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 in an asymptotic efficiency 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 under some conditions. We develop a measure, parametricness index (PI), to assess whether a model selected by a potentially consistent procedure can be practically treated as the true model, which also hints on AIC or BIC is better suited for the data for the goal of estimating the regression function. 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 [55] 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 (see [68] for references).

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 if interpretation and statistical inference based on the selected model are questionable due to model selection uncertainty.
4. It tells us whether AIC is likely better than BIC for the data for the purpose of estimating the regression function.
5. It can be used to approximately achieve the better estimation performance of AIC and BIC for both parametric and nonparametric scenarios.

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., [17, 27, 40, 44, 48, 50, 51, 52, 53, 56, 59, 63, 65, 69, 73, 76, 77]) have been established mostly in terms of either selection consistency or an asymptotic optimality. It is well-known that AIC [2], C_p [49], and FPE [1, 60] 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 (see [56, 68] for references).

Another direction of model selection theory focuses on oracle risk bounds (also called index of resolvability bounds). When the candidate models are constructed to work well for target function classes, this approach 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, 24, 43, 71], to name a few. In particular, AIC type of model selection methods are minimax-rate optimal for both parametric and nonparametric scenarios under square error loss for estimating the regression function (see [5, 68]). 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, 71, 66]).

While pointwise asymptotic results are certainly of interest, it is not surprising that the limiting behaviors can be very different from the finite-sample reality, especially when model selection is involved. (see e.g., [41, 21, 45]).

The general forms of AIC and BIC make it very clear that they and similar criteria (such as GIC in [54]) 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, 31, 34, 39, 57, 58, 70]). Yang [70] 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 [39]. However, Yang [68, 70] proved that no model selection procedure can be both consistent (or pointwise adaptive) and minimax-rate optimal at the same time. As will be seen, if we can properly distinguish between parametric and nonparametric scenarios, a consequent data-driven 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. 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 (see, e.g., [8, 13, 16, 30, 75, 61]). There has been a lot of debate from not only statistical but also philosophical perspectives, especially about the existence of a true model and the ultimate goal of statistical modeling. 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 the 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.

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 [29], 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.

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 the 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.

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 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 simulation 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.

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 $x_i = (x_{i1}, \cdots, x_{ip})$ is the value of a p-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 $\sigma^2 > 0$.

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 family of regression functions linear in the parameter θ_k 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., [72] 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 in our context 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 $\lambda_n \geq (\log n)^{-1}$ and a constant $d \geq 0$, let

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

when σ is known, and

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

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 potentially consistent model selection method (i.e., it will select the true model with probability going to 1 as $n \to \infty$ if the true model is among the candidates). 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*: the true model at sample size n is in Γ and denoted by k_n^* with $r_{k_n^*}$ assumed to be larger than 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 $\hat{\sigma}_n^2 = \frac{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{\hat{k}_n}\|^2}{n - r_{\hat{k}_n}}$. We also have

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

Remarks: 1. The conditions (P1) 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, 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.

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 $1 \leq a_n < b_n < n$ (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 $a_n \le j \le b_n$,

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

(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*

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

Remarks: 1. For nonparametric regression, for familiar model selection methods, the order of $r_{\hat{k}_n}$ can be identified (e.g., [39, 72]), sometimes loosing a logarithmic factor, and (N1) is satisfied in a typical nonparametric situation.

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 estimating f, 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 $d = 0$.

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

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

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 a constant $c_0 > 0$ such that for $a_n \le j \le b_n$, $N_j \cdot L_j \le c_0 e^{\rho_n E_{j,n}}$, 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*

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

3.3. *PI separates parametric and nonparametric scenarios.* The results in Sections 3.1 and 3.2 imply that starting with a potentially consistent model selection procedure (i.e., it will be consistent if one of the candidate models holds), 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: $r_{k_n^*}$ increases with n.

In this case, both $r_{k_n^*}$ and p_n go to infinity with 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. 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 orthonormal 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 $n^{-\frac{1}{4}}$.

(iii). Nonparametric.

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

(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 $\mathbf{\Phi} = \{\phi_1(x),\cdots,\phi_k(x),\cdots\}$ 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, Condition (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 [39], 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 fixed finite-dimensional correct models, which implies that BIC is asymptotically loss efficient [56].

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. Previous work on sharing the strengths of AIC and BIC utilized minimum description length criterion in an adaptive fashion ([7, 34]), or flexible priors in a Bayesian framework ([31, 26]). Ing [39] and Yang [70] established (independently) simultaneous asymptotic efficiency for both parametric and nonparametric scenarios.

2. Recently, Erven, Grünwald, de Rooij [26] 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 shown in [68] can be resolved by a Bayesian strategy that allows switching between models.

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 at the given sample size, 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, 38, 62]). In addition, given a function class defined in terms of approximation error decay behavior by an approximating system, rates of convergence of minimax risks have been established (see, e.g., [72]). 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) [72]. Furthermore, adaptive estimators that are simultaneously optimal for multiple function classes can be obtained by model selection or model combining (see, e.g, [5, 67] 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 (i.e., report standard errors and confidence intervals for parameters and make interpretations based on the selected model without assessing its reliability).

Closely related to the above discussion is the issue of model selection uncertainty (see, e.g., [12, 16]). 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 discovery of the relevance of some additional features and then we may be in a 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 a practical nonparametric scenario again, and so on.

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.

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 $d = 0$.

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.

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. 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.

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

We have investigated the effects of sample size and magnitude of the coefficients on PI. The results show that i) given the regression function and the noise level, the value of PI indicates whether the problem is 'practically' parametric/nonparametric at the current sample size; 2) 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. See [47] for details.

5.2. *Multiple predictors.* In the 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 [33] 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 [32]). Unless otherwise stated, in these examples, the sample size is 200 and we replicate 300 times. The first two examples were used in [64].

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

Example 4. Differences from Example 3: $\beta_j = .85, \forall j$ and $\sigma = 3$.

Example 5. $\beta = (0.9, 0.9, 0, 0, 2, 0, 0, 1.6, 2.2, 0, 0, 0, 0)^T$. There are 13 predictors and the 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 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 Examples 3 and 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.

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.)

More simulation results are given in [47]. First, an illuminating example shows that with specially chosen coefficients, PI switches positions several times, as they should, in declaring practical parametricness or nonparametricness as more and more information is available. Second, it is shown that PI is informative on reliability of inference after model selection. When PI is large (Example 3), confidence intervals based on the selected model are quite trustworthy, but when PI is small (Example 4), the actual coverage probability intended at 95% is typically around 65%. While it is now well known that model selection has an impact on subsequent statistical inferences (see, e.g., [74, 36, 28, 42]), 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. Third, it is shown that an adaptive choice between AIC and BIC based on the PI value (choose BIC when PI is larger than 1.2) indeed leads to nearly the better performance of AIC and BIC and thus beats both AIC and BIC in an overall sense. So PI provides helpful information regarding whether AIC or BIC works better (or they have similar performances) in risks of estimation. Therefore, PI can be viewed as a **P**erformance **I**ndicator of AIC versus BIC.

Based on our numerical investigations, in nested model problems (like order selection for series expansion), a cutoff point of $c = 1.6$ seems proper. In subset selection problems, since the infimum in computing PI is taken over many models, the cutoff point is expected to be smaller, and 1.2 seems to be quite good.

6. Real Data Examples.

In this section, we study three data sets: the Ozone data with 10 predictors and $n = 330$ (e.g. [11]), the Boston housing data with 13 predictors and $n = 506$ (e.g. [35]), and the Diabetes data with 10 predictors and $n = 442$ (e.g. [25]).

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.

Based on sub-sampling at the sample size 400, we found that 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. Similar to the simulation results in Section 5, by parametric bootstrap at the original sample size from the selected model, combining AIC and BIC based on PI shows good overall performance in estimating the regression function. The combined procedure has a statistical risk close to the better one of AIC and BIC in each case. Details can be found in [47].

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 model 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 potentially consistent model selection rule (i.e., it will be consistent if one of the candidate models is true) 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 of f is automatically asymptotically efficient for both parametric and nonparametric scenarios, which clearly cannot be achieved by any deterministic choice of the penalty parameter λ_n in the criteria of the form $-\log$-likelihood$+\lambda_n m_k$, 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.

When working with parametric candidate models, 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 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 coverage 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 performance in estimating the regression function. 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. For instance, the focused information criterion (FIC) [17, 18] emphasizes performance at a given estimand, and it seems interesting to understand when FIC improves over AIC and how to take advantages of both in an implementable fashion.

For the purpose of estimating the regression function, it has been suggested that AIC performs better for a nonparametric scenario and BIC better for a parametric one (see [70] 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. Our numerical results tend to suggest that 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.

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 fact will be used in our proofs (see [66]).

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

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

For the ease of notation, we denote $P_{k^{(s)},k} = M_k - M_{k^{(s)}}$ 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

$$\|(I_n - M_{k^{(s)}})e_n\|^2 = \|(I_n - M_k)e_n\|^2 + \|Pe_n\|^2 \tag{7.1}$$
$$\|(I_n - M_{k^{(s)}})f_n\|^2 = \|(I_n - M_k)f_n\|^2 + \|Pf_n\|^2 \tag{7.2}$$
$$rem_1(k^{(s)}) = rem_1(k) + e_n^T P f_n \tag{7.3}$$

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 $n > n_1$.

Since $\|\mathbf{Y}_n - \hat{\mathbf{Y}}_k\|^2 = \|(I_n - M_k)f_n\|^2 + \|(I_n - M_k)e_n\|^2 + 2rem_1(k)$, for any $k_n^{*(s)}$ being a sub-model of k_n^* with $r_{k_n^{*(s)}} = r_{k_n^*} - 1$, we know that $\frac{IC(k_n^{*(s)})}{IC(k_n^*)}$

is equal to

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

By the fact on χ^2 distribution,

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

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

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

Since for $Z \sim N(0,1)$, $\forall\, t > 0$, $P(|Z| \geq t) \leq e^{-t^2/2}$, we know that $\forall c > 0$,

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

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

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

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 $h_n n^{\frac{1}{2}+\tau}$ and for $c < 1/2$ (to be chosen), there exists

$n_3 > n_2$ such that $IC(k_n^{*(s)})$ is positive for $n > n_3$ and $\left|\frac{IC(k_n^{*(s)})}{IC(k_n^*)}\right|$ is no smaller than

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

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

Then for $n > n_3$, $\inf_{k_n^{*(s)}} \left|\frac{IC(k_n^{*(s)})}{IC(k_n^*)}\right|$ is lower bounded by

$$\frac{(1-2c)A_n - h_n n^{1/2+\tau} + (r_{k_n^*} - 1)(\lambda_n \log(n) - 1) + dn^{1/2}\log(n)}{h_n n^{1/2+\tau} + r_{k_n^*}\lambda_n \log(n) + dn^{1/2}\log(n)}$$

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

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 $2e^{-\frac{1}{8}n^{2\tau}h_n^2} + r_{k_n^*} \cdot (e^{-\frac{1}{8}n^{2\tau}h_n^2} + e^{-c^2 A_n/2}) \to 0$.

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

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 that $\forall M > 0, \exists\, n_4 > n_3$ such that $bound_n \geq M$ and $q_n \leq \epsilon$ for $n > n_4$. Thus $PI_n \xrightarrow{p} \infty$.

□

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.

As in the proof of Theorem 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, $\forall c_2 > 0$,

$$P\left(\frac{|e_n^T P f_n|}{B_{r_{k_n}}} \geq c_2\right) \leq e^{-\frac{c_2^2 B^2_{r_{k_n}}}{2\|Pf_n\|^2}} \leq e^{-c_2^2 B_{r_{k_n}}/2} \qquad (\text{if } \|Pf_n\|^2 \leq B_{r_{k_n}}), \tag{7.4}$$

$$P\left(\frac{|e_n^T P f_n|}{\|Pf_n\|^2} \geq c_2\right) \leq e^{-\frac{c_2^2 \|Pf_n\|^2}{2}} \leq e^{-c_2^2 B_{r_{k_n}}/2} \qquad (\text{if } \|Pf_n\|^2 > B_{r_{k_n}}). \tag{7.5}$$

Also,

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

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

$$P\left(\|(I_n - M_{k_n})e_n\|^2 - (n - r_{k_n}) \leq -\gamma_n B_{r_{k_n}}\right) \leq e^{-\frac{\gamma^2 B^2_{r_{k_n}}}{4(n - r_{k_n})}}. \tag{7.6}$$

For a sequence $D_n > 0$ (to be chosen), we have

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

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

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

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

$$\sum_{j=a_n}^{b_n} \left\{ N_j \cdot \left(e^{-\frac{\gamma^2 B^2_{j,n}}{4(n-j)}} + e^{-c_1^2 B_{j,n}/2} \right) + N_j \cdot L_j \cdot \left(e^{-D_n/2} + e^{-c_2^2 B_{j,n}/2} \right) \right\} \leq \epsilon. \tag{7.7}$$

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

$$a_n \le r_{\hat{k}_n} \le b_n$$

$$\|(I_n - M_{\hat{k}_n})e_n\|^2 - (n - r_{\hat{k}_n}) \ge -\gamma B_{r_{\hat{k}_n}}$$

$$\|P_{\hat{k}_n^{(s)},\hat{k}_n} e_n\|^2 \le 1 + D_n$$

$$|rem_1(\hat{k}_n)| \le c_1((\lambda_n \log(n) - 1)r_{\hat{k}_n} + \|(I_n - M_{\hat{k}_n})f_n\|^2 + dn^{\frac{1}{2}}\log(n))$$

$$|e_n^T P_{\hat{k}_n^{(s)},\hat{k}_n} f_n| \le c_2 B_{r_{\hat{k}_n}} \text{ or } |e_n^T P_{\hat{k}_n^{(s)},\hat{k}_n} f_n| \le c_2\|P_{\hat{k}_n^{(s)},\hat{k}_n} f_n\|^2.$$

Note that
(7.8)

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

Also with probability higher than $1-2\epsilon$, the denominator in equation (7.8) 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 (7.8) is positive.

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

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

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

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

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 (7.7). By Condition (N2), $\forall \epsilon > 0$, $\exists\ n_\epsilon$ such that for $n \ge n_\epsilon$, $\sum_{j=a_n}^{b_n} c_0 \cdot e^{-\frac{B_{j,n}^2}{10(n-j)}} < \epsilon/4$. Then for $n > n_\epsilon$,

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

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

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 (7.7) holds and this completes the proof. □

The proofs of the cases with unknown σ in Theorems 1 and 3 are almost the same as those when σ is known. Due to space limitation, we omit the details.

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 all the reviewers, associate and editors are appreciated.

SUPPLEMENTARY MATERIAL

Supplement A: Details and more numerical examples (http://lib.stat.cmu.edu/aoas/???/???). We provide complete descriptions and more results of our numerical work.

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

[18] Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging.* Cambridge University Press.

[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] Efroimovich, S. (1985). Nonparametric estimation of a density of unknown smoothness, *Theory Probab. Appl.*, **30**, 557-568.

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

[26] 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.

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

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

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

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

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

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

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

[34] 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.

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

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

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

[38] 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.

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

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

[41] Kabaila, P. (2002). On variable selection in linear regression. *Econometric Theory*, *18*, 913-925.

[42] 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.

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

[44] Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process, *Bernoulli*, **14**, 661-690.

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

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

[47] Liu, W. and Yang, Y. (2011). Supplement to "Parametric or nonparametric? A parametricness index for model selection".

[48] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, **37**, 3498-3528.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[77] 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