and the error variance, the issue of model selection does not seem to be too different from order selection for polynomial regression where a high polynomial power is needed. In our view, when analyzing data (in contrast to asymptotic analysis), if one explores over a number of parametric models, it is not necessarily proper to treat the situation as a parametric one, by which we mean the standard practice of reporting the standard errors and confidence intervals for parameters and making interpretations based on the selected model.

Closely related to the above discussion is the issue of model selection uncertainty (see, e.g., [12, 16]). It is now well recognized that model selection uncertainty should not be conveniently ignored as is still the dominating practice in real world statistical applications. It is an important issue to know when we are in a situation where a relatively simple and reliable model stands out in a proper sense and thus can be used as the "true" model for practical purposes, and when a selected model is just one out of multiple or even many possibilities among the candidates at the given sample size. In the first case, we would be willing to call it parametric (or more formally, practically parametric) and the latter (practically) nonparametric.

We should emphasize that in our review, our goal is not exactly finding out whether the underlying model is finite-dimensional (relative to the list of candidate models) or not. Indeed, we will not be unhappy to declare a truly parametric scenario nonparametric when around the current sample size no model selection criterion can possibly identify it with confidence and then take advantage of it, in which case, it seems better to view the models as approximations to the true one and we are just making a tradeoff between the approximation error and estimation error. In contrast, we will not be shy to continue calling a truly nonparametric model parametric should we be given that knowledge by an oracle if one model stands out at the current sample size and the contribution of the ignored features is so small that it is clearly better to be ignored at the time being. When the sample size is much increased, the enhanced information allows the discovering of the relevance of some additional features and then we may be in the practical nonparametric scenario. As the sample size further increases, it may well be that a parametric model stands out until reaching a larger sample size where we enter practical nonparametric scenario again, and so on. This point will be illustrated in the next section.

Based on hypothesis testing theories, obviously, at a given sample size, for any true parametric distribution in one of the candidate families from

which the data are generated, one has a nonparametric distribution (i.e., not in any of the candidate families) that cannot be distinguished from the true distribution. From this perspective, pursuing a rigid finite-sample distinction between parametric and nonparametric scenarios is improper.

PI is relative to the list of candidate models and the sample size. So it is perfectly possible (and fine) that for one list of models, we declare the situation to be parametric, but for a different choice of candidate list, we declare nonparametriness.

To summarize, for application,

1. We adopt a pragmatic view on the contrast between parametric and nonparametric scenarios when a number of parametric models are considered to analyze the data. In our view, when a candidate model stably stands out in proper measures, we can call it practically parametric, and in contrast, when the selected model is behaving similarly to some alternative models among the candidates and it is merely providing a sample-size sensitive balancing of the approximation and estimation error, we call it nonparametric to indicate the lack of an outstanding candidate.
2. Note that our assessment of practical parametricness/nonparamreticness depend only on data and the candidate models, but not on assumptions on the true model.
3. PI can be viewed as a model selection diagnostic measure (or practical identifiability of the best model), which tells us, roughly, for the time being and foreseeable future, if the selected model deserves to be called a parametric model based on which parameter estimates and standard errors are to be reported. When PI is close to 1, it indicates that it is not OK to treat the selected model as the "truth", and interpretations based on the selected model, no matter how convenient, should not be treated too seriously.

5. Simulation Results. In this section, we consider single-predictor and multiple-predictor cases, aiming at a serious understanding of the practical utility of PI. In all the numerical examples in this paper, we choose $\lambda_n = 1$ and $d = 0$. Note that the number of predictors is not very large relative to the sample size (see Section 3.5).

5.1. *Single predictor.*

Example 1. Compare two different situations:

Case 1: $Y = 3\sin(2\pi x) + \sigma_1\epsilon$,

Case 2: $Y = 3 - 5x + 2x^2 + 1.5x^3 + 0.8x^4 + \sigma_2\epsilon$, where $\epsilon \sim N(0, 1)$ and $x \sim N(0, 1)$.

BIC is used to select the order of polynomial regression between 1 and 30. The estimated σ from the selected model is used to calculate the PI. Representative scatterplots at $n = 200$ with $\sigma_1 = 3$, $\sigma_2 = 7$ can be found in Figure 1. Note that the function estimate based on the selected model by BIC is visually more different from that based on the smaller model with one fewer term for the parametric scenario than the nonparametric one.

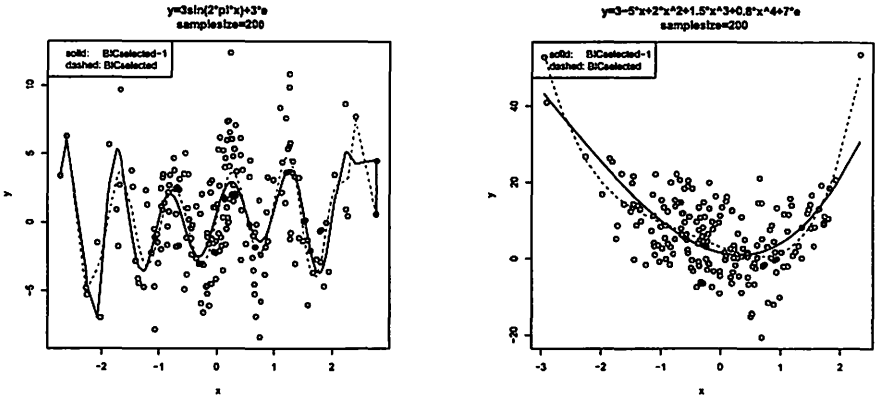

FIG 1. *Scatterplots For Example 1*

Quantiles for the PIs in both scenarios based on 300 replications are presented in Table 1.

TABLE 1
Percentiles of PI for Example 1

	case 1			case 2		
percentile	order selected	PI	$\hat{\sigma}$	order selected	PI	$\hat{\sigma}$
10%	1	0.47	2.78	4	1.14	6.53
20%	13	1.02	2.89	4	1.35	6.67
50%	15	1.12	3.03	4	1.89	6.96
80%	16	1.34	3.21	4	3.15	7.31
90%	17	1.54	3.52	4	4.21	7.49

Example 2. Compare the following two situations:

Case 1: $Y = 1 - 2x + 1.6x^2 + 0.5x^3 + 3\sin(2\pi x) + \sigma\epsilon$

Case 2: $Y = 1 - 2x + 1.6x^2 + 0.5x^3 + \sin(2\pi x) + \sigma\epsilon$.

The two mean functions are the same except the coefficient of the $\sin(2\pi x)$ term. Scatterplots and table similar to those for Example 1 are in Figure 2 and Table 2, respectively. As we can see from Table 2, although both cases are of a nonparametric nature, they have different behaviors in terms of model selection uncertainty and PI values. Case 2 can be called 'practically' parametric and the large PI values provide information in this regard.

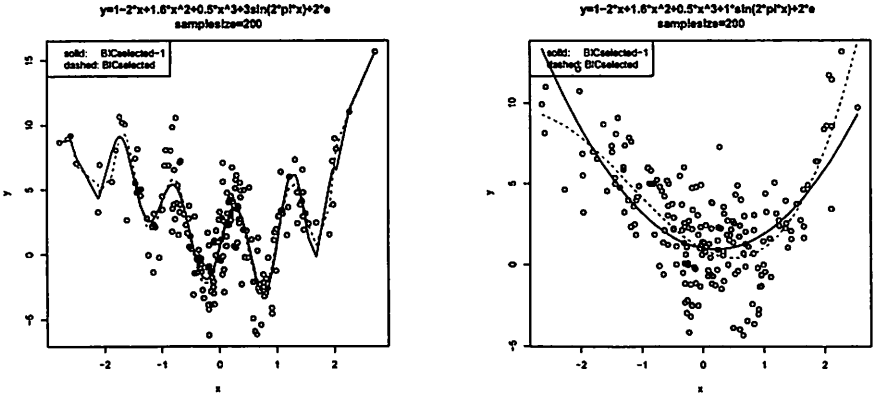

FIG 2. *Scatterplots For Example 2*

TABLE 2
Percentiles of PI for Example 2

	case 1			case 2		
percentile	order selected	PI	$\hat{\sigma}$	order selected	PI	$\hat{\sigma}$
10%	15	1.01	1.87	3	1.75	1.99
20%	15	1.05	1.92	3	2.25	2.03
50%	16	1.14	2.00	3	3.51	2.12
80%	17	1.4	2.11	3	5.33	2.22
90%	18	1.63	2.17	3	6.62	2.26

5.2. *Factors that influence PI.* As we know, most model selection problems, if not all, are affected by many factors like the regression function itself, the noise level, and the sample size. We expect that these factors influence the behavior of PI as well. We investigate the effects of these factors on PI and report some representative results below. As we will see, PI, as a

diagnostic measure, can tell us whether a problem is 'practically' parametric/nonparametric due to the influences of all the factors that affect model selection.

5.2.1. *The effect of sample size.* We calculated the PI at different sample sizes with 300 replications for each and report the results for Examples 1 and 2 in Figures 3 and 4.

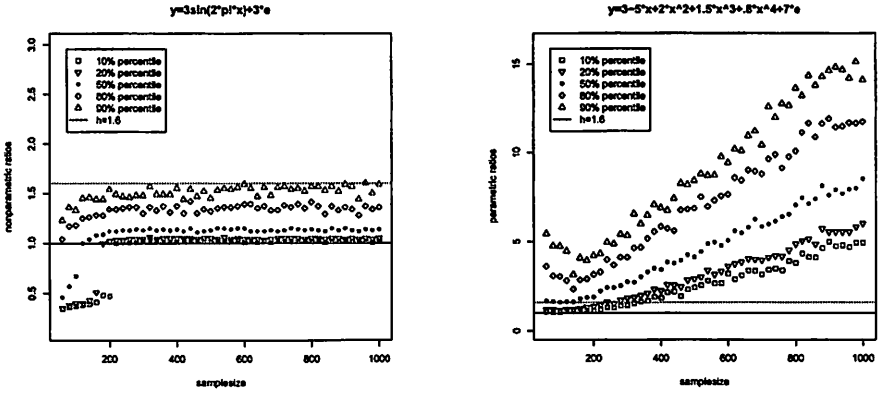

FIG 3. *Sample size effect for Example 1*

For Example 1, from Figure 3, we see the PIs in case 1 basically fall in between 1 and 1.6, whereas the PIs in case 2 become larger as sample size increases.

From Figure 4, in case 2 of Example 2, the PIs first increase and then drop down as the sample size increases. This is due to the fact that in the beginning, the sine term is better to be ignored due to lack of information, and when the sample size is bigger, say 300-400, the PI indicates a strong parametric scenario. With a sample size in this range, the problem is 'practically' parametric. With more and more data we are then gradually able to detect the signal of the $\sin(2\pi x)$ term, thus capturing the nonparametric nature of the mean function.

In case 1 of Example 2 we have the 90% percentiles slightly exceeding 1.6. Also notice that the percentiles in case 2 drop at different levels of sample sizes. For example, the 10% drops below 1.6 when the sample size is bigger than 400, while the 50% drops below 1.6 when the sample size is bigger than 800.

The examples show that given the regression function and the noise

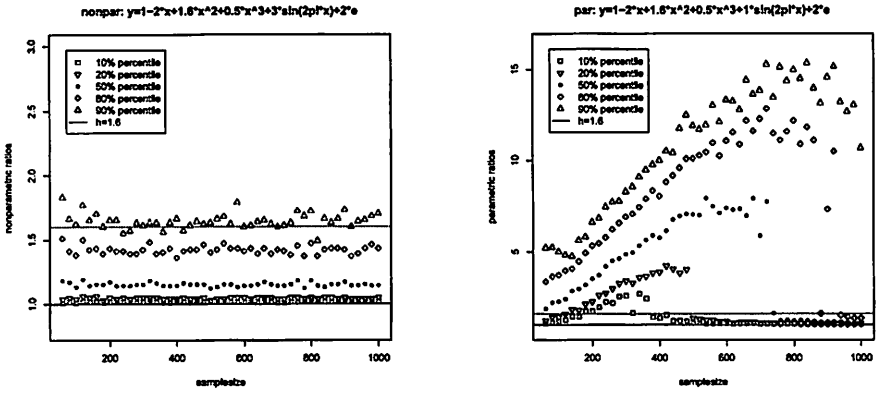

FIG 4. *Sample size effect for Example 2*

level, the value of PI indicates whether the problem is 'practically' parametric/nonparametric at the current sample size.

5.2.2. *The effect of coefficient.* We study the PIs for different values of the coefficient of the last term in case 2 of Example 1 and Example 2, respectively. The results are reported in Figure 5.

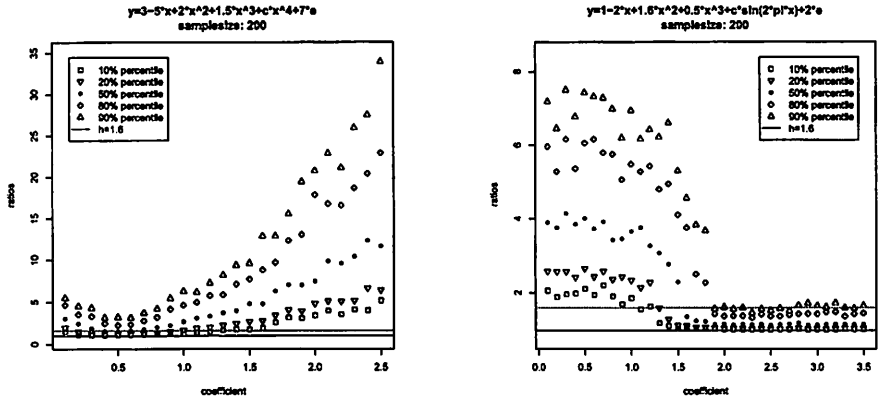

FIG 5. *Effect of coefficient*

For Example 1, the values of PI first decrease and then increase as the coefficient of x^4 increases. This is because when the coefficient for the term

of x^4 is small (less than .5), the true mean function behaves just like a polynomial of order 3 at the current sample size. As the coefficient gets slightly larger, there is no clear distinction between a polynomial of order 3 and a polynomial of order 4 at the current sample size. That is why we see the PIs drop a little in the beginning. However, when the coefficient gets bigger than .5 or .6, then we can detect the term of x^4 and the PIs increase with the coefficient. Overall, the PI values are mostly larger than 1.6 in this example.

For Example 2, the PIs drop as the coefficient of $\sin(2\pi x)$ increases. This is because as the coefficient gets larger, the nonparametric signal becomes stronger. When the coefficient is small (less than 1.3), most of the PIs are bigger than 1.6 and the problem is 'practically' parametric. When the coefficient is bigger than 1.9, most of the PIs fall in between 1 and 1.6 and the problem is 'practically' nonparametric.

The examples show that given the noise level and the sample size, when the nonparametric part is very weak, PI has a large value, which properly indicates that the nonparametric part is negligible; but as the nonparametric part gets strong enough, PI will drop close to 1, indicating a clear nonparametric scenario. For a parametric scenario, the stronger the signal, the larger PI as is expected.

5.3. *Multiple predictors.* Now we study several examples with multiple predictors. The first two examples were used in the original lasso paper [59].

Unlike what we did in the single predictor cases, in these multiple-predictor examples we are going to do all subset selection. We generate data from a linear model (except example 7):

$$Y = \beta^T \mathbf{x} + \sigma\epsilon,$$

where $\mathbf{x}$ is generated from a multivariate normal distribution with mean 0, variance 1, and correlation structure given in each example. For each generated data set, we apply the Branch and Bound algorithm [32] to do all subset selection by BIC and then calculate the PI value (part of our code is modified from the aster package of Geyer [31]). Unless otherwise stated, in these examples, the sample size is 200 and we replicate 300 times.

Example 3. In this example, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. We set $\sigma = 5$.

Example 4. This example is the same as example 3, but with $\beta_j = .85, \forall j$ and $\sigma = 3$.

Example 5. In this example, $\beta = (0.9, 0.9, 0, 0, 2, 0, 0, 1.6, 2.2, 0, 0, 0, 0)^T$. There are 13 predictors and the pairwise correlation between x_i and x_j is $\rho = 0.6$ and $\sigma = 3$.

Example 6. This example is the same as example 5 except that $\beta = (0.85, 0.85, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0)^T$ and $\rho = 0.5$.

Example 7. This example is the same as example 3 except that we add a nonlinear component in the mean function and $\sigma = 3$, i.e., $Y = \beta^T \mathbf{x} + \phi(u) + \sigma\epsilon$, where $u \sim uniform(-4, 4)$ and $\phi(u) = 3(1 - 0.5u + 2u^2)e^{-u^2/4}$. All subset selection is carried out with predictors $x_1, \cdots, x_8, u, \cdots, u^8$ which are coded as 1-8 and A-G in the Table 3.

TABLE 3
Proportion of selecting true model

Example	true model	proportion
3	125	0.82
4	12345678	0.12
5	12589	0.43
6	125	0.51
7	1259ABCEG*	0.21

TABLE 4
Quartiles of PIs

example	Q1	Q2	Q3
3	1.26	1.51	1.81
4	1.02	1.05	1.10
5	1.05	1.15	1.35
6	1.09	1.23	1.56
7	1.02	1.07	1.16

The selection behaviors and PI values are reported in Table 3 and Table 4, respectively. From those results, we see that the PIs are large for Example 3 and small for Example 4. Note that in Example 3 we have 82% chance selecting the true model, while in Example 4 the chance is only 12%. Although both Example 3 and Example 4 are of parametric nature, we would call Example 4 'practically nonparametric' in the sense that at the given sample size many models are equally likely and the issue is to balance the approximation error and estimation error. For Examples 5 and 6, the PI values are in-between, so are the chances of selecting the true models. Note that the median PI values in Examples 5 and 6 are around 1.2. These examples together show that the values of PI provide sensible information on how strong the parametric message is and that information is consistent with stability in selection. More discussions about these examples in terms of PI and statistical risks will be given later in this section. (In the lasso paper σ was chosen to be 3 for Example 3. But even with a higher noise level $\sigma = 5$, the parametric nature of this example is still obvious.)

Example 7 is quite interesting. Previously, without the $\phi(u)$ component, even at $\sigma = 5$, large values of PI are seen. Now with the nonparametric component present, the PI values are close to 1. (The asterisk mark (*) in Table 3 indicates the model is the most frequently selected one instead of being the true model.)

An illuminating example. We now look at a special example. We still generate data from a linear model with $\beta = (2, 2, 0.3, 0.3, 0.1, 0.1, 0, 0, 0, 0)^T$ and $\sigma = 2$. The pairwise correlation among the predictors is 0.5. For this example we do all-subset selection by BIC at different sample sizes. Our thinking is that since some of the coefficients are large and others are small, BIC is going to pick up the significant predictors gradually as the sample size increases. We expected to see both big and small PI values alternating to some degree when the sample size changes. In this example, we replicate 500 times for each sample size.

The results of median PIs at different sample sizes are shown in figure 6. From the plot we see PI first increases with the sample size, then decreases, then increases and decreases again, and finally increases. This is because when the sample size is small, most of the time BIC only picks up x_1 and x_2 and the PI increases with the sample size. As the sample size further increases, BIC finds the predictors x_3 and x_4 relevant and the PI then decreases since the coefficients for x_3 and x_4 are small (but not too small) so that BIC is not quite sure about the best model. When the sample size gets big enough so that most of the times BIC chooses x_1, x_2, x_3, and x_4, the PI increases again with sample size. A similar story repeats for the predictors x_5, and x_6. If we choose 1.2 as a cutoff point, we would see (practically) parametric and (practically) nonparametric scenarios alternating as the sample size changes.

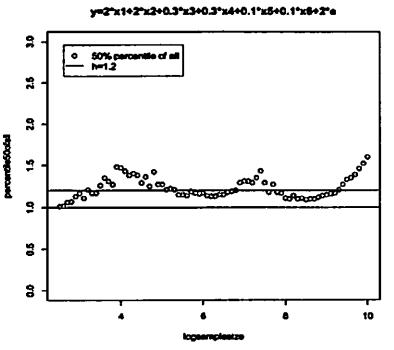

FIG 6. *Behavior of PI for the special example*

Inference after model selection (PI as practical identifiability). With Examples 3 and 4, we assess the accuracy of statistical inferences after model selection. We first generate an evaluation set of predictors from the same distribution as that for observations. Then for each replication, we generate data and do subset selection by BIC. After selecting a model, we get the resulting predicted value (estimated regression function at the given point), the standard error and the 95% confidence interval for the regression estimate at each of the new design points in the evaluation set. We replicate 500 times. Then at each new design point we calculate the actual standard

deviation (which is called se.fit.t in the output table) of the 500 regression estimates and compare that to quantiles of the 500 standard errors that are obtained based on the selected model. We also take a look at the actual coverage of the 95% CIs for the true mean of the regression function at each new design point. Results at 10 randomly chosen new design points are reported in Table 5 and Table 6 for the two examples.

TABLE 5
Reliability of inference for Example 3

	quantiles of standard errors of fit						
new design point	5%	25%	50%	75%	95%	se.fit.t	coverage
1	0.439	0.474	0.496	0.525	0.583	0.564	0.930
2	0.351	0.375	0.396	0.416	0.480	0.457	0.934
3	0.333	0.356	0.375	0.395	0.446	0.471	0.932
4	0.359	0.386	0.405	0.428	0.464	0.453	0.932
5	0.360	0.385	0.405	0.426	0.493	0.498	0.926
6	0.229	0.247	0.258	0.271	0.349	0.350	0.920
7	0.391	0.420	0.443	0.468	0.516	0.582	0.906
8	0.398	0.426	0.447	0.473	0.509	0.502	0.926
9	0.631	0.679	0.712	0.747	0.809	0.738	0.960
10	0.238	0.254	0.265	0.277	0.296	0.252	0.972

TABLE 6
Reliability of inference for Example 4

	quantiles of standard errors of fit						
new design point	5%	25%	50%	75%	95%	se.fit.t	coverage
1	0.448	0.551	0.637	0.716	0.795	1.340	0.626
2	0.368	0.537	0.667	0.721	0.783	1.250	0.656
3	0.727	0.920	1.097	1.200	1.304	2.122	0.660
4	0.298	0.458	0.516	0.551	0.600	0.941	0.662
5	0.618	0.728	0.798	0.856	0.946	1.365	0.758
6	0.353	0.411	0.438	0.463	0.501	0.654	0.796
7	0.543	0.683	0.773	0.830	0.914	1.471	0.672
8	0.393	0.457	0.507	0.560	0.610	0.991	0.684
9	0.537	0.624	0.676	0.727	0.795	1.130	0.740
10	0.566	0.688	0.786	0.867	0.959	1.720	0.634

From the results we can see the actual coverages in Example 3 are reasonably close to 95% while the ones in Example 4 are much worse than the nominal 95% level. Also the (simulated) actual standard errors of the regression estimation are quite close to the ones from the selected model in Example 3, and in contrast, in Example 4, the reported uncertainty of regression estimation is grossly under-estimated. We tried several evaluation data sets with different sizes, the results are similar.

It is now well known that model selection has an impact on subsequent statistical inferences (see, e.g., [69, 35, 27, 40]). For observational data, typically one cannot avoid making various modeling choices (such as which type of statistical analysis to pursue, which kind of models to consider) after seeing the data, but their effects are very difficult to quantify. Thus it can be very helpful to know when the choices have limited impact on the final results. The above results together with Tables 3 and 4 show that the value of PI can provide valuable information on the parametricness of the underlying regression function and hence on how confident we are on the accuracy of subsequent inferences.

Combining strengths of AIC and BIC based on PI. Still with Examples 3-7, we investigate the performance of an adaptive choice between AIC and BIC based on the PI value. Again we first generate an evaluation data set with 500 new design points from the same distribution as the one for observations. Then for each replication, we use both AIC and BIC to select a model. The combined procedure is BIC if the PI value is larger than a cutoff point (chosen as 1.2 in these examples) and AIC otherwise. Then for each procedure (AIC, BIC, and the combined) in each replication, we calculate the average squared error (which is the average squared difference between the true regression mean and the fitted value based on the selected model) at the new design points in the evaluation data. We replicate 500 times and the statistical risk is estimated to be the average of the 500 average squared errors. The risk ratios are reported in Table 7 with BIC as the reference.

TABLE 7
Statistical risks of AIC, BIC, and the Combined procedure

	Statistical Risk			Risk Ratio		
Example	AIC	BIC	Combined	AIC	BIC	Combined
3	0.335	0.227	0.230	1.474	1.000	1.014
4	0.543	1.045	0.680	0.520	1.000	0.651
5	0.562	0.513	0.564	1.096	1.000	1.098
6	0.502	0.402	0.459	1.250	1.000	1.142
7	0.835	0.927	0.899	0.901	1.000	0.969

From the results we see in all these examples the combined procedure shows capability of adaptation between AIC and BIC in terms of the statistical risk. We also see from Tables 7, 3 and 4 that in Examples 5 and 7, the PIs are roughly around 1.2 and AIC and BIC have similar performance in terms of statistical risks, while in the other examples the PIs are either large or small and correspondingly, either BIC or AIC has a smaller statistical risk. These results show that PI provides helpful information regarding

whether AIC or BIC works better or they have similar performances in statistical risks. Therefore, PI can be viewed as a **P**erformance **I**ndicator of AIC versus BIC.

5.4. *A summary.* From our simulation outcomes (some are not presented due to space limitation), we summarize a few points here.

1. Factors other than the nature of the regression function also influence the value of PI, including the sample size and the noise level. From a practical point of view, PI, as a diagnostic measure, indicates whether a specific problem is 'practically' parametric/nonparametric with the influences of all those factors.
2. Model selection effect on subsequent statistical inferences may or may not be reasonably ignored, and the value of PI provides useful information in that regard.
3. For Examples 3 and 4, both being parametric, one is practically parametric and the other practically nonparametric for $n = 200$. Correspondingly, BIC works better for the former and AIC for the latter in terms of risk for estimating the regression function. This phenomenon will be seen again in the next section.
4. Combining AIC and BIC based on the PI value shows adaptation capability in terms of statistical risk. That is, the composite rule yields a risk close to the better one of AIC and BIC.
5. In nested model problems (like order selection of series expansion), a cutoff point of $c = 1.6$ seems to be good. In subset selection problems, we expect the cutoff point to be smaller since the infimum is taken over many models. The choice of 1.2 seems to be reasonably good based on our numerical investigations, which is also supported by the observation that when PI is around 1.2, AIC and BIC perform similarly.

6. Real Data Examples. In this section, we study three data sets: the Ozone data (e.g. [11]), the Boston housing data (e.g. [34]), and the Diabetes data (e.g. [24]).

In these examples, we conduct all subset selection by BIC using the Branch and Bound algorithm. Besides finding the PI values for the full data, we also do the same with sub-samples from the original data at different sample sizes. In addition, we carry out a parametric bootstrap from the model selected by BIC based on the original data to assess the stability of model selection. (The design points of the predictors are randomly selected with replacement from the original data.) Like in the multiple-predictor sim-

ulation study, we also combine AIC and BIC based on the PI value when doing parametric bootstrap. Unless otherwise stated, the subsampling and the bootstrap are both replicated 500 times at each sample size. (In the results, the predictors are coded to be a single digit between 1 and 9 and then a single capital letter between 'A' and 'Z', i.e, letter 'A' stands for the 10th predictor, 'B' for the 11th, and so on.)

Ozone Data. There are 9 variables with 8 predictors and 330 observations. We followed the transformations of the predictors and the response suggested by Hawkins [36]. (In that paper a ninth predictor, day of the year, was also included. We left this predictor out as many others did. See [11].) After the transformations, we have 10 predictors with quadratic terms of two predictors added.

Boston Housing Data. The data consists of 14 variables (1 response and 13 predictors). There are 506 observations. We followed the transformations of the variables in Harrison and Rubinfeld's paper [34].

Diabetes Data. There are 11 variables with 10 predictors and 442 observations.

The PIs from the original data for these three examples are: 1.277 (ozone), 1.028 (Boston housing), and 1.298 (diabetes). The results of subsampling and bootstrap are reported in Tables 8-9 and Tables 10-11, respectively.

TABLE 8
Quartiles of PIs from subsamples of size 400

Data	Q1	Q2	Q3
Ozone	-	-	-
Boston	1.02	1.04	1.1
Diabetes	1.17	1.23	1.28

TABLE 9
Quartiles of PIs from subsamples of size 200

Data	Q1	Q2	Q3
Ozone	1.08	1.21	1.47
Boston	1.02	1.05	1.11
Diabetes	1.06	1.13	1.24

From the tables, we see the PIs for the ozone data are mostly larger than 1.2, while those for the Boston housing data are smaller than 1.2. Moreover, the parametric bootstrap suggests that for the Ozone data, the model selected from the full data still reasonably stands out even when the sample size is reduced to about 200 and noises are added (not all shown due to space limitation). For the Boston housing data, however, even at a sample size of 400 we only have 28% chance selecting the same model as the one selected with the full data. Interestingly, the diabetes data exhibit a parametric behavior when $n = 400$, but with the sample size reduced by half, it looks more like a nonparametric scenario.

TABLE 10
The 6 most frequently selected models and their frequencies with a sample size of 400

	Ozone		Boston Housing		Diabetes	
	model	proportion	model	proportion	model	proportion
1	-	-	145689ABCD	0.28	23479	0.732
2	-	-	15689ABCD	0.238	3479	0.078
3	-	-	15689ABD	0.092	349	0.058
4	-	-	145689ABD	0.084	23489	0.016
5	-	-	145689BCD	0.062	2349	0.012
6	-	-	14568BCD	0.046	3459	0.012

TABLE 11
The 6 most frequently selected models and their frequencies with a sample size of 200

	Ozone		Boston Housing		Diabetes	
	model	proportion	model	proportion	model	proportion
1	1269	0.474	15689ABCD	0.088	23479	0.318
2	126	0.248	15689ABD	0.088	349	0.17
3	1236	0.06	1568BD	0.07	39	0.128
4	1239	0.046	1589ABD	0.062	3479	0.102
5	167	0.028	14568BD	0.05	2349	0.042
6	12	0.012	1568BCD	0.044	379	0.042

TABLE 12
Combining AIC and BIC based on PI with full sample size

	Statistical Risk			Risk Ratio		
Data	AIC	BIC	Combined	AIC	BIC	Combined
Ozone	7.66e-4	6.44e-4	6.82e-4	1.189	1.000	1.060
Boston Housing	8.18e-4	1.05e-3	8.65e-4	0.779	1.000	0.824
Diabetes	63.05	57.42	58.19	1.098	1.000	1.014

Combining AIC and BIC based on PI. Similar to the simulation results in Section 5, by parametric bootstrap at the original sample size from the selected model, in these data examples, combining AIC and BIC based on PI shows good overall performance in terms of statistical risk (Table 12). The combined procedure has a statistical risk close to the better one of AIC and BIC in each case.

7. Conclusions. Parametric models have been commonly used to estimate a finite-dimensional or infinite-dimensional function. While there have been serious debates on which model selection criterion to use to choose a candidate and there has been some work on combining the strengths of very distinct model selection methods, there is a major lack of understanding on statistically distinguishing between scenarios that favor one method (say

AIC) and those that favor another (say BIC). To address this issue, we have derived a parametricness index (PI) that has the desired theoretical property: PI converges in probability to infinity for parametric scenarios and to 1 for nonparametric ones. The use of a consistent model selection rule in constructing PI effectively prevents overfitting when we are in a parametric scenario. The comparison of the selected model with a subset model separates parametric and nonparametric scenarios through the distinct behaviors of the approximation errors of these models in the two different situations.

One interesting consequence of the property of PI is that a choice between AIC and BIC based on its value ensures that the resulting regression estimator is automatically asymptotically efficient for both parametric and nonparametric scenarios, which clearly cannot be achieved by any deterministic choice of the penalty parameter in the criteria of the form $-\log\text{-likelihood} + \lambda m_k$, where m_k is the number of parameters in the model k. Thus an adaptive regression estimation to simultaneously suit parametric and nonparametric scenarios is realized through the information provided by PI.

We advocate a practical view on parametricness/nonparametricness. In our view, a parametric scenario is one where a relatively parsimonious model reasonably stands out. Otherwise, the selected model is most likely a tentative compromise between goodness of fit and model complexity, and the recommended model is most likely to change when the sample size is slightly increased. Our simulation and data examples suggest that for a practically parametric scenario, BIC tends to perform better, but for a practically nonparametric scenario, AIC does so in estimation.

Our numerical results seem to be very encouraging. PI is informative, giving the statistical user an idea on how much one can trust the selected model as the "true" one. When PI does not support the selected model as the "right" parametric model for the data, we have demonstrated that estimation standard errors reported from the selected model are often too small compared to the real ones, that the coverages of the resulting confidence intervals are much smaller than the nominal levels, and that mode selection uncertainty is high. In contrast, when PI strongly endorses the selected model, model selection uncertainty is much less a concern and the resulting estimates and interpretation are trustworthy to a large extent.

Identifying a stable and strong message in data as is expressed by a meaningful parametric model, if existing, is obviously important. In biological and social sciences, especially observational studies, a strikingly reliable parametric model is often too much to ask for. Thus, to us, separating scenarios where one model is reasonably standing out and is expected to shine over

other models for sample sizes not too much larger than the current one from those where the selected model is simply the lucky one to be chosen among multiple equally performing candidates is an important step beyond simply choosing a model based on one's favorite selection rule or, in the opposite direction, not trusting any post model selection interpretation due to existence of model selection uncertainty.

For the other goal of regression function estimation, in application, one typically applies a model selection method, or considers estimates from two (or more) model selection methods to see if they agree with each other. In light of PI (or similar model selection diagnostic measures), the situation can be much improved: one adaptively applies the better model selection criterion to improve estimation/prediction performance. We have focused on the competition between AIC and BIC, but similar measures may be constructed for comparing other model selection methods that are derived from different principles or under different assumptions.

It has been suggested that AIC performs better for a nonparametric scenario and BIC better for a parametric one (see [65] for a study on the issue in a simple setting). This is asymptotically justified but certainly not quite true in reality. Our numerical results have demonstrated that for some parametric regression functions, AIC is much better. On the other hand, for an infinite-dimensional regression function, BIC can give a much more accurate estimate. Regarding this discrepancy between asymptotics and finite-sample reality, one typically explains that of course finite-sample behaviors can be totally different from asymptotic ones. Our numerical results tend to suggest that a much more helpful statement is: when PI is high and thus we are in a practical parametric scenario (whether the true regression function is finite-dimensional or not), BIC tends to be better for regression estimation; when PI is close to 1 and thus we are in a practical nonparametric scenario, AIC tends to be better. We feel that the use of PI as a suitable indicator of the relative performance of AIC and BIC is a positive step forward towards a data-driven sound choice of a model selection method.

Finally, we point out some limitations of our work. First, our results address only linear models under Gaussian errors. Second, more understanding on the choices of λ_n, d, and the best cutoff value c for PI is needed. Although the choices recommended in this paper worked very well for the numerical examples we have studied, different values may be proper for other situations (e.g., when the predictors are highly correlated and/or the number of predictors is comparable to the sample size).

APPENDIX

The following two facts will be used in our proofs (see [61]).

Fact 1. If $Z \sim N(0, 1)$, then $P(|Z| \geq t) \leq e^{-t^2/2}$, $\forall t > 0$.

Fact 2. If $Z_m \sim \chi^2_m$, then

$$P(Z_m - m \geq \kappa m) \leq e^{-\frac{m}{2}(\kappa - \ln(1+\kappa))}, \qquad \forall \kappa > 0.$$
$$P(Z_m - m \leq -\kappa m) \leq e^{-\frac{m}{2}(-\kappa - \ln(1-\kappa))}, \qquad \forall 0 < \kappa < 1.$$

Before the proofs, let us look at the relationship between the projection matrices, $M_{k^{(s)}}$ and M_k, of two nested models as following.

$$\text{Model } k^{(s)}: \quad Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + \epsilon,$$

$$\text{Model } k: \quad Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + \beta_k x_k + \epsilon.$$

Fact 3. Denote $\alpha = (x_{1,k}, \cdots, x_{n,k})^T$, $\alpha_{New} = (I_n - M_{k^{(s)}})\alpha$ and $P_{k^{(s)},k} = M_k - M_{k^{(s)}}$, then we have

$$P_{k^{(s)},k} = \alpha_{New} \cdot \alpha_{New}^T / \|\alpha_{New}\|^2. \tag{7.1}$$

Note that the space that α_{New} spans is the orthogonal complement of the space spanned by the columns of X_{k-1} in the whole space spanned by the columns of X_k, and $P_{k^{(s)},k}$, with rank 1, is the projection matrix onto the vector of α_{New}.

For the ease of notation, we denote $P_{k^{(s)},k}$ by P, $rem_1(k) = e_n^T(f_n - M_k f_n)$, and $rem_2(k) = \|(I_n - M_k)e_n\|^2/\sigma^2 - n$ in the proofs . Then

$$\|(I_n - M_{k^{(s)}})e_n\|^2 = \|(I_n - M_k)e_n\|^2 + \|Pe_n\|^2 \tag{7.2}$$
$$\|(I_n - M_{k^{(s)}})f_n\|^2 = \|(I_n - M_k)f_n\|^2 + \|Pf_n\|^2 \tag{7.3}$$
$$rem_1(k^{(s)}) = rem_1(k) + e_n^T P f_n \tag{7.4}$$

For the proofs of the Theorems in the case of σ known, without loss of generality, we assume $\sigma^2 = 1$. In all the proofs, we denote $IC_{\lambda_n,\, d}(k)$ by $IC(k)$.

Proof of Theorem 1 (parametric, σ known). Under the assumption that $P(\hat{k}_n = k_n^*) \to 1$, we have $\forall \epsilon > 0$, $\exists\, n_1$ such that $P(\hat{k}_n = k_n^*) > 1 - \epsilon$ for $n > n_1$.

Now we consider $\frac{IC(k_n^{*(s)})}{IC(k_n^*)}$ for any $k_n^{*(s)}$ being a sub-model of k_n^* with $r_{k_n^{*(s)}} = r_{k_n^*} - 1$.

Observe

$$\|\mathbf{Y}_n - \hat{\mathbf{Y}}_k\|^2 = \|(I_n - M_k)f_n\|^2 + \|(I_n - M_k)e_n\|^2 + 2rem_1(k). \tag{7.5}$$

Then

$$\begin{aligned}
&\frac{IC(k_n^{*(s)})}{IC(k_n^*)} \\
= &\frac{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{k_n^{*(s)}}\|^2 + \lambda_n \log(n) r_{k_n^{*(s)}} - n + dn^{1/2}\log(n)}{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{k_n^*}\|^2 + \lambda_n \log(n) r_{k_n^*} - n + dn^{1/2}\log(n)} \\
= &\frac{\|(I_n - M_{k_n^{*(s)}})f_n\|^2 + rem_2(k_n^{*(s)}) + 2rem_1(k_n^{*(s)}) + \lambda_n \log(n) r_{k_n^{*(s)}} + dn^{\frac{1}{2}}\log(n)}{\|(I_n - M_{k_n^*})f_n\|^2 + rem_2(k_n^*) + 2rem_1(k_n^*) + \lambda_n \log(n) r_{k_n^*} + dn^{\frac{1}{2}}\log(n)} \\
= &\frac{\|(I_n - M_{k_n^{*(s)}})f_n\|^2 + rem_2(k_n^{*(s)}) + 2rem_1(k_n^{*(s)}) + \lambda_n \log(n)(r_{k_n^*} - 1) + dn^{\frac{1}{2}}\log(n)}{rem_2(k_n^*) + \lambda_n \log(n) r_{k_n^*} + dn^{\frac{1}{2}}\log(n)}.
\end{aligned}$$

By Fact 2,

$$P(\|(I_n - M_{k_n^*})e_n\|^2 - (n - r_{k_n^*}) \geq \kappa(n - r_{k_n^*})) \leq e^{-\frac{n - r_{k_n^*}}{2}(\kappa - \ln(1+\kappa))} \text{ for } \kappa > 0,$$

and $$P(\|(I_n - M_{k_n^*})e_n\|^2 - (n - r_{k_n^*}) \leq -\kappa(n - r_{k_n^*})) \leq e^{-\frac{n - r_{k_n^*}}{2}(-\kappa - \ln(1-\kappa))} \text{ for } 0 < \kappa < 1.$$

For the given $\tau > 0$, let $\kappa = \frac{n^{\frac{1}{2}+\tau} h_n}{n - r_{k_n^*}}$ for some $h_n \to 0$. Note that when n is large enough, say $n > n_2 > n_1$, we have $0 < \kappa = \frac{n^{\frac{1}{2}+\tau} h_n}{n - r_{k_n^*}} < 1$.

Since $x - \log(1+x) \geq \frac{1}{4}x^2$ and $-x - \log(1-x) \geq \frac{1}{4}x^2$ for $0 < x < 1$, we have

$$P\left(\left|\|(I_n - M_{k_n^*})e_n\|^2 - (n - r_{k_n^*})\right| \geq h_n n^{\frac{1}{2}+\tau}\right) \leq 2e^{-\frac{n - r_{k_n^*}}{8}\kappa^2} \leq 2e^{-\frac{1}{8}n^{2\tau}h_n^2}.$$

By Fact 1, $\forall c > 0$,

$$\begin{aligned}
P\left(\frac{|rem_1(k_n^{*(s)})|}{\|(I_n - M_{k_n^{*(s)}})f_n\|^2} \geq c\right) &= P\left(\frac{|rem_1(k_n^{*(s)})|}{\|(I_n - M_{k_n^{*(s)}})f_n\|} \geq c\|(I - M_{k_n^{*(s)}})f_n\|\right) \\
&\leq e^{-c^2\|(I - M_{k_n^{*(s)}})f_n\|^2/2}.
\end{aligned}$$

Thus $\left|\frac{IC(k_n^{*(s)})}{IC(k_n^*)}\right|$ is no smaller than

$$\frac{\left|\|(I_n - M_{k_n^{*(s)}})f_n\|^2 + rem_2(k_n^{*(s)}) + 2rem_1(k_n^{*(s)}) + \lambda_n \log(n)(r_{k_n^*} - 1) + dn^{\frac{1}{2}}\log(n)\right|}{h_n n^{1/2+\tau} + r_{k_n^*}(\lambda_n \log(n) - 1) + dn^{1/2}\log(n)}$$

with probability higher than $1 - 2e^{-\frac{1}{8}n^{2\tau}h_n^2}$.

Note that $IC(k_n^{*(s)})$ is no smaller than

$$(1-2c)\|(I_n-M_{k_n^{*(s)}})f_n\|^2-h_nn^{1/2+\tau}+(r_{k_n^*}-1)(\lambda_n\log(n)-1)+dn^{1/2}\log(n)$$

with probability higher than $1-e^{-\frac{1}{8}n^{2\tau}h_n^2}-e^{-c^2\|(I-M_{k_n^{*(s)}})f_n\|^2/2}$. Since A_n is of order higher than $h_nn^{\frac{1}{2}+\tau}$ and for $c<1/2$ (to be chosen), there exists $n_3>n_2$ such that $IC(k_n^{*(s)})$ is positive for $n>n_3$ with probability higher than $1-e^{-\frac{1}{8}n^{2\tau}h_n^2}-e^{-c^2\|(I-M_{k_n^{*(s)}})f_n\|^2/2}$.

Thus for $n>n_3$, $\left|\frac{IC(k_n^{*(s)})}{IC(k_n^*)}\right|$ is no smaller than

$$\frac{(1-2c)\|(I_n-M_{k_n^{*(s)}})f_n\|^2-h_nn^{1/2+\tau}+(r_{k_n^*}-1)(\lambda_n\log(n)-1)+dn^{1/2}\log(n)}{h_nn^{1/2+\tau}+r_{k_n^*}\lambda_n\log(n)+dn^{1/2}\log(n)}$$

with probability higher than $1-2e^{-\frac{1}{8}n^{2\tau}h_n^2}-(e^{-\frac{1}{8}n^{2\tau}h_n^2}+e^{-c^2\|(I-M_{k_n^{*(s)}})f_n\|^2/2})$.

And for $n>n_3$,

$$\begin{aligned}
&\inf_{k_n^{*(s)}}\left|\frac{IC(k_n^{*(s)})}{IC(k_n^*)}\right| \\
\geq\ &\inf_{k_n^{*(s)}}\frac{(1-2c)\|(I_n-M_{k_n^{*(s)}})f_n\|^2-h_nn^{1/2+\tau}+(r_{k_n^*}-1)(\lambda_n\log(n)-1)+dn^{\frac{1}{2}}\log(n)}{h_nn^{1/2+\tau}+r_{k_n^*}\lambda_n\log(n)+dn^{1/2}\log(n)} \\
=\ &\frac{(1-2c)A_n-h_nn^{1/2+\tau}+(r_{k_n^*}-1)(\lambda_n\log(n)-1)+dn^{1/2}\log(n)}{h_nn^{1/2+\tau}+r_{k_n^*}\lambda_n\log(n)+dn^{1/2}\log(n)}
\end{aligned}$$

with probability higher than $1-2e^{-\frac{1}{8}n^{2\tau}h_n^2}-r_{k_n^*}\cdot(e^{-\frac{1}{8}n^{2\tau}h_n^2}+e^{-c^2A_n/2})$.

According to Conditions (P1) and (P2), $r_{k_n^*}=o(n^{\frac{1}{2}+\tau})/(\lambda_n\log(n))$ and A_n is of order $n^{1/2+\tau}$ or higher, we can choose h_n such that $2e^{-\frac{1}{8}n^{2\tau}h_n^2}+r_{k_n^*}\cdot(e^{-\frac{1}{8}n^{2\tau}h_n^2}+e^{-c^2A_n/2})\to 0$.

For example, taking $h_n=n^{-\tau/3}$, then

$$\begin{aligned}
\inf_{k_n^{*(s)}}\left|\frac{IC(k_n^{*(s)})}{IC(k_n^*)}\right| &\geq \frac{(1-2c)A_n-n^{1/2+2\tau/3}+(r_{k_n^*}-1)\lambda_n\log(n)+dn^{1/2}\log(n)}{n^{1/2+2\tau/3}+r_{k_n^*}\lambda_n\log(n)+dn^{1/2}\log(n)} \\
&:= bound_n
\end{aligned}$$

with probability higher than $1-2e^{-\frac{1}{8}n^{4\tau/3}}-r_{k_n^*}(e^{-\frac{1}{8}n^{4\tau/3}}+e^{-c^2A_n/2}):=1-p_n$.

With $c<1/2$, A_n of order $n^{1/2+\tau}$ or higher, and $r_{k_n^*}\lambda_n\log(n)=o(A_n)$, we have $\forall M>0,\exists n_4>n_3$ such that $bound_n\geq M$ and $p_n\leq\epsilon$ for $n>n_4$.

Therefore, $\forall M > 0$, $\epsilon > 0$, when $n > n_4$

$$P(|PI_n| \geq M) \geq 1 - 2\epsilon.$$

That is, $PI_n \xrightarrow{p} \infty$.

□

Proof of Theorem 2 (nonparametric, σ known). Similar to the proof of Theorem 1, consider $\frac{IC(\hat{k}_n^{(s)})}{IC(\hat{k}_n)}$ for any $\hat{k}_n^{(s)}$ being a sub-model of $\hat{k}_n$ with one fewer term, and we have

$$\frac{IC(\hat{k}_n^{(s)})}{IC(\hat{k}_n)} = 1 + \frac{\|Pf_n\|^2 + \|Pe_n\|^2 + e_n^T P f_n - \lambda_n \log(n)}{\|(I_n - M_{\hat{k}_n})f_n\|^2 + rem_2(\hat{k}_n) + 2rem_1(\hat{k}_n) + \lambda_n \log(n) r_{\hat{k}_n} + dn^{\frac{1}{2}} \log(n)}.$$

Next consider the terms in the above equation for any model k_n. For ease of notation, we write $B_{r_{k_n},n} = B_{r_{k_n}}$, where r_{k_n} is the rank of the projection matrix of model k_n.

By Fact 1, $\forall c_1 > 0$,

$$\begin{aligned} & P\left(\frac{|rem_1(k_n)|}{(\lambda_n \log(n) - 1)r_{k_n} + \|(I_n - M_{k_n})f_n\|^2 + dn^{1/2}\log(n)} \geq c_1\right) \\ \leq\ & e^{-c_1^2 \frac{(\lambda_n \log(n)-1)r_{k_n} + \|(I_n - M_{k_n})f_n\|^2 + dn^{1/2}\log(n)}{2}} \leq e^{-c_1^2 B_{r_{k_n}}/2}. \end{aligned}$$

Similarly, $\forall c_2 > 0$,

$$P\left(\frac{|e_n^T P f_n|}{B_{r_{k_n}}} \geq c_2\right) \leq e^{-\frac{c_2^2 B_{r_{k_n}}^2}{2\|Pf_n\|^2}} \leq e^{-c_2^2 B_{r_{k_n}}/2} \quad (\text{if } \|Pf_n\|^2 \leq B_{r_{k_n}}), \tag{7.6}$$

$$P\left(\frac{|e_n^T P f_n|}{\|Pf_n\|^2} \geq c_2\right) \leq e^{-\frac{c_2^2 \|Pf_n\|^2}{2}} \leq e^{-c_2^2 B_{r_{k_n}}/2} \quad (\text{if } \|Pf_n\|^2 > B_{r_{k_n}}). \tag{7.7}$$

By Fact 2,

$$P\left(\|(I_n - M_{k_n})e_n\|^2 - (n - r_{k_n}) \leq -\kappa(n - r_{k_n})\right) \leq e^{-\frac{n - r_{k_n}}{2}(-\kappa - \log(1-\kappa))}.$$

We can choose κ such that $\kappa(n - r_{k_n}) = \gamma B_{r_{k_n}}$ for some $0 < \gamma < 1$. Note that $-x - \log(1 - x) > x^2/2$ for $0 < x < 1$. Then

$$P\left(\|(I_n - M_{k_n})e_n\|^2 - (n - r_{k_n}) \leq -\gamma_n B_{r_{k_n}}\right) \leq e^{-\frac{\gamma^2 B_{r_{k_n}}^2}{4(n - r_{k_n})}}. \tag{7.8}$$

Still by Fact 2, for a sequence $D_n > 0$ (to be chosen), we have

$$P\left(\|Pe_n\|^2 - 1 \geq D_n\right) \leq e^{-(D_n - \log(1+D_n))}.$$

Note for $x > 1$, $x - \log(1+x) > x/2$. Thus, $P\left(\|Pe_n\|^2 - 1 \geq D_n\right) \leq e^{-D_n/2}$ for $D_n > 1$.

Since $\hat{k}_n$ is random, we apply union bounds on the exception probabilities. According to Condition (N1), for any $\epsilon > 0$, there exists n_1 such that $P(a_n \leq r_{\hat{k}_n} \leq b_n) \geq 1 - \epsilon$ for $n > n_1$. As will be seen, when n is large enough, the following quantities can be arbitrarily small for appropriate choice of γ, D_n, c_1 and c_2:

$$\sum_{j=a_n}^{b_n} N_j \cdot e^{-\frac{\gamma^2 B_{j,n}^2}{4(n-j)}}, \quad \sum_{j=a_n}^{b_n} N_j \cdot L_j \cdot e^{-D_n/2}, \quad \sum_{j=a_n}^{b_n} N_j \cdot e^{-c_1^2 B_{j,n}/2}, \quad \sum_{j=a_n}^{b_n} N_j \cdot L_j \cdot e^{-c_2^2 B_{j,n}/2}.$$

More precisely, we claim that there exists $n_2 > n_1$ such that for $n \geq n_2$,

(7.9)
$$\sum_{j=a_n}^{b_n} \left\{ N_j \cdot \left(e^{-\frac{\gamma^2 B_{j,n}^2}{4(n-j)}} + e^{-c_1^2 B_{j,n}/2} \right) + N_j \cdot L_j \cdot \left(e^{-D_n/2} + e^{-c_2^2 B_{j,n}/2} \right) \right\} \leq \epsilon.$$

Then for $n > n_2$ with probability higher than $1 - 2\epsilon$,

$$\begin{aligned}
a_n \leq r_{\hat{k}_n} &\leq b_n \\
\|(I_n - M_{\hat{k}_n})e_n\|^2 - (n - r_{\hat{k}_n}) &\geq -\gamma B_{r_{\hat{k}_n}} \\
\|P_{\hat{k}_n^{(s)},\hat{k}_n} e_n\|^2 &\leq 1 + D_n \\
|rem_1(\hat{k}_n)| &\leq c_1((\lambda_n \log(n) - 1) r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n}) f_n\|^2 + dn^{\frac{1}{2}} \log(n)) \\
|e_n^T P_{\hat{k}_n^{(s)},\hat{k}_n} f_n| &\leq c_2 B_{r_{\hat{k}_n}} \text{ or } |e_n^T P_{\hat{k}_n^{(s)},\hat{k}_n} f_n| \leq c_2 \|P_{\hat{k}_n^{(s)},\hat{k}_n} f_n\|^2.
\end{aligned}$$

Note that

(7.10)
$$PI_n = 1 + \inf_{\hat{k}_n^{(s)}} \frac{\|Pf_n\|^2 + \|Pe_n\|^2 + e_n^T P f_n - \lambda_n \log(n)}{\|(I_n - M_{\hat{k}_n}) f_n\|^2 + rem_2(\hat{k}_n) + 2rem_1(\hat{k}_n) + \lambda_n \log(n) r_{\hat{k}_n} + dn^{1/2} \log(n)}.$$

Also with probability higher than $1 - 2\epsilon$, the denominator in equation (7.10) is bigger than $(1-2c_1)\left[\|(I_n - M_{\hat{k}_n}) f_n\|^2 + (\lambda_n \log(n) - 1) r_{\hat{k}_n} + dn^{1/2} \log(n)\right] - \gamma B_{r_{\hat{k}_n}}$. Thus when $2c_1 + \gamma < 1$, the denominator in (7.10) is positive.

Then for $n > n_2$, with probability at $1 - 2\epsilon$ we have

$$PI_n = 1 + \frac{\inf_{\hat{k}_n^{(s)}} (\|Pf_n\|^2 + \|Pe_n\|^2 + e_n^T P f_n - \lambda_n \log(n))}{\|(I_n - M_{\hat{k}_n}) f_n\|^2 + rem_2(\hat{k}_n) + 2rem_1(\hat{k}_n) + \lambda_n \log(n) r_{\hat{k}_n} + dn^{1/2} \log(n)}.$$

For $n > n_2$ with probability higher than $1-2\epsilon$, if $\|Pf_n\|^2 \le B_{r_{\hat{k}_n}}$, then

$$PI_n - 1 \le \frac{\inf_{\hat{k}_n^{(s)}} \|Pf_n\|^2 + 1 + D_n + c_2 B_{r_{\hat{k}_n}} + \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n}) f_n\|^2 + dn^{\frac{1}{2}} \log(n))}$$

$$\text{and} \quad PI - 1 \ge \frac{\inf_{\hat{k}_n^{(s)}} \|Pf_n\|^2 - 1 - D_n - c_2 B_{r_{\hat{k}_n}} - \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n}) f_n\|^2 + dn^{\frac{1}{2}} \log(n))},$$

$$\text{otherwise,} \quad PI_n - 1 \le \frac{\inf_{\hat{k}_n^{(s)}} \|Pf_n\|^2 + 1 + D_n + c_2 \|Pf_n\|^2 + \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n}) f_n\|^2 + dn^{\frac{1}{2}} \log(n))}$$

$$\text{and} \quad PI_n - 1 \ge \frac{\inf_{\hat{k}_n^{(s)}} \|Pf_n\|^2 - 1 - D_n - c_2 \|Pf_n\|^2 - \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n}) f_n\|^2 + dn^{\frac{1}{2}} \log(n))}.$$

Next we focus on the case $\|Pf_n\|^2 \le B_{r_{\hat{k}_n}}$. The case of $\|Pf_n\|^2 > B_{r_{\hat{k}_n}}$ can be similarly handled. Note that $\sup_{a_n \le j \le b_n} \frac{B_{j,n}}{n-j} := \zeta_n' \to 0$. Let $\zeta_n'' = \zeta_n + \zeta_n'$. Taking $\gamma = \sqrt{4/5}$, $D_n = 4\zeta_n'' B_{r_{k_n}}$, $c_2 = 2\sqrt{\zeta_n''}$, $0 < c_1 < \frac{1-\gamma}{2}$, then

$$\begin{aligned}
& PI_n - 1 \\
\le\ & \frac{\inf_{\hat{k}_n^{(s)}} \|Pf_n\|^2 + 1 + 4\zeta_n'' B_{r_{\hat{k}_n}} + 2\sqrt{\zeta_n''} B_{r_{\hat{k}_n}} + \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n}) f_n\|^2 + dn^{1/2} \log(n))} \\
\le\ & \sup_{a_n \le r_{k_n} \le b_n} \frac{\inf_{k_n^{(s)}} \|Pf_n\|^2 + 1 + 4\zeta_n'' B_{r_{k_n}} + 2\sqrt{\zeta_n''} B_{r_{k_n}} + \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{k_n} + \|(I_n - M_{k_n}) f_n\|^2 + dn^{\frac{1}{2}} \log(n))} \\
:=\ & Upperbound_n \\
\to\ & 0 \text{ according to (N3) and the fact that } \zeta_n'' \to 0 \text{ as } n \to \infty.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& PI_n - 1 \\
\ge\ & -\frac{1 + 4\zeta_n'' B_{r_{k_n}} + 2\sqrt{\zeta_n''} B_{r_{k_n}} + \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n}) f_n\|^2 + dn^{1/2} \log(n))} \\
\ge\ & -\sup_{a_n \le r_{k_n} \le b_n} \frac{1 + 4\zeta_n'' B_{r_{k_n}} + 2\sqrt{\zeta_n''} B_{r_{k_n}} + \lambda_n \log(n)}{(1-2c_1-\gamma)((\lambda_n \log(n)-1) r_{k_n} + \|(I_n - M_{k_n}) f_n\|^2 + dn^{\frac{1}{2}} \log(n))} \\
:=\ & Lowerbound_n \\
\to\ & 0 \text{ according to } (N3) \text{ and the fact that } \zeta_n'' \to 0.
\end{aligned}$$

Therefore, $\forall \delta > 0, \exists\, n_3$ such that $Upperbound_n \le \delta$ and $Lowerbound_n \ge -\delta$ for $n > n_3$. Thus, $\forall \epsilon > 0, \delta > 0, \exists N = \max(n_2, n_3)$ such that $P(|PI_n - 1| \le \delta) \ge 1 - 2\epsilon$ for $n > N$. That is, $PI_n \xrightarrow{p} 1$.

To complete the proof, we just need to check the claim of (7.9). By Condition (N2), $\forall \epsilon > 0$, $\exists\, n_\epsilon$ such that for $n \geq n_\epsilon$, $\sum_{j=a_n}^{b_n} c_0 \cdot e^{-\frac{B_{j,n}^2}{10(n-j)}} < \epsilon/4$. Then for $n > n_\epsilon$,

$$\sum_{j=a_n}^{b_n} N_j \cdot e^{-\frac{\gamma^2 B_{j,n}^2}{4(n-j)}} \leq \sum_{j=a_n}^{b_n} c_0 \cdot e^{\frac{B_{j,n}^2}{10(n-j)}} \cdot e^{-\frac{\gamma^2 B_{j,n}^2}{4(n-j)}} \leq \sum_{j=a_n}^{b_n} c_0 \cdot e^{-\frac{B_{j,n}^2}{10(n-j)}} < \epsilon/4$$

$$\sum_{j=a_n}^{b_n} N_j \cdot L_j \cdot e^{-D_n/2} = \sum_{j=a_n}^{b_n} N_j \cdot L_j \cdot e^{-2\zeta_n'' B_{j,n}} \leq \sum_{j=a_n}^{b_n} c_0 \cdot e^{-\zeta_n'' B_{j,n}} < \frac{\epsilon}{4}.$$

Similarly,

$$\sum_{j=a_n}^{b_n} N_j \cdot e^{-c_1^2 B_{j,n}/2} < \frac{\epsilon}{4}, \qquad \sum_{j=a_n}^{b_n} N_j \cdot L_j \cdot e^{-c_2^2 B_{j,n}/2} < \frac{\epsilon}{4}.$$

Thus claim (7.9) holds and this completes the proof. □

Proof of Theorem 1 (parametric, σ unknown). The proof of the case of unknown σ is almost the same as the one for the case where σ is known.

Note that $\hat{\sigma}_n^2 = \frac{1}{n-r_{k_n^*}} \|\mathbf{Y}_n - \hat{\mathbf{Y}}_{k_n^*}\|^2 = \frac{1}{n-r_{k_n^*}} \|(I_n - M_{k_n^*})e_n\|^2$, and $\frac{IC(k_n^{*(s)})}{IC(k_n^*)}$ is equal to

$$= \frac{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{k_n^{*(s)}}\|^2 + \lambda_n \log(n) r_{k_n^{*(s)}} \hat{\sigma}_n^2 - n\hat{\sigma}_n^2 + dn^{1/2}\log(n)\hat{\sigma}_n^2}{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{k_n^*}\|^2 + \lambda_n \log(n) r_{k_n^*} \hat{\sigma}_n^2 - n\hat{\sigma}_n^2 + dn^{1/2}\log(n)\hat{\sigma}_n^2}$$

$$= \frac{\frac{\|(I_n - M_{k_n^{*(s)}})f_n\|^2}{\sigma^2} + \frac{\|(I_n - M_{k_n^{*(s)}})e_n\|^2}{\sigma^2} + 2\frac{rem_1(k_n^{*(s)})}{\sigma^2} + \lambda_n \log(n)(r_{k_n^*} - 1)\frac{\hat{\sigma}_n^2}{\sigma^2} - n\frac{\hat{\sigma}_n^2}{\sigma^2}}{[(\lambda_n \log(n) - 1)r_{k_n^*} + dn^{\frac{1}{2}}\log(n)]\frac{\hat{\sigma}_n^2}{\sigma^2}}$$

$$+ \frac{dn^{\frac{1}{2}}\log(n)}{(\lambda_n \log(n) - 1)r_{k_n^*} + dn^{\frac{1}{2}}\log(n)}.$$

Then similarly, we can bound the stochastic terms of $\|(I_n - M_{k_n^{*(s)}})e_n\|^2$, $rem_1(k_n^{*(s)})$, and $\frac{\hat{\sigma}_n^2}{\sigma^2}$. We shall omit the rest of the proof.

□

Proof of Theorem 3 (nonparametric, σ unknown). The proof of Theorem 3 is similar to that of Theorem 2. We only point out the major difference. Note that

$$\hat{\sigma}_n^2 = \frac{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{\hat{k}_n}\|^2}{n - r_{\hat{k}_n}} = \frac{1}{n - r_{\hat{k}_n}} \left[\|(I_n - M_{\hat{k}_n})f_n\|^2 + \|(I_n - M_{\hat{k}_n})e_n\|^2 + 2rem_1(\hat{k}_n) \right].$$

Then

$$\frac{IC(\hat{k}_n^{(s)})}{IC(\hat{k}_n)} = \frac{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{\hat{k}_n^{(s)}}\|^2 + \lambda_n \log(n) r_{\hat{k}_n^{(s)}} \hat{\sigma}_n^2 - n\hat{\sigma}_n^2 + dn^{\frac{1}{2}} \log(n)\hat{\sigma}^2}{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{\hat{k}_n}\|^2 + \lambda_n \log(n) r_{\hat{k}_n} \hat{\sigma}_n^2 - n\hat{\sigma}_n^2 + dn^{\frac{1}{2}} \log(n)\hat{\sigma}^2}$$

$$= 1 + \frac{\|Pf_n\|^2 + \|Pe_n\|^2 + e_n^T P f_n - \lambda_n \log(n)\hat{\sigma}_n^2}{[(\lambda_n \log(n) - 1) r_{\hat{k}_n} + dn^{\frac{1}{2}} \log(n)]\hat{\sigma}_n^2},$$

and $PI_n - 1 = \frac{\inf_{\hat{k}_n^{(s)}} (\|Pf_n\|^2 + \|Pe_n\|^2 + e_n^T P f_n)}{[(\lambda_n \log(n) - 1) r_{\hat{k}_n} + dn^{\frac{1}{2}} \log(n)]\hat{\sigma}_n^2} - \frac{\lambda_n \log(n)}{(\lambda_n \log(n) - 1) r_{\hat{k}_n} + dn^{\frac{1}{2}} \log(n)}$.

The rest of the proof follows similarly.

□

Acknowledgments. The authors thank Dennis Cook, Charles Geyer, Wei Pan, Hui Zou and the participants at a seminar that one of the authors gave in Department of Statistics at Yale University for helpful comments and discussions. Comments from three reviewers and an AE are appreciated.

REFERENCES

[1] Akaike, H. (1969). Fitting autoregressive models for regression. *Ann. Inst. Statist. Math.*, **21**, 243-247.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceed. 2nd Int. Symp. on Infor. Theory*, Ed. B. N. Petrov and F. Csaki. Budapest: Akademia Kiado. 267-281

[3] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM - Probability and Statistics*, **6**, 127-146.

[4] Barron, A. (1994). Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning*, **14**, 115-133.

[5] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection by penalization. *Prob. Theory and Related Fields*, **113**, 301-413.

[6] Barron, A. and Cover, T. (1991). Minimum complexity density estimation. *IEEE Trans. on Infor. Theory*, **37**, 1034-1054.

[7] Barron, A.R., Yang, Y., Yu, B. (1994). Asymptotically optimal function estimation by minimum complexity criteria. *Proceed. 1994 Int. Symp. Info. Theory, Trondheim, Norway: IEEE Info. Theory Soc.*, 38.

[8] Berger, J. O. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison, in *Model Selection*, ed. P. Lahiri, Institute of Mathematical Statistics Lecture Notes – Monograph Series, **38**, Beachwood Ohio, 135–207.

[9] Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange Facts. *Prob. Theory and Related Fields*, **71**, 271-291.

[10] Birgé, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales de l'institut Henri Poincare (B) Prob. and Statist.*, **42**, 273-325.

[11] Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, **80**, 580-598.

[12] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350-2383.

[13] Burnham, K.P. and Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods Research*, **33**, 167 - 187.

[14] Bunea, F., Tsybakov, A., Wegkamp, M. (2006). Aggregation and sparsity via l_1 penalized least squares. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **4005 LNAI**, 379-391.

[15] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2392-2404

[16] Chatfield, C. (1995). Model uncertainty, data mining, and statistical inference (with discussion). *J. Roy. Statist. Soc. Ser. A.*, **158**, 419-466.

[17] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 1-13.

[18] Claeskens, G., Hjort, N. (2003). The Focused Information Criterion. *J. Amer. Statist. Assoc.*, **98**, 900-916.

[19] Cook, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statistical Science*, **22**, 1-26.

[20] Cox, D. (1995). Model uncertainty, data mining, and statistical inference: discussion. *J. Roy. Statist. Soc. Ser. A.*, **158**, 455-456.

[21] Danilov, D. and Magnus, J. (2004). On the harm that ignoring pretesting can cause. *J. Econometrics*, **122**, 27-46.

[22] Devroye, L., Gyrfi, L., and Lugosi, G. (1997). *A Probabilistic Theory of Pattern Recognition. Series: Stochastic Modelling and Applied Probability.*, Springer, **31**.

[23] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508-539.

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

[25] Erven, T., Grünwald, P., and de Rooij, S. (2008). Catching up faster by switching sooner: a prequential solution to the AIC-BIC dilemma, Arxiv preprint arXiv:0807.1005.

[26] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

[27] Faraway, J.J. (1992). On the Cost of Data Analysis. *J. Computational and Graphical Statist.*, **1**, 213229.

[28] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*, **222**, 309-368.

[29] Freedman, D. (1995). Some issues in the foundation of statistics. *Foundations of Science*, **1**, 19-83.

[30] George, E. and Foster, D. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731-747.

[31] Geyer, C. and Shaw, R. (2008). Model selection in estimation of fitness landscapes. *Technical Report.*, University of Minnesota.

[32] Hand, D. J. (1981). Branch and bound in statistical data analysis. *The Statistician.*, **30**, 1-13.

[33] Hansen, M. and Yu, B. (1999). Bridging AIC and BIC: an MDL model selection criterion. *In Proceed. of IEEE Infor. Theory Workshop on Detection, Estimation, Classification and Imaging,* 63.

[34] Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *J. Environmental Economics Management.*, **5**, 81-102.

[35] Hurvich, C. M., and Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, **44**, 214217.

[36] Hawkins, D. (1989). Flexible parsimonious smoothing and additive modeling: discussion. *Technometrics*, **31**, 31-34.

[37] Ibragimov, I.A., Hasminskii, R.Z. (1977). On the estimation of an infinite-dimensional parameter in Gaussian white noise. *Soviet Math. Dokl.*, **18**, 1307-1309.

[38] Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Statist.*, **35**, 1238-1277.

[39] Ing, C.-K., Wei, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Annals of Statistics*, **33**, 2423-2474.

[40] Kabaila P., Leeb H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association*, **101**, 619-629.

[41] Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, **34**, 2593-2656.

[42] Leeb, H. and Pötscher, B. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, **34**, 2554-2591.

[43] Li, K.-C. (1987). Asymptotic optimality for Cp, CL, cross-validation and generalized crossvalidation: discrete index set. *Ann. Statist.*, **15**, 958-975.

[44] Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

[45] McQuarrie, A. and Tsai, C.L. (1998). *Regression and Time Series Model Selection.* World Scientific: Singapore.

[46] Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436-1462.

[47] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758-765.

[48] Pötscher, B.M. (1991). Effects of model selection on inference. *Econometric Theory*, **7**, 163185.

[49] Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369-374.

[50] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

[51] Shao, J. (1997). An asymptotic theoy for linear model selection (with discussion). *Statist. Sinica*, **7**, 221-264.

[52] Shen X., Huang H.-C. (2006). Optimal model assessment, selection, and combination *J. Amer. Statist. Assoc.*, **101**, 554-568.

[53] Shen, X. and Ye, J. (2002). Adaptive model selection. *J. Amer. Statist. Assoc.*, **97**, 210-221.

[54] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45-54.

[55] Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43-49.

[56] Sober, E. (2004). The contest between parsimony and likelihood. *Systematic Biology*, **53**, 644-653.

[57] Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040-1053.

[58] Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. Roy. Statist. Soc. Ser. B*, **41**,276-278.

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

[60] Wang, H., Li, R., and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.

[61] Yang, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica*, **9**, 475-499.

[62] Yang, Y. (2000). Combining different procedures for adaptive regression. *J. Multivariate Analysis* , **74**, 135-161.

[63] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, **92**, 937-950.

[64] Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, **35**, 2450-2473.

[65] Yang, Y. (2007) Prediction/Estimation With Simple Linear Models: Is It Really That Simple? *Econometric Theory*, **23**, 1-36.

[66] Yang, Y. and Barron, A. (1998). An asymptotic property of model selection criteria. *IEEE Trans. on Infor. Theory*, **44**, 95-116.

[67] Yang, Y. and Barron, A. (1999). Information theoretic determination of minimax rates of convergence. *Ann. Statist.*, **27**, 1564-1599.

[68] Zhang, C. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567-1594.

[69] Zhang, P. (1990). Inference after variable selection in linear regression models. *Biometrika*, **79**, 741-746.

[70] Zhang, P. (1997). An asymptotic theoy for linear model selection: discussion. *Statist. Sinica*, **7**, 254-258.

[71] Zhao, P. and Yu, B. (2006). On Model selection consistency of Lasso. *J. Machine Learning Research*, **7**, 2541-2563.

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

WEI LIU
SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
313 FORD HALL
224 CHURCH STREET S.E.
MINNEAPOLIS, MN 55455, US
E-MAIL: william050@stat.umn.edu

YUHONG YANG
SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
313 FORD HALL
224 CHURCH STREET S.E.
MINNEAPOLIS, MN 55455, US
E-MAIL: yyang@stat.umn.edu
HTTP://WWW.STAT.UMN.EDU/~YYANG