Received: 17 August 2022 | Revised: 25 December 2022 | Accepted: 16 January 2023

DOI: 10.1002/wics.1607

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS

WILEY

Information criteria for model selection

Jiawei Zhang | Yuhong Yang | Jie Ding

School of Statistics, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA

Correspondence
Jie Ding, School of Statistics, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA.
Email: dingj@umn.edu

Funding information
National Science Foundation, Grant/Award Numbers: DMS-2134148, ECCS-2038603

Edited by: James E. Gentle, Commissioning Editor and David W. Scott, Review Editor and Co-Editor-in-Chief

Abstract

The rapid development of modeling techniques has brought many opportunities for data-driven discovery and prediction. However, this also leads to the challenge of selecting the most appropriate model for any particular data task. Information criteria, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC), have been developed as a general class of model selection methods with profound connections with foundational thoughts in statistics and information theory. Many perspectives and theoretical justifications have been developed to understand when and how to use information criteria, which often depend on particular data circumstances. This review article will revisit information criteria by summarizing their key concepts, evaluation metrics, fundamental properties, interconnections, recent advancements, and common misconceptions to enrich the understanding of model selection in general.

This article is categorized under:

Data: Types and Structure > Traditional Statistical Data
Statistical Learning and Exploratory Methods of the Data Sciences > Modeling Methods
Statistical and Graphical Methods of Data Analysis > Information Theoretic Methods
Statistical Models > Model Selection

KEYWORDS

Akaike information criterion, Bayesian information criterion, information criteria, model selection, variable selection

1 | INTRODUCTION

Data-driven discovery and prediction have been the backbone of many science and engineering domains. Multiple candidate models are often considered when we establish models for a particular data task. Examples include selecting the most appropriate biological model to understand a disease, choosing the time window to predict economic trends, and ranking the models in an online data competition. Model selection is such an area that studies principled machinery for scientists to obtain a reliable model for interpretation and prediction purposes. Information criteria represent a broad class of model selection criteria, which have found extensive applications in fields such as economics (Pesaran, 1974), social studies (Raftery, 1995), psychology (Zucchini, 2000), ecology (Johnson & Omland, 2004), epidemiology

(Walsh, 2007), and engineering (Hong et al., 2008). There have been several information criteria proposed from different perspectives. Among them, two representative information criteria, Akaike information criterion (AIC) (Akaike, 1974, 1998) and Bayesian information criterion (BIC) (Schwarz, 1978), are perhaps the most influential and have been built into the standard tool-kits of data science software such as R, STAT, SPSS, and SAS.

In application, since different information criteria typically give different results, one may eventually have to choose one for reaching a final model. Unfortunately, this often puzzles data analysts in multiple aspects: (1) Which one is the best among various information criteria available? (2) How to correctly interpret the obtained model selection results? (3) Is the final model reliable? Indeed, when and how to use information criteria appropriately is a complex problem, and the uncertainty of model selection should be properly assessed. In the past few decades, researchers have found that the successful use of each method may depend on many factors, including the underlying data-generating process, postulated models (in particular, how close they are to the data-generating process), sample size, and evaluation metrics. Overall, it is impossible to have a single criterion that fits all circumstances. As such, it is crucial to understand the theoretical foundations of information criteria and their practical implications.

We see the emerging need for reviewing information criteria and related methods with increasing data and modeling problems. While there are several articles on the overview of general model selection techniques (see, e.g., Ding et al. (2018b) and the references therein), a summary of foundational aspects of information criteria is relatively lacking. Also, there have been new developments in model selection and assessment based on information criteria. These have motivated our review article on information criteria, which aims to enrich its understanding of the following elements:

- Model selection objectives, evaluation metrics, challenges, and insights.
- Insights into the theoretical properties of AIC and BIC.
- Connections between information criteria and other related methods.
- Recent research advancements on the use of information criteria.
- Clarification of misleading folklores and practical guidelines.

The information criteria were historically developed for traditional statistical models, such as linear regression and autoregression models. Here, we use "traditional statistical models" to generally refer to candidate models that are basically pre-determined, have a relatively small set of continuously-valued unknown parameters, satisfy some regularity conditions (e.g., smoothness with respect to the parameters), and have many existing well-understood technical analyses. Needless to say, semi-automated and highly complex machine learning procedures, which we generally refer to as "blackbox methods," have become increasingly popular in many applications. For example, nowadays, tree ensembles and deep neural networks have been frequently applied to regression and classification problems, especially when both the sample size and input dimension are considerable. Nevertheless, studying model selection in traditional statistical models is still critically important for the following reasons. First, traditional statistical models, with advantages in interpretability and uncertainty quantification, continue to be appropriate tools for scientific understanding beyond pure prediction. For example, Ye et al. (2018) illustrated how linear regression could provide more reliable descriptions of variable importance than random forest. Second, when the sample size is relatively small (e.g., hundreds or thousands), the traditional models often perform better or much better in prediction (Bartol et al., 2022; Dudoit et al., 2002). Third, the developed understanding based on traditional statistical models may apply to blackbox methods since their underlying principles such as the bias-variance tradeoff may be the same. Finally, choosing between a regular statistical model and a blackbox method when both are applicable for prediction is also a model selection problem, which can be addressed by proper cross-validations, as will be reviewed later. This provides an angle into how model selection principles initially developed for traditional models can leverage the power of blackbox methods.

We emphasize that considering traditional statistical models as candidate models does not mean that the underlying data-generating process is restricted to belong to any candidate model. In other words, the models may not be well-specified—an essential aspect of model selection that will be presented in this review.

The outline of the article is as follows. Section 2 introduces the model selection settings, including common objectives, desirable properties, and an example to illustrate the key challenges. Sections 3 and 4 introduce AIC and BIC, respectively, which have played foundational roles in developing model selection literature. We will review their standard forms, theoretical properties, and intuitive explanations of asymptotic behaviors. Section 5 introduces some other information criteria that represent different perspectives. Section 6 reviews the connections between information criteria and other methods for model selection. Section 7 discusses recent research advancements in bridging the gap between AIC and BIC (and their variants). Section 8 presents information criteria for high-dimensional regression.

Section 9 focuses on the model selection uncertainty and introduces information criteria-based tools to assess and control the uncertainty. Section 10 clarifies some common misconceptions about information criteria. Section 11 provides concluding remarks of this review article.

2 | MODEL SELECTION FRAMEWORKS

Suppose we observe data $\mathcal{D}$ with an unknown probability density function p_* with respect to a -finite measure. In this work, any specification of a set of probability density functions for $\mathcal{D}$ is called a model. To estimate p_*, we consider a pre-determined set of parametric models $\{\mathscr{M}_m : m \in \mathbb{M}\}$, where $\mathbb{M}$ is an index set, $\mathscr{M}_m$ can be written as $\{p_{\theta_m} : \theta_m \in \mathscr{H}_m\}$, θ_m is a continuously valued parameter that uniquely determines the function p_{θ_m}, and $\mathscr{H}_m$ is a d_m-dimensional parameter space of θ_m. Different models may overlap in the sense of sharing some density functions. Given a specific model, θ_m can be estimated by a proper method, such as the maximum likelihood estimation and the method of moments. The core interest of model selection is to choose $\mathscr{M}_m$ ($m \in \mathbb{M}$) most suitable for the objectives for modeling the data $\mathcal{D}$. Next, we will elaborate on some common model selection problems in Section 2.1, discuss different objectives and evaluation metrics in Section 2.2, and highlight the challenges through a specific example in Section 2.3.

2.1 | Common problem formulations for model selection

We first consider model selection in regression models. Suppose

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $y_i \in \mathbb{R}$ is the response, $x_i \in \mathbb{R}^d$ is the vector of the predictor variables, $f : \mathbb{R}^d \to \mathbb{R}$ is the underlying regression function, and ε_i represents independent random noise with variance σ^2. Since in regression problems, one typically does not study the distribution of the predictor variables, a postulated model $\mathscr{M}_m \triangleq \{f_{\theta_m} : \theta_m \in \mathscr{H}_m\}$ often focuses on representing the regression function together with a specification of the error distribution. Given the observed data $D_n \triangleq \{z_i\}_{i=1}^n$, where $z_i \triangleq (y_i, x_i)$, the estimated parameters of $\mathscr{M}_m$ are usually obtained by solving

$$\widehat{\theta}_m \triangleq \arg\min_{\theta_m \in \mathscr{H}_m} \sum_{i=1}^{n} s(f_{\theta_m}, z_i), \tag{2}$$

where s is a proper scoring function, such as the quadratic loss $s : (f, [y, x]) \mapsto (f(x) - y)^2$. The performance of the models is evaluated by the out-sample prediction loss for a future observation z that follows p_*:

$$\mathscr{L}_n(m) \triangleq \mathbb{E}_*\left(s\left(f_{\widehat{\theta}_m}, z\right) \mid D_n\right) - \sigma^2, \tag{3}$$

where $\mathbb{E}_*$ denotes the expectation with respect to the data-generating distribution (random-design settings). Here, the uncontrollable future noise variance is subtracted to sharpen the evaluation metric in distinguishing different estimators. In the case of quadratic loss, we have

$$\mathscr{L}_n(m) \triangleq \mathbb{E}_*\left(\left(y - f_{\widehat{\theta}_m}(x)\right)^2 | D_n\right) - \sigma^2 = \mathbb{E}_*\left(\left(f(x) - f_{\widehat{\theta}_m}(x)\right)^2 | D_n\right),$$

where x is treated as a random vector. For fixed-design regression settings, we consider the estimation loss (in terms of estimating the regression function at fixed design points):

$$\mathscr{L}_n(m) \triangleq n^{-1} \sum_{i=1}^{n} \left(f(x_i) - f_{\widehat{\theta}_m}(x_i)\right)^2. \tag{4}$$

Next, we introduce two other commonly studied model selection settings: order selection in time series and model selection in density estimation. In history, many model selection methods have been initially developed for selecting the lag order in autoregressive models. More specifically, suppose the data-generating process is $y_t = \sum_{i=1}^{\infty} \beta_i y_{t-i} + \varepsilon_t$, where ε_t is random noise with variance σ^2. A candidate model $\mathscr{M}_m$ may be formulated as

$$y_t = f_{\theta_m}\left(y_{t-1}, \ldots, y_{t-d_m}\right) + \varepsilon_t,$$

where $f_{\theta_m}\left(y_{t-1}, \ldots, y_{t-d_m}\right) \triangleq \sum_{k=1}^{d_m} \theta_{m,k} \cdot y_{t-k}$, and $\theta_m \triangleq \left[\theta_{m,1}, \ldots, \theta_{m,d_m}\right]^{\mathrm{T}} \in \mathscr{H}_m \subseteq \mathbb{R}^{d_m}$ is the parameter vector that can be estimated by the quadratic loss for time series:

$$s : \left(f_{\theta_m}, \left[y_t, \ldots, y_{t-d_m}\right]\right) \mapsto \left(y_t - f_{\theta_m}\left(y_{t-1}, \ldots, y_{t-d_m}\right)\right)^2.$$

The performance of a time series model can be evaluated by the one-step prediction loss:

$$\mathscr{L}_n(m) \triangleq \mathbb{E}_*\left(s\left(f_{\widehat{\theta}_m}, \left[y_{n+1}, \ldots, y_{n+1-d_m}\right]\right) | y_1, \ldots, y_n\right) - \sigma^2.$$

In order selection problems, we consider the set of nested models $\mathscr{M}_1, \ldots, \mathscr{M}_{d_{\max}}$ with increasing orders where $1 < d_{\max} < n$.

Model selection is generally used for density estimation where $y_1, \ldots, y_n$ are independent and identically distributed (IID) observations following a density function p_*. We postulate a set of candidate models $\{\mathscr{M}_m : m \in \mathbb{M}\}$, where $\mathscr{M}_m \triangleq \left\{p_{\theta_m} : \theta_m \in \mathscr{H}_m\right\}$ represents a parametric family of density functions $\left\{p_{\theta_m}, \theta_m \in \mathscr{H}_m\right\}$ with $\mathscr{H}_m \subseteq \mathbb{R}^{d_m}$. The parameters are often estimated by

$$\widehat{\theta}_m \triangleq \arg\min_{\theta_m \in \mathscr{H}_m} \sum_{i=1}^{n} s\left(p_{\theta_m}, y_i\right),$$

where $s(p,y) : (p,y) \mapsto -\log p(y)$ is the logarithmic loss. The performance of $\mathscr{M}_m$ can be evaluated by $\mathscr{L}_n(m) \triangleq \mathbb{E}_*\left(s\left(p_{\widehat{\theta}_m}, y\right) | y_1, \ldots, y_n\right)$ for a future observation y, which is equivalent to the Kullback–Leibler loss

$$\mathbb{E}_*\left(\log \frac{p_*(y)}{p_{\widehat{\theta}_m}(y)} | y_1, \ldots, y_n\right).$$

2.2 | Model selection objectives and evaluations

In practice, there may be various goals of model selection. We roughly categorize them into two types: (1) *prediction* that aims at selecting the best model in terms of the out-sample prediction loss (e.g., minimize the one-step prediction loss in time series modeling), and (2) *inference* that focuses on improving the explainability and interpretability of the unknown data-generating process (e.g., selecting essential variables). For prediction, it may not be a concern if the model selection result keeps switching between some candidate models with similar performances when the sample size or signal-to-noise ratio is slightly changed (see Zhang et al. (2023) for an example). In contrast, for interpretation, the selection result must be stable to characterize the data-generating process reliably. Let $\widehat{m}$ denote the index of the model obtained from a model selection method. In line with the above goals, the following notions characterize the desirable model selection properties from different perspectives.

To simplify the notation, let us consider a regression setting and let f denote the underlying regression function and $\mathbb{M}$ the set of postulated candidate models. We define the set of "well-specified" models to be

$$\mathbb{M}_w \triangleq \{m \in \mathbb{M} : \exists \theta_m \in \mathscr{H}_m, \ s.t. \ f = f_{\theta_m}\}, \tag{5}$$

which may contain none, one, or more elements. For example, suppose the considered candidate models are polynomial regression functions of a variable $x \in \mathbb{R}$ with degrees ranging from 1 to $d_{\max}$, namely, $f_{\theta_m}(x) = \sum_{i=0}^{d_m} \theta_{m,i} \cdot x^i$ with $d_m = 1, ..., d_{\max}$. Then, $\mathbb{M}_w$ is empty if the underlying data-generating regression function is $f(x) = e^x$, $\mathbb{M}_w$ contains one element if $f(x) = x^{d_{\max}}$, and $\mathbb{M}_w$ contains multiple elements if $f(x) = x^d$ with $d < d_{\max}$. When $\mathbb{M}_w$ contains more than one element, it is desirable to identify the most parsimonious model, denoted by $\mathscr{M}_{m_*}$, namely $m_* = \arg\min_{m \in \mathbb{M}_w} d_m$. Such a model is often called the *true model* or *data-generating model*.

For the following definition, we assume $\mathscr{M}_{m_*}$ exists and is unique. The property of consistency aims to select a reliable model for inference purposes.

Definition 1. (Consistency). A model selection method is consistent if it selects a model $\widehat{m}$ that satisfies

$$\mathbb{P}(\widehat{m} = m_*) \to 1, as\, n \to \infty.$$

From a prediction perspective, we may not require the existence of $\mathscr{M}_{m_*}$. Instead, we may only need the selected model to predict as accurately as the theoretically best model (asymptotically), even if the selection result itself may not be stable.

Definition 2. (Efficiency). A model selection method is asymptotically efficient if it selects a model $\widehat{m}$ that satisfies

$$\frac{\mathscr{L}_n(\widehat{m})}{\min_{m \in \mathbb{M}} \mathscr{L}_n(m)} \to_p 1 \, as\, n \to \infty.$$

Whether a model selection method has the above properties depends on the underlying data-generating distribution and the candidate model class. Such dependence can be roughly categorized into the following two scenarios.

Definition 3. (Parametric scenario). In a parametric scenario, there is at least one well-specified candidate model, in the sense that $\mathbb{M}_w$ defined in Equation (5) is not empty.

A cautious reader may question whether any theoretical understanding of model selection under this definition is of practical relevance if "all models are wrong." It is conceivable that we are rarely in a "true" parametric scenario (except, e.g., in well-controlled physical/chemical studies, where parametric models are known to be reliable). Nevertheless, in practice, the developed understanding does not necessarily require the underlying data-generating distribution to sit in the model class in the strict sense. The theoretical knowledge acquired from a parametric scenario may also apply to a "practically parametric" scenario, depending on the sample size and underlying distribution, which we will discuss in Section 10.2.

Definition 4. (Nonparametric scenario). In a nonparametric scenario, it is assumed that the data are generated in a way that cannot be fully characterized by any candidate models, namely $\mathbb{M}_w = \emptyset$, and the models may provide better and better approximations to the underlying data-generating process when the model complexity increases.

We will illustrate that an efficient information criterion in a nonparametric scenario may not be efficient in a parametric scenario and vice versa. It is worth noting that the above notions of parametric and nonparametric scenarios correspond respectively to the M-closed and M-open views in a Bayesian framework (see, e.g., Section 6.1.2 of Bernardo and Smith (2009) for more details).

Next, we provide an alternative notion, "minimax rate optimality," to evaluate model selection from a prediction perspective. Compared with the above definition of efficiency, this definition will focus on the worst-case scenario. More specifically, while efficiency characterizes a model selection method by assuming a fixed underlying data-generating distribution, the minimax rate optimality provides a uniform guarantee of its performance on a set of possible data-generating distributions. We will introduce formal notions by focusing on regression models.

In the parametric scenario, given model $m \in \mathbb{M}$, we define the minimax risk

$$\mathscr{R}_n(m) \triangleq \inf_{\widehat{f}} \sup_{\theta_m \in \mathscr{H}_m} \mathbb{E}_{\theta_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f_{\theta_m}(x_i) - \widehat{f}(x_i) \right)^2 \right\},$$

where $\widehat{f}$ is over all estimators of f based on the data, and $\mathbb{E}_{\theta_m}$ denotes the expectation with respect to θ_m being the true regression parameter. If m_* were known, for the estimation of f, we can achieve the smallest worst-case risk by using the minimax estimator of m_*. With m_* unknown and $\widehat{m}$ selected, we hope the resulting estimator converges at the minimax rate $\mathscr{R}_n(m_*)$.

In the nonparametric scenario, consider an infinite-dimensional class of regression functions $\mathscr{F}$ as the target collection of regression functions to learn. To that end, a countable list of finite-dimensional models is used to approximate the target functions. The minimax risk for estimating $f \in \mathscr{F}$ is defined as

$$\mathscr{R}_n(\mathscr{F}) \triangleq \inf_{\widehat{f}} \sup_{f \in \mathscr{F}} \mathbb{E}_f \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \widehat{f}(x_i) \right)^2 \right\},$$

where $\widehat{f}$ is over all estimators of f based on the data, and $\mathbb{E}_f$ denotes the expectation with respect to f being the true regression function. For the selected approximating model $\widehat{m}$, we hope the associated estimator achieves the same convergence rate as $\mathscr{R}_n(\mathscr{F})$.

Definition 5. (Minimax rate optimality). A model selection method is minimax rate optimal if it selects a model $\widehat{m}$ that satisfies:

1. In the parametric scenario, for all $m \in \mathbb{M}$,

$$\sup_{\theta_m \in \mathscr{H}_m} \mathbb{E}_{\theta_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f_{\theta_m}(x_i) - f_{\widehat{\theta}_{\widehat{m}}}(x_i) \right)^2 \right\} \leq C \cdot \mathscr{R}_n(m)$$

for some constant $C > 0$ for all $n \geq 1$.

2. In the nonparametric scenario,

$$\sup_{f \in \mathscr{F}} \mathbb{E}_f \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - f_{\widehat{\theta}_{\widehat{m}}}(x_i) \right)^2 \right\} \leq \widetilde{C} \cdot \mathscr{R}_n(\mathscr{F})$$

for some constant $\widetilde{C} > 0$ for all $n \geq 1$.

We focused on regression with a fixed design under the squared error loss in the above definition. Proper modifications are needed for random design and other loss functions.

2.3 | Fundamental challenges through an example

To highlight the fundamental challenges in model selection, we consider a very specific setting where the underlying data-generating process is a fixed design linear regression with the underlying function $f(x)$, $x \in \mathbb{R}^d$, and independent noise ε with a known variance σ^2. Let $\boldsymbol{y}_n \triangleq [y_1,...,y_n]^{\mathrm{T}}$, $\boldsymbol{f}_n \triangleq [f(x_1),...,f(x_n)]^{\mathrm{T}}$, and $\boldsymbol{\varepsilon}_n \triangleq [\varepsilon_1,...,\varepsilon_n]^{\mathrm{T}}$. We consider the candidate models $\mathscr{M}_m$ consisting of the regression functions $f_{\theta_m}(x) = x_m^{\mathrm{T}}\theta_m$ where x_m is a subset of variables in x with size $d_m \leq d$ and θ_m is the unknown parameter. Note that the true f may or may not belong to any of the candidate models. Denote the projection matrix of $\boldsymbol{y}_n$ onto the column span of the predictors in model $\mathscr{M}_m$ by $F_n(m)$ and let $\boldsymbol{f}_{\widehat{\theta}_m} \triangleq F_n(m)\boldsymbol{y}_n$ be the fitted values from $\mathscr{M}_m$. Then, the expected prediction loss (also known as the risk under the average squared error), up to the difference of a factor n, can be written as

$$\mathbb{E}_* \|\boldsymbol{f}_n - \boldsymbol{f}_{\widehat{\theta}_m}\|_2^2 = \mathrm{bias}^2 + \mathrm{variance}, \tag{6}$$

where $\mathbb{E}_*$ is the expectation taken under the data-generating distribution with respect to ε_n, $\|\cdot\|_2$ denotes the ℓ_2 norm of vectors, bias $\triangleq \|\boldsymbol{f}_n - F_n(m)\boldsymbol{f}_n\|_2$, and variance $\triangleq \mathbb{E}_*\left(\varepsilon_n^{\mathrm{T}}F_n(m)\varepsilon_n\right) = d_m\sigma^2$. It indicates that d_m should be properly chosen since a small d_m may lead to a large bias due to missing useful predictors and a large d_m will increase the variance. However, we do not have future observations to evaluate the out-sample prediction loss. Instead, we only have access to the in-sample loss $RSS_m \triangleq \|\boldsymbol{y}_n - \boldsymbol{f}_{\widehat{\theta}_m}\|_2^2$, which favors an over-complicated model that fits not only $\boldsymbol{f}_n$ but together with the noise ε_n. More specifically, we have

$$\begin{aligned}\mathrm{RSS}_m &= \|\varepsilon_n\|_2^2 + \|\boldsymbol{f}_n - \boldsymbol{f}_{\widehat{\theta}_m}\|_2^2 + 2\varepsilon_n^{\mathrm{T}}\left(\boldsymbol{f}_n - \boldsymbol{f}_{\widehat{\theta}_m}\right)\\ &= \|\varepsilon_n\|_2^2 + \|\boldsymbol{f}_n - \boldsymbol{f}_{\widehat{\theta}_m}\|_2^2 - 2\varepsilon_n^{\mathrm{T}}F_n(m)\varepsilon_n + 2\varepsilon_n^{\mathrm{T}}(I - F_n(m))\boldsymbol{f}_n.\end{aligned}$$

Note that the last term $2\varepsilon_n^{\mathrm{T}}(I - F_n(m))\boldsymbol{f}_n$ has a zero mean. The formula suggests that if our goal is to estimate the prediction loss from RSS_m (up to a quantity that does not depend on m), we may add a bias correction term

$$2\mathbb{E}_*\left(\varepsilon_n^{\mathrm{T}}F_n(m)\varepsilon_n - \varepsilon_n^{\mathrm{T}}(I - F_n(m))\boldsymbol{f}_n\right) = 2\mathbb{E}_*\left(\varepsilon_n^{\mathrm{T}}F_n(m)\varepsilon_n\right) = 2d_m\sigma^2. \tag{7}$$

Nevertheless, having a reasonable estimate of the prediction loss does not necessarily guarantee a selection of the best model. The intuition is that for selection purposes, we need to compare the relative performance between models instead of their absolute performance. For example, consider the setting where two nested models $\mathscr{M}_{m_1}$ and $\mathscr{M}_{m_2}$ ($d_{m_2} > d_{m_1}$) are both well-specified. It can be shown that $\mathscr{M}_{m_1}$ has a smaller expected loss value than $\mathscr{M}_{m_2}$ thanks to its fewer parameters to estimate. Also, it can be verified that in this case

$$\mathrm{RSS}_m = \|\varepsilon_n\|_2^2 - \varepsilon_n^{\mathrm{T}}F_n(m)\varepsilon_n, \quad m \in \{m_1, m_2\}, \tag{8}$$

and thus

$$\mathrm{RSS}_{m_1} - \mathrm{RSS}_{m_2} = \varepsilon_n^{\mathrm{T}}(F_n(m_2) - F_n(m_1))\varepsilon_n \sim \chi^2_{d_{m_2} - d_{m_1}}$$

with a positive probability of being larger than the bias correction term $2(d_{m_2} - d_{m_1})\sigma^2$ obtained in Equation (7), where χ_k^2 denotes the chi-squared distribution with k degrees of freedom. Consequently, we cannot attain selection consistency or efficiency by the penalty $2d_m\sigma^2$ for fixed d_{m_2} and d_{m_1}. We need a model complexity penalty term that goes to infinity to sufficiently discourage the choice of the larger model. In summary, we see two major challenges in model selection. First, we may need to approximate the out-sample prediction loss from the sample already used for estimating the candidate models and use that as a basis to select models. Second, an accurate approximation of the prediction loss may not directly lead to selecting the best model since the selection results may also depend on the relative performances between the candidate models.

3 | AKAIKE INFORMATION CRITERION

We will first introduce the Akaike information criterion (AIC) (Akaike, 1974, 1998), which is the earliest information criterion with a profound influence on the development of model selection techniques. The original form of AIC selects a model $m \in \mathbb{M}$ for the density estimation problem by minimizing

$$\mathrm{AIC}_m \triangleq -2\sum_{i=1}^{n}\log p_{\widehat{\theta}_m}(y_i)+2d_m, \tag{9}$$

where $\widehat{\theta}_m$ denotes the maximum likelihood estimate under the model m. The above formula can be extended to regression and time series settings by replacing $p_{\widehat{\theta}_m}$ by the conditional density of $y \mid x$ or $y_t \mid y_{t-1},\ldots,y_{t-d_m}$ estimated from the model, respectively.

To develop insights into Formula (9), let us consider the Kullback–Leibler (KL) divergence

$$D_{KL}\left(p_{\theta_*}\middle\|p_{\theta_m}\right)=\mathbb{E}_*\left(\log\left(p_{\theta_*}(y)\right)\right)-\mathbb{E}_*\left(\log\left(p_{\theta_m}(y)\right)\right)=-\mathbb{E}_*\left(\log\left(p_{\theta_m}(y)\right)\right)+c_*,$$

where c_* does not depend on m. Then, minimizing the KL divergence is equivalent to minimizing the expected logarithmic loss for each model m. The quantity of AIC_m (after re-scaling) can be regarded as an estimate of $-\mathbb{E}_*\left(\log\left(p_{\theta_m}(y)\right)\right)$ using its sample analog $n^{-1}\sum_{i=1}^{n}\log\left(p_{\widehat{\theta}_m}(y_i)\right)$ plus a bias-correction term $n^{-1}d_m$. Notably, the bias-correction term is needed because $n^{-1}\sum_{i=1}^{n}\log\left(p_{\widehat{\theta}_m}(y_i)\right)$ has overused the observed data (once for the sample average and once for estimating $\widehat{\theta}_m$). To see this, we assume $\mathscr{M}_m$ is well-specified and use second-order Taylor expansion to obtain

$$\mathbb{E}_*\left(\log\left(p_{\widehat{\theta}_m}(y)\right)\right)\approx\mathbb{E}_*\left(\log\left(p_{\theta_*}(y)\right)-\frac{1}{2n}G_n^{\mathrm{T}}JG_n\right), \tag{10}$$

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(p_{\widehat{\theta}_m}(y_i)\right)\approx\frac{1}{n}\sum_{i=1}^{n}\log\left(p_{\theta_*}(y_i)\right)+A-\frac{1}{2n}G_n^{\mathrm{T}}JG_n, \tag{11}$$

where $G_n \triangleq \sqrt{n}\left(\widehat{\theta}_m-\theta_*\right)$ is expected to converge to $\mathcal{N}(0,J^{-1})$ in distribution as $n\to\infty$, $J \triangleq \mathbb{E}_*\left\{(\nabla_\theta \log p_{\theta^*}(y))(\nabla_\theta \log p_{\theta^*}(y))^{\mathrm{T}}\right\}$, and

$$A \triangleq n^{-3/2}G_n\cdot\sum_{i=1}^{n}\nabla_\theta \log p_{\theta^*}(y_i)\approx n^{-1}G_n^{\mathrm{T}}JG_n. \tag{12}$$

Here, $\mathcal{N}$ denotes a Gaussian distribution and ∇_θ denotes the gradient with respect to θ. From asymptotic normality of $\widehat{\theta}_m$, it can be verified that $G_n^{\mathrm{T}}JG_n$ follows the chi-squared distribution with degrees of freedom d_m and $\mathbb{E}_*\left(G_n^{\mathrm{T}}JG_n\right)=d_m$. Therefore, by taking the expectation of the right-hand side of (11) and comparing it with (10), we have

$$\mathbb{E}_*\left(\log\left(p_{\widehat{\theta}_m}(y)\right)\right)\approx n^{-1}\sum_{i=1}^{n}\log\left(p_{\widehat{\theta}_m}(y_i)\right)-n^{-1}\cdot d_m,$$

which leads to the formula of AIC_m that aims to approximate $-2n\mathbb{E}_*\left(\log\left(p_{\widehat{\theta}_m}(y)\right)\right)$.

Next, we will first illustrate the properties of AIC under a linear regression setting and then present its properties under general settings.

3.1 | AIC for regression

For regression models, AIC is often known in the following particular form

$$\mathrm{AIC}_m \triangleq n\log\left(n^{-1}\mathrm{RSS}_m\right)+2d_m, \tag{13}$$

which is derived from (9) by assuming $y\,|\,x\sim\mathcal{N}(f(x),\sigma^2)$. In fact, it can be calculated that

$$\mathrm{AIC}_m = n\log(2\pi)+n\log\widehat{\sigma}_m^2+\frac{\mathrm{RSS}_m}{\widehat{\sigma}_m^2}+2d_m. \tag{14}$$

where $\widehat{\sigma}_m^2=n^{-1}\mathrm{RSS}_m$. By rearranging it and dropping model-independent terms, we obtain the formula in (13).

To illustrate the properties of AIC in terms of consistency and efficiency, we consider the linear regression example in Section 2.3 with known σ^2 and an alternative form of AIC:

$$\mathrm{AIC}_m \triangleq \mathrm{RSS}_m+2d_m\sigma^2, \tag{15}$$

which is obtained by multiplying the right-hand side of (14) by σ^2 and dropping model-independent terms. We let $\mathbb{M}=\{m_1,m_2\}$ and focus on the following three scenarios.

1. The parametric scenario with both candidate models well-specified and $\mathcal{M}_{m_1}$ being nested in $\mathcal{M}_{m_2}$: Here, "nested" means any element in $\mathcal{M}_{m_1}$ also belongs to $\mathcal{M}_{m_2}$. Based on the result in Equation (8)

   $$\mathrm{AIC}_{m_2}-\mathrm{AIC}_{m_1}=-\varepsilon_n^{\mathrm{T}}(F_n(m_2)-F_n(m_1))\varepsilon_n+2(d_{m_2}-d_{m_1})\sigma^2.$$

   Therefore, the probability of the event $\{\mathrm{AIC}_{m_2}-\mathrm{AIC}_{m_1}<0\}$ is lower bounded by a positive constant, and AIC cannot choose the data-generating model with probability going to one.
2. The parametric scenario when $\mathcal{M}_{m_1}$ is misspecified and $\mathcal{M}_{m_2}$ is well-specified: By Equation (8)

   $$\mathrm{AIC}_{m_2}=\|\varepsilon_n\|_2^2-\varepsilon_n^{\mathrm{T}}F_n(m_2)\varepsilon_n+2d_{m_2}=\|\varepsilon_n\|_2^2+O_p(d_{m_2}). \tag{16}$$

   Also, it can be shown under mild conditions (Shao, 1993) that

   $$\mathrm{RSS}_m=\|\varepsilon_n\|_2^2+\|\boldsymbol{f}_n-\boldsymbol{f}_{\widehat{\theta}_m}\|_2^2-2\sigma^2d_m+o_p\left(\|\boldsymbol{f}_n-\boldsymbol{f}_{\widehat{\theta}_m}\|_2^2\right) \tag{17}$$

   for misspecified m, and $\|\boldsymbol{f}_n-\boldsymbol{f}_{\widehat{\theta}_{m_1}}\|_2^2\to_p\infty$ goes to infinity at the rate of n as $n\to\infty$. Thus,

   $$\mathrm{AIC}_{m_1}=\|\varepsilon_n\|_2^2+\|\boldsymbol{f}_n-\boldsymbol{f}_{\widehat{\theta}_{m_1}}\|_2^2+o_p\left(\|\boldsymbol{f}_n-\boldsymbol{f}_{\widehat{\theta}_{m_1}}\|_2^2\right). \tag{18}$$

   Combining Equations (16) and (18) gives $\mathrm{AIC}_{m_1}-\mathrm{AIC}_{m_2}\to_p\infty$, which further implies that AIC favors the well-specified model with probability going to one as $n\to\infty$.
3. The nonparametric scenario where both models are misspecified: In this case, the form in Equation (18) applies to both AIC_{m_1} and AIC_{m_2} (with m_1 replaced with m_2). Therefore, comparing AIC_{m_1} and AIC_{m_2} is asymptotically equivalent with comparing $\|\boldsymbol{f}_n-\boldsymbol{f}_{\widehat{\theta}_{m_1}}\|_2^2$ and $\|\boldsymbol{f}_n-\boldsymbol{f}_{\widehat{\theta}_{m_2}}\|_2^2$, namely, the out-sample prediction losses. This suggests that AIC is efficient.

From the above, we see that as long as there are at least two correct candidate models, AIC has a non-vanishing probability of selecting the larger one. Hence, AIC is inconsistent in selection in general.

To provide insights into the minimax rate optimality, which is less transparent to see, we first restrict the setting to the simple linear regression where $\mathscr{M}_{m_1}$ is the model with intercept only and $\mathscr{M}_{m_2}$ has one predictor. Based on theorem 2 of Yang (2007), it can be seen that for a criterion that minimizes $\mathrm{RSS}_m + \lambda_n d_m \sigma^2$, it is minimax rate optimal if and only if λ_n is bounded in the context. It implies the minimax optimality of AIC in such a setting. More generally, due to its bias-correction nature, AIC is minimax rate optimal for both parametric and nonparametric situations. We refer to proposition 1 in Yang (2005) for such a result and section 4 in Yang and Barron (1999) for an understanding of why the bias correction of AIC leads to minimax rate optimality for nonparametric regression.

3.2 | AIC for general models

In general settings, AIC has similar properties to those in Section 3.1. Under a more general setting of the nonparametric scenario where the size of $\mathbb{M}$ increases as $n \to \infty$, it has been shown that AIC is asymptotically efficient (Ing et al., 2012; Ing & Wei, 2005; Shibata, 1980, 1981). However, it is neither efficient nor consistent in the parametric scenario when there is more than one well-specified candidate model (Nishii, 1984; Shao, 1997). Also, AIC is minimax rate optimal in both parametric and nonparametric scenarios (Barron et al., 1999; Yang, 2005; Yang & Barron, 1998).

4 | BAYESIAN INFORMATION CRITERION

The Bayesian information criterion (BIC) (Schwarz, 1978) is another cornerstone of information criteria. For the density estimation problem, BIC selects a model that minimizes

$$\mathrm{BIC}_m \triangleq -2\sum_{i=1}^{n} \log p_{\widehat{\theta}_m}(y_i) + d_m \log n, \tag{19}$$

where $\widehat{\theta}_m$ denotes the maximum likelihood estimate under the model m. Similar to AIC, BIC can be applied to linear regression and time series under distributional assumptions on ε. It has a nice Bayesian interpretation of choosing the model with the largest posterior probability

$$p(\mathscr{M}_m | D_n) = p(D_n | \mathscr{M}_m) \cdot \frac{p(\mathscr{M}_m)}{p(D_n)}, \text{where} \tag{20}$$

$$p(D_n | \mathscr{M}_m) = \int_{\mathbb{R}^{d_m}} \exp\left(\sum_{i=1}^{n} \log(p_{\theta_m}(y_i))\right) p_m(\theta_m) d\theta_m, \tag{21}$$

and $p_m(\theta_m)$ denotes the prior density of θ_m under model $\mathscr{M}_m$. To see the interpretation, we use Laplace approximation to approximate $p(D_n | \mathscr{M}_m)$ as

$$\exp\left(\sum_{i=1}^{n} \log\left(p_{\widehat{\theta}_m}(y_i)\right)\right) \cdot p_m\left(\widehat{\theta}_m\right) \cdot \int_{\mathbb{R}^{d_m}} \exp\left(-\frac{1}{2}\left(\widehat{\theta}_m - \theta_m\right)^{\mathrm{T}} B\left(\widehat{\theta}_m - \theta_m\right)\right)) d\theta_m$$

$$= \exp\left(\sum_{i=1}^{n} \log\left(p_{\widehat{\theta}_m}(y_i)\right)\right) \cdot p_m\left(\widehat{\theta}_m\right) \cdot (2\pi)^{d_m/2} \cdot \det(B)^{-1/2},$$

where $B \triangleq -\sum_{i=1}^{n} \nabla_\theta^2 \log p_{\widehat{\theta}_m}(y_i)$, and $\det(\cdot)$ denotes the determinant of a matrix. Here, with the sample size suitably large, the prior $p_m(\theta_m)$ is expected to play a less important role compared with the likelihood and thus $\widehat{\theta}_m$ approximately minimizes the integrand in (21). As n becomes larger, B/n gets closer to a constant matrix, say B_0, under some regularity conditions, so $\det(B)^{-1/2} \approx n^{-d_m/2} \det(B_0)^{-1/2}$. Therefore, taking a logarithm of $p(D_n | \mathscr{M}_m)$ gives

$$\log p(D_n|\mathscr{M}_m) = -\frac{1}{2}\mathrm{BIC}_m + o_p(d_m \log n),$$

where $o_p(d_m \log n)$ denotes a term asymptotically negligible compared with the BIC penalty $d_m \log n$. Thus, it follows from Equation (20) that

$$\log p(\mathscr{M}_m|D_n) = -\frac{1}{2}\mathrm{BIC}_m + o_p(d_m \log n) - \log p(D_n).$$

Since $\log p(D_n)$ is a term that does not depend on the model, minimizing BIC is asymptotically close to choosing the model with the largest posterior probability. The above argument also applies to the selection principle based on maximizing the marginal likelihood or Bayes factors, namely based on directly maximizing $p(D_n|\mathscr{M}_m)$ over $m \in \mathbb{M}$. As will be discussed in Section 8, the prior on the models needs to be brought to light in case of addressing exponentially many candidate models. We will introduce more properties of BIC in the following subsections.

4.1 | BIC for regression

Like AIC, BIC in regression takes the form of

$$\mathrm{BIC}_m \triangleq n \log\left(n^{-1}\mathrm{RSS}_m\right) + d_m \log n.$$

To illustrate its properties in terms of consistency and efficiency, we consider the same parametric settings in Section 3.1 for AIC, where BIC has the alternative form $\mathrm{BIC}_m \triangleq \mathrm{RSS}_m + d_m \sigma^2 \log n$.

1. The parametric scenario with $\mathscr{M}_{m_1}$ nested in $\mathscr{M}_{m_2}$ and both being correct. We have

$$\mathrm{BIC}_{m_2} - \mathrm{BIC}_{m_1} = -\varepsilon_n^{\mathrm{T}}(F_n(m_2) - F_n(m_1))\varepsilon_n + (d_{m_2} - d_{m_1})\sigma^2 \log n.$$

   Since the term $(d_{m_2} - d_{m_1})\sigma^2 \log n$ is dominating, BIC selects $\mathscr{M}_{m_1}$ with probability going to one.
2. For the parametric scenario with $\mathscr{M}_{m_1}$ misspecified and $\mathscr{M}_{m_2}$ well-specified, by a similar argument as the point 2 in Section 3.1, BIC selects the well-specified model with probability going to one as $n \to \infty$.

From the above, we see that with the penalty $d_m \log n$, since $\log n \to \infty$ as $n \to \infty$, it prevents over-selection. For minimax rate optimality, recall that we require λ_n defined in Section 3.1 to be upper bounded. Since $\log n$ does not meet this requirement, BIC is not minimax optimal, even for the parametric scenario.

4.2 | BIC for general models

For linear regression, BIC was shown to be consistent in the parametric scenario but not efficient under the nonparametric scenario (Nishii, 1984; Shao, 1997). It is not minimax rate optimal (Foster & George, 1994; Shao, 1997; Yang, 2005).

5 | OTHER INFORMATION CRITERIA

This section reviews some of the most popular information criteria apart from AIC and BIC. We only list some of them that represent different perspectives.

- Takeuchi's information criterion (TIC) by Takeuchi (1976) is a surrogate of AIC for possibly misspecified models. It calculates

$$\mathrm{TIC}_m \triangleq -2\sum_{i=1}^{n} \log p_{\widehat{\theta}_m}(y_i) + 2\mathrm{tr}\left(V_n^{-1}(m) J_n(m)\right),$$

where 'tr' denotes the trace of a matrix,

$$J_n(m) \triangleq \frac{1}{n}\sum_{i=1}^{n} \left(\nabla_\theta \log p_{\widehat{\theta}_m}(y_i)\right)\left(\nabla_\theta \log p_{\widehat{\theta}_n}(y_i)\right)^{\mathrm{T}},$$

$$V_n(m) \triangleq -\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta^2 \log p_{\widehat{\theta}_m}(y_i).$$

The derivation of TIC follows the same argument as in Section 3 except that G_n converges in distribution to $\mathcal{N}(0, V^{-1}JV^{-1})$, where J is as before and $V \triangleq \mathbb{E}_*\left(-\nabla_\theta^2 \log p_{\theta^*}(y)\right)$. Notably, when $\mathcal{M}_m$ is well-specified, $J = V$, and the argument reduces to the same as AIC. The penalty term in TIC_m is a sample-analog of $2\mathrm{tr}(V^{-1}J)$. The asymptotic efficiency of TIC and its generalizations was established by Ding et al. (2021).

- Hannan–Quinn (HQ) information criterion (Hannan & Quinn, 1979) was proposed for order selection in time series. It selects a model that minimizes

$$\mathrm{HQ}_m \triangleq n\log\left(\widehat{\sigma}_m^2\right) + 2\ c\ d_m \log\log n,$$

where $\widehat{\sigma}_m^2$ is the estimated variance from $\mathcal{M}_m$ and c is an arbitrary constant greater than one. It can be shown under some conditions that $\log\log n$ from the penalty term has the slowest possible rate of increase to guarantee strong consistency, in the sense that $\mathcal{M}_{m_*}$ will be eventually selected as $n \to \infty$ with probability one.
- Finite sample corrected AIC (AICc) (Hurvich & Tsai, 1989) adds a correction term $2d_m(d_m+1)/(n-d_m-1)$ to AIC_m by assuming Gaussian noise in linear regressions. When n is large relative to d_m, the correction term is small compared with the AIC penalty, and AICc tends to have the same selection result as AIC.
- The risk inflation criterion (RIC) (Foster & George, 1994) aims to select a linear regression model that minimizes the relative risk inflation

$$\sup_{\theta \in \mathbb{R}^d} \frac{\mathbb{E}_\theta\left\{n^{-1}\sum_{i=1}^{n}\left(f_\theta(x_i) - f_{\widehat{\theta}_m}(x_i)\right)^2\right\}}{\mathbb{E}_\theta\left\{n^{-1}\sum_{i=1}^{n}\left(f_\theta(x_i) - f_{\widehat{\theta}}(x_i)\right)^2\right\}},$$

where d is the total number of predictors, $f_\theta : x \mapsto \theta^{\mathrm{T}}x$ is the regression function with coefficient vector $\theta \in \mathbb{R}^d$, $\widehat{\theta}_m$ denotes the estimated coefficients from the model $\mathcal{M}_m$, $\widehat{\theta}$ denotes the estimated coefficients from the regression model that only includes the predictors with nonzero coefficients in θ, and $\mathbb{E}_\theta$ represents the expectation with respect to the distribution with regression function f_θ. A suggested form of RIC by Foster and George (1994) is to select a model that minimizes $\mathrm{RSS}_m + 2d_m\widehat{\sigma}_d^2 \log d$, where $\widehat{\sigma}_d^2$ is the estimator of the unknown variance using all the d variables. Compared with AIC, its penalty also depends on the size of the largest candidate model (namely d).
- The focused information criterion (FIC) (Claeskens & Hjort, 2003) focuses on the ℓ_2-loss of an estimator of any particular estimand obtained from the model. As such, the optimal candidate model is not necessarily consistent with the one selected from the likelihood-based approach like AIC and BIC. For example, a well-specified model may be undesirable due to the estimation error from considering too many parameters unrelated to the estimand. More specifically, FIC considers the setting where $y_1, \ldots, y_n$ are IID sampled from a density function $p_\beta(y)$, where $\beta = [\theta, \gamma_0 + \delta/\sqrt{n}]$, θ and δ are unknown vectors, and γ_0 is assumed to be known. An estimand is a function of θ and

δ. Each candidate model consists of θ and a subvector of δ. We refer to Claeskens and Hjort (2003) for the detailed formula of FIC.

- The deviance information criterion (DIC) (Spiegelhalter et al., 2002) is developed for Bayesian hierarchical modeling, where the number of parameters is not clearly defined. To measure the model complexity, DIC calculates

$$p_D(m) \triangleq \overline{D(\theta_m)} - D(\overline{\theta}_m),$$

where $D(\theta_m) \triangleq -2\sum_{i=1}^n \log p_{\theta_m}(y_i) + c$ with c being a term taking the same value for all candidate models, and $\overline{D(\theta_m)}$ and $\overline{\theta}_m$ are the posterior means of $D(\theta_m)$ and θ_m, respectively. DIC selects $\mathscr{M}_m$ that minimizes

$$\mathrm{DIC}_m \triangleq D(\overline{\theta}_m) + 2p_D(m).$$

One can show that DIC is closely related to AIC. More specifically, it was shown in (Spiegelhalter et al., 2002) that $p_D(m)$ is approximately the trace of the product of the Fisher information matrix and the posterior covariance matrix. Thus, the Bernstein–von Mises Theorem (e.g., Ghosh & Ramamoorthi, 2003, theorem 1.4.2) implies that $p_D(m)$ is asymptotically the model dimension under some regularity conditions. In addition, the posterior mean $\overline{\theta}_m$ is asymptotically close to the maximum likelihood estimate up to $o_p(n^{-1/2})$ under standard regularity conditions (e.g., Ghosh & Ramamoorthi, 2003, theorem 1.4.3). As such, DIC may be regarded as a Bayesian counterpart of AIC. In applications, $p_D(m)$ is often numerically calculated using Markov chain Monte Carlo techniques.

- Arlot and Massart (2009) proposed a model selection criterion under the least-square regression framework where the penalty is adaptively determined by the data. It chooses $\mathscr{M}_m$ that minimizes

$$\mathrm{crit}(m) \triangleq \mathrm{RSS}_m + K \cdot \mathrm{pen}(m), \tag{22}$$

where $\mathrm{pen}(m) : \mathbb{M} \to \mathbb{R}^+$ is a known and properly chosen penalty function and K is data-driven. For example, one may take $\mathrm{pen}(m) = d_m$, in which case $\mathrm{crit}(m)$ is similar to AIC except for a rescaled penalty. Let $m(K)$ denote the minimum of (22) over $m \in \mathbb{M}$. Arlot and Massart (2009) suggested the following procedure to obtain the best K, say $\widehat{K}$: First choose a $K_{\min} > 0$ such that $d_{m(K)}$ is "large" when $K < K_{\min}$ and "small" when $K > K_{\min}$. Then, choose $\widehat{K}$ to be $2K_{\min}$. This algorithm is based on the "slope heuristic" (Birgé & Massart, 2007). It considers the approximation of the prediction loss $\mathbb{E}_* s\left(f_{\widehat{\theta}_m}, z\right)$ by $\mathrm{RSS}_m + \mathrm{pen}_1(m) + \mathrm{pen}_2(m)$ plus a negligible term, where s denotes the quadratic loss,

$$\mathrm{pen}_1(m) \triangleq \ \mathbb{E}_*\left(s\left(f_{\widehat{\theta}_m}, z\right) - s\left(f_{\theta_m^*}, z\right)\right),$$

$$\mathrm{pen}_2(m) \triangleq \ n^{-1}\sum_{i=1}^n \left(s\left(f_{\theta_m^*}, z_i\right) - s\left(f_{\widehat{\theta}_m}, z_i\right)\right),$$

and θ_m^* minimizes $\mathbb{E}_* s(f_{\theta_m}, z)$ over $\mathscr{H}_m$. It can be shown that $\mathrm{pen}_2(m)$ is a "minimal penalty" for the prediction loss not to blow up and under some modeling framework, for example, regressograms (Tukey et al., 1961), $\mathrm{pen}_1(m) \approx \mathrm{pen}_2(m)$. These motivated the use of a penalty in the form of $2\mathrm{pen}_2(m)$, which corresponds to choosing $\widehat{K} = 2K_{\min}$ for model selection.

- The bridge criterion (BC) (Ding et al., 2018a) is an information criterion developed to combine the strengths of AIC and BIC. BC aims to adaptively attain the properties of BIC in the parametric scenario and AIC in the nonparametric scenario. In contrast to information criteria whose penalty terms are linear in d_m, BC uses a penalty term proportional to the harmonic number of order d_m, which is the sum of the reciprocals of the first d_m positive integers for model m. This penalty is designed to select the most appropriate model in a way adaptive to the underlying scenario, with the following intuition. The penalty term in BC increases with the model dimension, but the rate of increase depends on the model specification. In the parametric scenario, data tend to favor candidate models with relatively small or constant dimensions, where the penalty increases more quickly, similar to BIC. In the nonparametric scenario, competitive models tend to have larger dimensions, where the penalty rises more slowly, similar to AIC. It allows BC to adaptively choose the best model for the data without knowing whether we are in parametric or nonparametric scenarios.

We refer to Ding et al. (2018a) for more detailed interpretations and analysis of how BC works. A suggested form of BC is to select a model m that minimizes

$$-2\sum_{i=1}^{n} \log p_{\widehat{\theta}_m}(y_i) + n^{2/3}\sum_{k=1}^{d_m}\frac{1}{k},$$

where the dimension of model m, d_m, is restricted to be no more than that of the AIC-selected model. BC is the only information criterion known so far to simultaneously attain asymptotic efficiency in both parametric and nonparametric scenarios (and consistency in the parametric scenario).

6 | CONNECTIONS BETWEEN INFORMATION CRITERIA AND OTHER METHODS

We present the connections between the earlier information criteria, cross-validation, penalized regression, minimum description length, and other methods in Sections 6.1, 6.2, 6.3, and 6.4, respectively.

6.1 | Information criteria and cross-validations

We first provide a brief review of cross-validations, which has many connections with information criteria. It is a model-free approach based on splitting the data into two parts: the first part ("training set") for estimating $\widehat{\theta}_m$ and the second part ("test set") for calculating the validation loss

$$CV_m \triangleq n_v^{-1}\sum_{i=1}^{n_v}\left(y_i - f_{\widehat{\theta}_m}(x_i)\right)^2 \tag{23}$$

that approximates the out-sample prediction loss, where (y_i, x_i) with $i = 1, \ldots, n_v$ constitute the test set. To reduce the variation of the validation results, one typically repeats the above data-splitting process many times to obtain an average validation result. Notably, the right-hand side of (23) may involve an additional weighting factor (also known as a targeted CV) (Zhang et al., 2023) to accommodate, for example, new evaluation goals, data distributional shifts, and focused sub-populations.

For example, leave-one-out CV (Allen, 1974; Geisser, 1975; Stone, 1974) uses $n_v = 1$ and requires n possible data splittings. Another popular CV method is the leave-n_v-out CV (Shao, 1993; Zhang, 1993) that takes the test set size $1 < n_v < n$ and considers all $\binom{n}{n_v}$ data splittings. Since it requires computing the results from all possible data-splittings, leave-n_v-out can be computationally infeasible. Instead, one may randomly choose a certain number of data-splittings, for example, using data-splittings without replacement (Breiman et al., 1984; Burman, 1989; Zhang, 1993) and with replacement (Picard & Cook, 1984). We refer to Arlot and Celisse (2010); Ding et al. (2018b) for more CV-related references.

Like the penalty terms in AIC and BIC, the validation proportion n_v/n plays a vital role in the asymptotic property of CV methods. It is known that the leave-one-out CV is asymptotically equivalent to AIC under some regularity conditions (Stone, 1977), and it is indeed efficient in the nonparametric scenario (Shao, 1997) as AIC. Leave-n_v-out CV is shown to share a similar asymptotic property as BIC when $n_v/n \to 1$ (Shao, 1997) in terms of achieving consistency in the parametric scenario. In time series settings, a predictive validation approach, which selects a model based on the accumulated prediction loss (Rissanen, 1986a), is often used since data are not permutable. The choice of the data window to use plays a counterpart role as the data-splitting in CV, and its intimate connections with AIC and BIC have been given in, for example, Wei (1992) and Ing (2007).

In summary, in the use of CV for comparing parametric models, the data splitting ratio n_v/n drives the selection result to resemble AIC or BIC or in between. Since no single data splitting ratio generally works well, Zhan and

Yang (2022) proposed to examine a profile of the performances of the candidates with multiple data splitting ratios considered to make a better informed and adaptively well decision. It is worth mentioning that data splitting is used not only in cross-validation but also in model diagnostics (Zhang et al., 2022). But data splitting ratios for desired theoretical properties there may need to be chosen in the opposite direction as required for selection consistency in CV.

As mentioned in the Introduction, besides the advantage of simplicity and interpretability for the traditional models, even for pure prediction purposes, they may often perform better than sophisticated blackbox procedures such as the random forest when the sample size is relatively small. Interestingly, CV can be used to select among modeling procedures, particularly when choosing between traditional modeling and blackbox approaches. For example, one procedure uses BIC to choose a linear regression model, and the other is a random forest. If CV clearly prefers the first procedure, the data analyst has reasonable confidence that the selected linear model provides a sensible explanation of the regression relationship on top of its prediction accuracy advantage. In contrast, if random forest wins the CV competition, the linear models have likely missed important nonlinear effects such as interactions and higher-order terms. Finally, it is helpful to point out that in the context of comparing general modeling procedures, the data splitting ratio in the use of CV can be very different from that for comparing parametric models, as described earlier. We refer to Ding et al. (2018b); Zhang and Yang (2015) for details and more references.

6.2 | Information criteria and penalized regression

Penalized regression is a popular tool for analyzing high-dimensional data where the sample size is moderate or small compared with the number of predictors. We typically have only one full parameter vector θ for penalized regression instead of pre-specified subvectors that represent candidate models. The estimated regression coefficients are obtained by

$$\widehat{\theta} \overset{\Delta}{=} \arg\min_{\theta \in \mathbb{R}^d} \left(\sum_{i=1}^{n} s(f_\theta, z_i) + \sum_{j=1}^{d} v(|\theta_j|; \lambda, \psi) \right),$$

where $\theta = [\theta_1, \ldots, \theta_d]^{\mathrm{T}}$ is the unknown parameter vector with d possibly much larger than n, $\tau \mapsto v(\tau; \lambda, \psi)$ is a penalty function, and λ and ψ are tuning parameters. When all subsets of variables are considered as candidates, information criteria may be viewed as penalized regression methods with the ℓ_0 penalty, namely $v(\theta_j; \lambda) = \lambda \| \theta_j \|_0$ with

$$\| \theta_j \|_0 \overset{\Delta}{=} \begin{cases} 0 & \text{if } \theta_j = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{24}$$

In particular, $\lambda = 2\sigma^2$ and $\sigma^2 \log n$ correspond to AIC and BIC, respectively, when σ^2 is known.

Penalties based on ℓ_q norms with $q \geq 1$ are widely studied due to their convexity and subdifferentiability. One of the most popular methods is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) that uses $v(|\theta_j|; \lambda, \psi) = \lambda | \theta_j |$. The tuning parameter λ controls $\| \widehat{\theta} \|_1$ and determines the strength of the penalization. A problem of LASSO is that it over-penalizes parameters with large values, which leads to substantial biases in $\widehat{\theta}_m$. The method of smoothly clipped absolute deviation (SCAD) (Fan & Li, 2001) uses the penalty

$$v(\theta_j; \lambda, \psi) = \begin{cases} \lambda |\theta_j| & \text{if} |\theta_j| \leq \lambda, \\ \dfrac{2\psi\lambda|\theta_j| - \theta_j^2 - \lambda^2}{2(\psi - 1)} & \text{if} \lambda < |\theta_j| \leq \psi\lambda, \\ \dfrac{\lambda^2(\psi + 1)}{2} & \text{if} |\theta_j| > \psi\lambda. \end{cases}$$

It corrects the biases in LASSO by assigning constant penalty $\lambda^2(\psi + 1)/2$ to large parameter values. A similar method is the minimax concave penalty (MCP) (Zhang, 2010) that uses the penalty

$$v(\theta_j;\lambda,\gamma) = \begin{cases} \lambda|\theta_j| - \dfrac{\theta_j^2}{2\psi} & \text{if } |\theta_j| \le \psi\lambda, \\ \dfrac{\psi\lambda^2}{2} & \text{if } |\theta_j| > \psi\lambda. \end{cases}$$

Other penalized regression methods include elastic net (Zou & Hastie, 2005) with $q(\theta_j;\lambda_1,\lambda_2) = \lambda_1 |\theta_j| + \lambda_2\theta_j^2$ that addresses highly correlated predictors, group LASSO (Yuan & Lin, 2006) to select predictors in pre-specified subgroups, and adaptive LASSO (Zou, 2006) that adaptively assigns weights to each predictor to achieve variable selection consistency. We refer to Ding et al. (2018b) for theoretical results on the penalized regression methods for model selection.

6.3 | Minimum description length criteria

The minimum description length (MDL) is a generally applicable principle for model selection from an information-theoretical perspective (Barron, 1985; Barron et al., 1998; Hansen & Yu, 2001; Rissanen, 1978, 1982; Rissanen, 1986b). This principle has been applied in a variety of fields, including engineering. It states that the best model is the one that can describe the observed data in the most efficient way, using the minimum amount of space in coding.

To apply the MDL principle, let y denote the observed data. For simplicity, assume y is discrete (otherwise, a suitable discretization needs to be performed), and we want to represent y via binary coding. If the probability mass function of y is known to be $p(y)$, the Shannon code can describe y in a binary sequence of length $-\log_2 p(y)$ (ignoring rounding). If $p(y)$ is unknown but known to be in $\{p_\theta(y), \theta \in \mathscr{H}\}$, we may first describe which θ is to be used and then describe y with the Shannon code based on $p_\theta(y)$. When describing/encoding θ, a proper discretization of $\mathscr{H}$ is needed and one may, without any prior preference, code the discretized values with equal length. Thus, the total code length for describing y is the code length for describing θ plus $-\log_2 p_\theta(y)$. Consequently, minimizing the total code lengths is equivalent to maximizing the log-likelihood $\log p_\theta(y)$, which yields the maximum likelihood estimation (up to the discretization error).

Now instead of a single given model, we consider a set of parametric models $\{p_{\theta_m}(y), \theta_m \in \mathscr{H}_m, m \in \mathbb{M}\}$. Recall that d_m denotes the dimension of $\mathscr{H}_m$. With the added layer of modeling uncertainty, to describe the observed y, we may first describe the model index m, then describe the parameter θ_m to be used, and finally describe y with code length $-\log_2 p_{\theta_m}(y)$. It is intuitively clear that in this case, the description lengths of the parameters θ_m can be very different: a larger model (in d_m) may require longer codes to describe its parameters. Specifically, for each parameter, the discretization precision at the familiar rate $1/\sqrt{n}$ leads to the code length of order $(1/2)\log_2 n$ for describing the parameter. With d_m parameters, we have code length of order $(d_m/2)\log_2 n$. In the case where the number of models to be compared is relatively small, the code length for describing the model index m is negligible and the shortest total code length for model m is roughly

$$-\log_2 p_{\widehat{\theta}_m}(y) + \frac{d_m}{2}\log_2 n, \tag{25}$$

where $\widehat{\theta}_m$ is the maximum likelihood estimate for model m. When minimizing the above total description length over the models, we have arrived at the familiar BIC, but from an information-theoretical viewpoint in terms of coding. Thus, when MDL is applied as above, it shares the theoretical properties of BIC.

In the case of a large number of candidate models being considered, the code length to describe the model index m becomes important. For example, for all subset selection in regression with many explanatory variables, the coding of the model index should naturally favor sparse models (see, e.g., Yang and Barron et al. (1998)). Note also that the description length of the model index in the MDL approach corresponds to the model index descriptive complexity in the high-dimensional information criteria in Section 8.1.

Consistency and rates of convergence for estimators based on MDL have been well studied. For example, Barron and Cover (1991) pioneered the examination of MDL in the context of density estimation based on a countable list of candidate densities. It is shown that the MDL density estimator has a Hellinger loss upper bounded by an index of resolvability that provides the best trade-off between approximation error (in Kullback–Leibler divergence) and description complexity of the candidate density. It results in optimal or near-optimal density estimators in both parametric and nonparametric settings. Other results in various settings and applications can be found in a review article by Barron et al. (1998) and Hansen and Yu (2001).

We emphasize again that MDL is not merely a single model selection criterion. The versatile principle of MDL can be applied in different ways that lead to different theoretical properties. For example, in some cases, with proper constraints on the parameters, unlike the BIC form in (2.5), the MDL principle can actually produce a minimax optimal criterion of a form with the penalty term of the same order as that in AIC (Barron et al., 1994), which is crucial to attaining the minimax rate of convergence when the target function is in typical nonparametric function classes.

6.4 | Information criteria and other selection criteria

Information criteria are also connected to several other model selection methods. The generalized CV (GCV) (Golub et al., 1979) is a convenient approximation to the leave-one-out CV. Leeb (2008) showed that GCV could outperform AIC, AIC$_c$, and BIC in terms of efficiency when the sample size is small relative to the complexity of the data-generating model and the dimension of the best candidate model is large; The Akaike's final prediction error (FPE) (Akaike, 1970), Mallows' C_p (Mallows, 1973), and the prediction sum of squares (PSS) (Allen, 1971) are shown to be asymptotically equivalent to AIC (Nishii, 1984; Shao, 1997). For variable selection in the generalized estimating equation approach, Pan (2001) developed a modification of AIC by replacing the likelihood with the quasi-likelihood constructed under the working independence model.

7 | BRIDGING AIC AND BIC

As presented in previous sections, one group of model selection methods, including AIC, Mallows' C_p, and leave-one-out CV may be efficient in nonparametric scenarios but not consistent or efficient in the parametric scenarios, while another group that includes BIC and leave-n_v-out CV with $n_v/n \to 1$ have the reversed properties. A natural question is whether we can share the strength of these two groups of model selection methods. To illustrate this problem, we focus on AIC and BIC due to their fundamental roles in model selection. Recall that AIC is also minimax rate optimal in both parametric and nonparametric scenarios. However, for the question of whether we can share the strength between AIC and BIC in terms of consistency and minimax rate optimality in the parametric scenario, the answer is negative. Yang (2005) has proved that any consistent model selection method cannot be minimax optimal under rather general settings. Intuitively speaking, for a criterion that minimizes $\text{RSS}_m + \lambda_n d_m \sigma^2$, consistency requires $\lambda_n \to \infty$ as $n \to \infty$, which conflicts with the minimax rate optimality that requires λ_n to be upper bounded.

Another direction is attaining efficiency in both parametric and nonparametric scenarios. Compared with minimax rate optimality, efficiency is more optimistic in the sense that it only needs a “point-wise” convergence of the out-sample prediction loss to the best possible one rather than a “uniform” guarantee over different data-generating models. Fortunately, adaptively achieving efficiency in both scenarios is shown to be possible. One way is to identify the underlying scenario and decide which information criterion to use. In this direction, Liu and Yang (2011) has developed a parametricness index that goes to infinity in the parametric scenario and converges to one in probability in the nonparametric scenario. A similar notion of parametricness index was developed based on an adaptive model selection criterion Ding et al. (2018a). A second way is to perform a level-two model selection in the sense of selecting between AIC and BIC via cross-validation. For regression with orthogonal basis expansion, Zhang and Yang (2015) has shown that this approach is efficient with splitting ratio $n_t = o(n_v)$ where $n_t = n - n_v$ is the training size. Another direction of thought is to develop new information criteria that simultaneously attain the desirable properties of AIC and BIC in a way adaptive to the underlying data-generating distribution and model specification. An example of methods in this direction is the bridge criterion (Ding et al., 2018a) introduced in Section 5.

8 | HIGH-DIMENSIONAL INFORMATION CRITERIA

The traditional information criteria of AIC, BIC, and the like were derived in the framework that the number of candidate models to be compared is essentially small relative to the sample size. More technically speaking, they typically address the following example scenarios: (1) The models are nested (e.g., autoregression of various lag orders, function estimation from ordered basis functions), and the number of most relevant models is much smaller than n. In these cases, there is only a single model for each candidate model dimension. (2) The total number of predictors, d, is fixed and small compared with n, and all subset models are considered as candidates. Here, although we have multiple candidate models of the same dimension, the number of such models is still considered small.

However, the above frameworks become improper in various situations because there may be many relevant models of the same dimension. Examples include (1) The number of predictors d is close to or larger (or even much larger) than n, and the predictors are not pre-ordered by their importance. This is called a high-dimensional regression setting. (2) The number of predictors is moderate (quite a bit smaller than n), but higher-order interaction terms are considered. Even if we consider only two-way interactions, there are $\binom{d}{2}$ many interaction terms, which is fairly large when d is larger than $n^{2/3}$, for instance. It is conceivable that in some applications, one may need to explore higher-order interactions, which makes the complexity of the candidate model class even higher. In these situations, a key feature is that there are many candidate models of the same or similar dimensions and a comparison of many such models of similar dimensions may lead to the so-called selection bias (see, e.g., Yang and Barron (1998); Barron et al. (1999)), meaning that the probability of selecting the true or best model cannot be guaranteed high no matter how one designs the selection criterion.

We review the modified information criteria below to address the selection bias issue in high-dimensional linear regression. From both information-theoretic and Bayesian viewpoints, on top of the traditional information criteria, it is natural to add a penalty in the form of

$$\gamma \log \binom{d}{d_m}, \tag{26}$$

to describe the model class complexity, where d_m is the model dimension and γ is a positive constant (Barron et al., 1999; Yang & Barron, 1998). From a Bayesian point of view, this penalty means that all the subset models of the same dimension d_m are assigned with the same prior probability. From an information-theoretic perspective, $\log \binom{d}{d_m}$ represents the order of complexity associated with encoding the index of a specific model of dimension d_m.

A drawback of the above penalty is that when $d_m > d/2$, the penalty becomes decreasing as d_m increases. This is clearly undesirable from a sparse modeling perspective. Consequently, it may be better to replace the penalty with a nondecreasing upper bound

$$\gamma d_m \log \left(1 + \frac{d}{d_m}\right), \tag{27}$$

which can be found in, for example, section 2.5 and remark 4 of Wang et al. (2011).

The constant γ in (26) or (27) and in the criteria in the later two subsections needs to be chosen in application. Theoretical results on regression function estimation and variable selection consistency typically require γ to be a large enough constant. From this angle, the role of γ here is very different from the tuning parameters in LASSO, SCAD and the like, where the optimal tuning parameters depend on the sample size and are not of the constant order. Clearly, a larger choice of γ in (27) leads to a sparser selected model and the optimal choice depends on the specific situation of applications. The literature in this area seems to suggest that a (more or less) natural choice of $\gamma = 1$ or $\gamma = 1/2$ often works quite well based on numerical studies.

8.1 | Minimax rate optimality for estimation and prediction

In the high-dimensional linear regression case with d predictors and all-subset models considered, for the purpose of estimating the regression function or prediction, it is known that the selection bias (which corresponds to the "searching price" of finding out the best subset of predictors) dominates the estimation error. It results in an interesting phenomenon that the minimax rate of convergence for estimating the regression function is no longer d_{m_*}/n, but is determined by the searching price (Wang et al., 2014). For high-dimensional linear regression with soft sparsity, where the vector of coefficients from the data-generating model has its ℓ_q $(0 < q \leq 1)$ norm upper bounded, the minimax rate is achieved at an "effective model size" (Raskutti et al., 2011; Wang et al., 2014) that depends on d, n, and the sparsity level.

To achieve the minimax rate, Yang (1999) proposed the ABC criterion that selects a model $\mathscr{M}_m$ that minimizes

$$\mathrm{ABC}_m \triangleq \mathrm{RSS}_m + 2d_m\sigma^2 + \gamma\sigma^2 C_m,$$

where γ is a positive constant and C_m is interpreted as the model index descriptive complexity that satisfies $C_m > 0$ and $\sum_{m \in \mathbb{M}} e^{-C_m} \leq 1$. A particular choice is $C_m = d_m(1 + \log(d/d_m)) + 2\log(d_m + 2)$. When σ^2 is unknown, we may use its estimate in the criterion. A stochastic search (e.g., via a genetic algorithm) or a greedy procedure may be used instead of enumerating all the subset models for implementing ABC. As shown by Wang et al. (2011), the ABC criterion leads to a minimax rate optimal estimation of the regression function for both hard (meaning that the number of nonzero coefficients in the data-generating model is upper bounded) and soft (in the sense that the ℓ_q norm with $0 < q \leq 1$ of the coefficients from the data-generating model is upper bounded) sparse linear models adaptively.

8.2 | Consistency

Apart from achieving the minimax rate optimality, selection bias is also critical to deal with when model selection consistency is pursued.

8.2.1 | Extended BIC

Chen and Chen (2008) considers high-dimensional settings where the number of predictors d may increase at a polynomial rate of n. It selects a model that minimizes

$$\mathrm{EBIC}_m \triangleq n\log \mathrm{RSS}_m + d_m \log n + 2\gamma \log\binom{d}{d_m}, \quad 0 \leq \gamma \leq 1.$$

where γ is a tuning parameter. Compared with BIC, extended BIC (EBIC) has the additional penalty term $\log\binom{d}{d_m}$. From a Bayesian perspective, among models of dimension no larger than $d/2$, it assigns larger prior probabilities to smaller models compared with the uniform prior (as in BIC), which is essential to ensure variable selection consistency in high-dimensional regression under conditions on the hard sparsity, magnitudes of the true coefficients, and dependencies of the predictors.

Let $m \backslash m_*$ denote the variables in model $\mathscr{M}_m$ but not in $\mathscr{M}_{m_*}$. Chen and Chen (2008) showed that under some strong conditions, such as when $\mathscr{M}_{m_*}$ does not vary with n and at least one column of the design matrix associated with $\mathscr{M}_{m_*}$ is favorably away from the linear span corresponding to $m \backslash m_*$, EBIC selects m_* with probability going to one as n goes to infinity.

8.2.2 | BIC-p

In the previous subsection, EBIC considers a fixed $\mathscr{M}_{m_*}$. In contrast, BIC-p (Chen et al., 2022; Nan & Yang, 2014) allows the size of $\mathscr{M}_{m_*}$ to grow with n. The criterion becomes

$$\mathrm{BIC}-p_m \overset{\Delta}{=} n\log \mathrm{RSS}_m + d_m \log n + \gamma C_m,$$

where γ is a positive constant and $C_m = d_m(1+\log(d/d_m)) + 2\log(d_m+2)$. For practical implementation, choosing $\gamma = 1$ was found to work generally well (Nan & Yang, 2014).

Besides model selection, BIC-p criterion can be nautically used to construct the weights in model averaging, whose details will be elaborated in Section 9. Chen et al. (2022) showed that the BIC-p weighting is consistent in the sense that the data-generating model receives weight approaching one in probability under some conditions on the data-generating coefficients and given that the penalty constant γ is not too small. The result readily implies that BIC-p, when used for selecting a model, is consistent under the same conditions.

8.3 | Asymptotic efficiency

In the parametric scenario, model selection consistency directly implies asymptotic efficiency. Therefore with the modified EBIC or BIC-p, when the conditions for model selection consistency hold, the regression estimator is expected to achieve the smallest loss/risk among the subset models. It should be noted that the conditions are really strong and often are unlikely to be met in many applications. Consequently, such an asymptotic efficiency result may not be applicable in practice.

To the best of our knowledge, there is no general result on high-dimensional asymptotic efficiency in the nonparametric scenario. Although no formal negative results have been given in the literature (to our understanding), it seems intuitively convincing that the asymptotic efficiency may be out of reach unless unusually strong assumptions are made on the data-generating model, or some specific settings are under consideration. A particular setting is selecting the number of iterations in a step-wise variable selection algorithm for high-dimensional time series (Ing, 2020), where the author developed a high-dimensional AIC and showed its efficiency in the sense that the selected model performs closely to the performance under the best possible number of iterations.

9 | MODEL SELECTION UNCERTAINTY

Regardless of which model selection criterion is chosen for the data, a final model is selected. An important question is: How reliable is the selected model? Is it genuinely the best or data-generating model? Or is it among relatively few top-performing models? Or is it just one of many models that provide some prediction power? If it is in the first case, the selected model can be safely used for interpretation, inference, and prediction. In the second case, the selected model is close to the data-generating model, and we may not be sure whether a few variables are in or not in that model. In the third case, the selected model may serve the prediction goal quite well, but using it for inference may not be wise since it can lead to unreliable conclusions. With the above, assessing model selection uncertainty is crucial in model selection. Information criteria and related criteria can be used for model selection diagnostics. The key idea is to incorporate model weights based on model averaging in the model assessment process instead of relying on the selected model alone.

9.1 | Selection instability

We may apply instability measures to assess the stability of model selection methods, for example, information criteria, under slight perturbations of the dataset. For example, the sequential stability (Chen et al., 2007) randomly removes a small portion of the data and applies the model selection method to the remaining data. One can compare the selection results by calculating the symmetric difference between the sets of selected predictors. Bootstrap re-sampling can also assess the instability by regenerating the dataset (Breiman, 1996; Buckland et al., 1997; Diaconis & Efron, 1983). For instance, to assess the variable selection stability in linear regression by parametric bootstrap, we first obtain the fitted values from the selected model and then add random noise to obtain the bootstrapped datasets. A summary statistic can be obtained from the symmetric differences between the set of selected predictors in each bootstrapped dataset and the one from the original dataset. The method of perturbation instability in variable selection (Nan & Yang, 2014) is similar to the above methods, except that it generates new data by perturbing the response of the original dataset.

9.2 | Model averaging for reducing instability

Instead of selecting a single model, we may consider the combined model

$$\widehat{f}_w(x) = \sum_{m \in \mathbb{M}} w_m \cdot f_{\widehat{\theta}_m}(m),$$

where

$$w_m \triangleq \frac{\exp(-\mathcal{I}_m/2)}{\sum_{m' \in \mathbb{M}} \exp(-\mathcal{I}_{m'}/2)}. \tag{28}$$

Here, the index set $\mathbb{M}$ is assumed to be finite, and $\mathcal{I}_m$ is an information criterion value (e.g., AIC_m or BIC_m). For high-dimensional regression where the number of candidate models rapidly grows with n, we may use $\mathcal{I}_m = \text{BIC} - p_m$ (Chen et al., 2022). For prediction, when the candidate models are hard to distinguish, model averaging may outperform model selection due to its increased stability (Yang, 2003). In addition to improving model selection in estimation or prediction accuracy, we can use the model averaging weights to perform model selection diagnosis, as seen in the next subsections.

9.3 | Variable selection deviation

An important but ignored aspect of the model selection instability measurements is the quality of the selected variables compared with the data-generating model. In the worst case, for example, a model selection method may constantly select one predictor with high stability, which is irrelevant to the data-generating model. The variable selection deviation (VSD) is proposed to address this issue. With model averaging methods like adaptive regression by mixing (ARM) (Yang, 2001) and Bayesian model averaging (with weights of the form in Formula (28)), VSD identifies the number of predictors in the selected model that are not in $\mathcal{M}_{m_*}$ (defined for the parametric setting) and those in the reversed case.

9.4 | *F*- and *G*-measures

When variable selection consistency is of primary interest, it is crucial to quantify the quality of a selection result. While one may assess false positives and false negatives separately, as in VSD, it is helpful to have a measure that combines the under- and over-selection aspects. *F*- and *G*-measures are popular options in this regard.

Let $\mathcal{M}_{m_*}$ and $\mathcal{M}_{\widehat{m}}$ denote the sets of the variables in the data-generating and selected models, respectively. Then, the precision (or positive predictive value) is defined as the fraction of selected variables in $\mathcal{M}_{\widehat{m}}$ that are true variables, and recall (also known as sensitivity) is defined as the fraction of the true variables that are selected. Then, the *F*-measure is the harmonic mean of precision and recall and the *G*-measure is the geometric mean of precision and recall. Combining precision and recall into one measure may describe the overall accuracy of a given variable selection method. The *F*- and *G*-measures are between 0 and 1, and a higher value generally indicates a better selection performance. If the *F*- and *G*-measures are very low, say 0.1, the selection result is far from the truth and may not be used to guide which variables should be included in the data-generating model.

In practical data applications, *F*- and *G*-measures cannot be computed since the data-generating model is unknown. However, we can use information criteria or other methods to calculate the weights of the candidate models and use them to approximate *F*- and *G*-measures. The idea is to pretend each candidate model to be $\mathcal{M}_{m_*}$, calculate the *F*- and *G*-measures from the selected model and then average those measures according to a model averaging weighting. Yu et al. (2022) showed that this simple method leads to a consistent estimator of the actual *F*- and *G*-measures uniformly over all $m \in \mathbb{M}$ as long as the model average weighting is weakly consistent, in the sense that the weights of significantly misspecified models diminish as $n \to \infty$.

With the estimated *F*- and *G*-measures, one may have a good sense of the reliability of the selected model. For instance, if the estimated *F* value of the selection result is 0.9, there is little concern in treating it as a reliable description of the data-generating model. Suppose the estimated F value is 0.1. Even if there is good predictive power, the variables chosen are unlikely to be the most important ones that a scientist hopes to verify in future studies.

9.5 | Variable importance

Variable importance can be used for a prescreening procedure in high-dimensional regression to reduce costs in data analysis and improve stability. Also, it can be applied to improve the scientific understanding of the predictors. One method to evaluate variable importance, which is related to information criteria, is sparsity-oriented importance learning (SOIL) (Ye et al., 2018). For a predictor indexed by j and candidate models indexed by $\mathbb{M}$, SOIL calculates $S_j \triangleq \sum_{m \in \mathbb{M}} w_m \cdot \mathbb{I}\{\text{predictor}\, j\, \text{is in} \mathscr{M}_m\}$, j where the weight w_m can be obtained from Formula (28). The SOIL importance is consistent, meaning that it converges to one in probability for a variable in the data-generating model and to zero otherwise. While nonparametric variable importance measures such as those based on random forests (Gregorutti et al., 2017) are more widely applicable, when parametric modeling is proper, SOIL may produce more reliable and informative variable importance values (Ye et al., 2018).

10 | MISLEADING FOLKLORES

Although the existing asymptotic analysis of information criteria has provided many insights into their fundamental properties, misconceptions exist about their applications.

10.1 | Folklore one

> *AIC should be preferred for prediction and BIC for explanation.*

The derivations of AIC and BIC at the beginning of Sections 3 and 4 are from different angles: AIC is derived to predict future data accurately, and BIC is derived to find the actual data-generating model. This fact is often emphasized in the literature, for example, Shmueli (2010). These distinct angles indeed provide a nice insight into how different model selection objectives could motivate specific penalties of model complexity. However, the folklore that AIC should be preferred for prediction and BIC for explanation is an oversimplified view since it overlooks the difference between the parametric and nonparametric scenarios. For AIC, we mentioned in Section 3 that it could fail to lead to the best prediction result even when $n \to \infty$. For BIC, the consistency of selecting the data-generating model may not be well-defined in the nonparametric scenario. More closely related discussions are given in the following subsection.

10.2 | Folklore two

> *AIC should be preferred since the nonparametric scenario is more realistic in practice* (Aho et al., 2014).

More specifically, the viewpoint that the data-generating model has a complicated unknown structure is prevalent in, for example, biological studies (Anderson & Burnham, 2002; Burnham et al., 2011; Johnson & Omland, 2004). The above statement overlooks that the performance of information criteria in practice may also be affected by the sample size. For instance, consider linear regression in the nonparametric scenario where only a few coefficients are significant at the given sample size. BIC, which selects a standing-out model, may be preferred in this setting. In contrast, in the parametric scenario where the coefficients are all small with different magnitudes and the sample size is insufficient to estimate them accurately, AIC performs better in selecting among competing candidate models with similar performances. The above two scenarios are referred to as "practically parametric" and "practically nonparametric" (Ding et al., 2018b; Liu & Yang, 2011). Zhang and Yang (2015) gave an example where the relative performance in terms of out-sample prediction loss between AIC and BIC switches as the sample size increase for a fixed data-generating model.

In the (practically) parametric case, which is applicable in many applications, BIC is asymptotically efficient and may be much better in prediction than AIC. Even for explanation purposes, AIC may be preferred sometimes. For example, suppose our goal is to identify potentially important genes for treating disease based on a pilot study with a limited sample size. In this case, the use of BIC may be too conservative in terms of missing important genes that may be revealed in a follow-up study with a much larger sample size. In contrast, although AIC may choose "noise variables," it is much less likely to miss "true variables" with relatively small magnitudes in relation to the pilot sample size.

In practice, in line with the above discussion, it is promising to apply methods in Section 7, which can be adaptive to "practically parametric" and "practically nonparametric" scenarios. For interpretation, BIC is preferred for its stability when there is evidence for a "practically parametric" scenario. For example, in the study to identify genes with important roles in a certain disease, we hope the variable identification result from the selected models can be generalized to future observations, and BIC that penalizes over-selection is required. In contrast, in exploratory studies, AIC may be preferred since it is less conservative to find possibly important variables where over-selection is accepted.

10.3 | Folklore three

> *The ℓ_0 penalty is not as good as a smoothed penalty such as LASSO, SCAD, and MCP penalty because it is discontinuous.*

Recall that information criteria can be viewed as a penalized regression with the ℓ_0 penalty (as defined in (24)). Whether a penalty is "good" depends on the evaluation goal. In terms of stability, the claim seems plausible since the objective function with the ℓ_0 penalty is indeed discontinuous, and the discontinuity may lead to undesirable instability of the selection result. But this is not true. First, even for a fixed tuning parameter, there is no known relationship between the continuity/discontinuity of the penalty function and the behavior of the selection result. Second, with the tuning parameter selected based on data, there is even less reason to believe the other penalties lead to better performance than the ℓ_0 penalty.

For adaptive LASSO, SCAD, and MCP, which are known to produce oracle estimators under different conditions, Leeb and Pötscher (2008) showed that the estimation and prediction risk has a supremum that diverges to infinity, implying that the selection results may be unstable under a tiny change in data-generating parameters. On the contrary, AIC and high-dimensional modifications are known to be minimax rate optimal. In fact, to our knowledge, the ℓ_0 penalty approach yields minimax rate optimality under the least stringent conditions than other penalties. While there is much work to be done on the advantages and disadvantages of the different penalties, there seems to be the increasingly popular view that penalties of LASSO, SCAD, MCP, and so on are appropriate relaxations of the ℓ_0 penalty for computational considerations.

11 | CONCLUDING REMARKS

In summary, in model selection, the model dimension d_m needs to be appropriately chosen to strike the best bias-variance tradeoff for optimal prediction. Identifying the most appropriate model for inference is essential as well. Information criteria address these tasks by combining the in-sample loss with a penalty term. Their performance depends on the goal of model selection and the relationship between the data-generating process and postulated models, categorized as parametric and nonparametric scenarios. AIC represents a group of information criteria that are efficient in the nonparametric scenario in the sense that the prediction performance of the selected model is asymptotically close to the best among the candidate models. However, they may fail to be consistent in choosing the most parsimonious well-specified model if it exists. It is because their penalties are too small to distinguish between two well-specified models and thus will lead to an over-selection of d_m. In contrast, BIC represents another group of information criteria more suitable for a parametric scenario. But they may penalize too much to enjoy the efficiency in the nonparametric scenario.

Finally, based on the previous results and our experience, we make the following suggestions for the practice of model selection.

- If the goal of model selection is accurate prediction, we suggest either (1) choosing between AIC-type and BIC-type methods based on parametricness index or cross-validation, or (2) using adaptive information criteria such as BC to combine the strength of AIC and BIC.

- If the main goal is variable selection for interpretation, namely one hopes to reproduce the selected variables in a follow-up study with a similar sample size, it is better to use BIC-type methods to avoid including variables unlikely to be shown to be significant.
- If protection of the worst-case prediction accuracy is essential for an application, AIC is preferred due to its minimax rate optimality. BIC can be arbitrarily worse than AIC in risk ratio but not the other way around. In addition, for exploratory studies to find possibly relevant variables, even though AIC-type methods may over-select, they are safe in terms of not missing important variables that one may verify in follow-up studies with large sample sizes.
- When the number of predictors, d, is not small compared with the sample size n and all subsets of the d variables are considered, it is better to use a high-dimensional AIC or BIC to address the potential severe selection bias.
- Perform model selection diagnosis. Instability, VSD, F- and G-measures, variable importance, and so on may be investigated to quantify the reliability of model selection results. If diagnostic results are concerning, we may not trust the selection result and keep exploring other methods.
- When model selection instability is high, for prediction purposes, one may consider model averaging.

AUTHOR CONTRIBUTIONS

Jiawei Zhang: Conceptualization (equal); writing – original draft (equal); writing – review and editing (equal). **Yuhong Yang:** Conceptualization (equal); writing – original draft (equal); writing – review and editing (equal). **Jie Ding:** Conceptualization (equal); writing – original draft (equal); writing – review and editing (equal).

ACKNOWLEDGMENTS

We sincerely thank the Editors-in-Chief, James Gentle and David Scott, for handling the manuscript and two anonymous reviewers for their constructive comments that have improved the article.

FUNDING INFORMATION

J. Zhang was supported in part by the National Science Foundation under grant number ECCS-2038603. J. Ding was supported in part by the National Science Foundation under grant number DMS-2134148.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Jie Ding https://orcid.org/0000-0002-3584-6140

REFERENCES

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, *22*(1), 203–217.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen (Ed.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, *13*(3), 469–475.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, *16*(1), 125–127.

Anderson, D., & Burnham, K. (2002). *Model selection and multimodel inference, a practical information-theoretic approach* (2nd ed.). Springer.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Arlot, S., & Massart, P. (2009). Data-driven calibration of penalties for leastsquares regression. *Journal of Machine Learning Research*, *10*(2), 245–279.

Barron, A. R. (1985). Logistically smooth density estimation (PhD thesis). Department of Electrical Engineering, Stanford, CA.

Barron, A. R., Birgé, L., & Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, *113*(3), 301–413.

Barron, A. R., & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, *37*(4), 1034–1054.

Barron, A. R., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*(6), 2743–2760.

Barron, A. R., Yang, Y., & Yu, B. (1994). Asymptotically optimal function estimation by minimum complexity criteria. In Proceedings of 1994 IEEE international symposium on information theory, p. 38.

Bartol, K., Bojanić, D., Petković, T., Peharec, S., & Pribanić, T. (2022). Linear regression vs. deep learning: A simple yet effective baseline for human body measurement. *Sensors*, *22*(5), 1885.

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory* (Vol. 405). John Wiley & Sons.

Birgé, L., & Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, *138*(1), 33–73.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*(6), 2350–2383.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees. Wadsworth statistics/probability series*. Wadsworth Advanced Books and Software.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*(2), 603–618.

Burman, P. (1989). A comparative study of ordinary cross-validation, ν-fold cross-validation and the repeated learning-testing methods. *Biometrika*, *76*(3), 503–514.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Chen, L., Giannakouros, P., & Yang, Y. (2007). Model combining in factorial data analysis. *Journal of Statistical Planning and Inference*, *137*(9), 2920–2934.

Chen, Z., Zhang, J., Xu, W., & Yang, Y. (2022). Consistency of BIC model averaging. *Statistica Sinica*, *32*, 635–640.

Claeskens, G., & Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, *98*(464), 900–916.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, *248*(5), 116–131.

Ding, J., Diao, E., Zhou, J., & Tarokh, V. (2021). On statistical efficiency in learning. *IEEE Transactions on Information Theory*, *67*(4), 2488–2506.

Ding, J., Tarokh, V., & Yang, Y. (2018a). Bridging AIC and BIC: A new criterion for autoregression. *IEEE Transactions on Information Theory*, *64*(6), 4024–4043.

Ding, J., Tarokh, V., & Yang, Y. (2018b). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, *35*(6), 16–34.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, *97*(457), 77–87.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, *22*(4), 1947–1975.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328.

Ghosh, J., & Ramamoorthi, R. (2003). *Bayesian nonparametrics*. Springer.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B*, *41*(2), 190–195.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, *96*(454), 746–774.

Hong, X., Mitchell, R. J., Chen, S., Harris, C. J., Li, K., & Irwin, G. W. (2008). Model selection approaches for non-linear system identification: A review. *International Journal of Systems Science*, *39*(10), 925–946.

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

Ing, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *The Annals of Statistics*, *35*(3), 1238–1277.

Ing, C.-K. (2020). Model selection for high-dimensional linear regression with dependent observations. *The Annals of Statistics*, *48*(4), 1959–1980.

Ing, C.-K., Sin, C.-Y., & Yu, S.-H. (2012). Model selection for integrated autoregressive processes of infinite order. *Journal of Multivariate Analysis*, *106*, 57–71.

Ing, C.-K., & Wei, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *The Annals of Statistics*, *33*(5), 2423–2474.

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology & Evolution*, *19*(2), 101–108.

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*(3), 661–690.

Leeb, H., & Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, *142*(1), 201–211.

Liu, W., & Yang, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *The Annals of Statistics*, *39*(4), 2074–2102.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*(4), 661–675.

Nan, Y., & Yang, Y. (2014). Variable selection diagnostics measures for highdimensional regression. *Journal of Computational and Graphical Statistics*, *23*(3), 636–656.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, *12*(2), 758–765.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*(1), 120–125.

Pesaran, M. H. (1974). On the general problem of model selection. *The Review of Economic Studies*, *41*(2), 153–171.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Transactions on Information Theory*, *57*(10), 6976–6994.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Rissanen, J. (1982). Estimation of structure by minimum description length. *Circuits, Systems and Signal Processing*, *1*(3), 395–406.

Rissanen, J. (1986a). A predictive least-squares principle. *IMA Journal of Mathematical Control and Information*, *3*(2-3), 211–222.

Rissanen, J. (1986b). Stochastic complexity and modeling. *The Annals of Statistics*, *14*, 1080–1100.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, *7*, 221–242.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, *8*, 147–164.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, *68*(1), 45–54.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B*, *64*(4), 583–639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society. Series B*, *36*(2), 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by crossvalidation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B*, *39*(1), 44–47.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, *153*, 12–18.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, *58*(1), 267–288.

Tukey, J. W., et al. (1961). Curves as parameters, and touch estimation. In Proceedings of the 4th Berkeley symposium on mathematical statistics and probability, vol. 1, pp. 681–694.

Walsh, L. (2007). A short review of model selection techniques for radiation epidemiology. *Radiation and Environmental Biophysics*, *46*(3), 205–213.

Wang, Z., Paterlini, S., Gao, F., & Yang, Y. (2011). Adaptive minimax estimation over sparse l_q-hulls. *arXiv Preprint*. https://doi.org/10.48550/arXiv.1108.1961

Wang, Z., Paterlini, S., Gao, F., & Yang, Y. (2014). Adaptive minimax regression estimation over sparse l_q-hulls. *Journal of Machine Learning Research*, *15*(50), 1675–1711.

Wei, C.-Z. (1992). On predictive least squares principles. *The Annals of Statistics*, *20*(1), 1–42.

Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, *9*, 475–499.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, *96*(454), 574–588.

Yang, Y. (2003). Regression with multiple candidate models: Selecting or mixing? *Statistica Sinica*, *13*(3), 783–809.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, *92*(4), 937–950.

Yang, Y. (2007). Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory*, *23*(1), 1–36.

Yang, Y., & Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, *44*(1), 95–116.

Yang, Y., & Barron, A. R. (1999). Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, *27*, 1564–1599.

Ye, C., Yang, Y., & Yang, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association*, *113*(524), 1797–1812.

Yu, Y., Yang, Y., & Yang, Y. (2022). Performance assessment of high-dimensional variable identification. *Statistica Sinica*, *32*, 695–718.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, *68*(1), 49–67.

Zhan, Z., & Yang, Y. (2022). Profile electoral college cross-validation. *Information Sciences*, *586*, 24–40.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zhang, J., Ding, J., & Yang, Y. (2022). Is a classification procedure good enough?—A goodness-of-fit assessment tool for classification learning. *Journal of the American Statistical Association*, 1–11. https://www.tandfonline.com/doi/abs/10.1080/01621459.2021.1979010

Zhang, J., Ding, J., & Yang, Y. (2023). Targeted cross-validation. *Bernoulli*, *29*(1), 377–402.

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, *21*(1), 299–313.

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, *67*(2), 301–320.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*(1), 41–61.

How to cite this article: Zhang, J., Yang, Y., & Ding, J. (2023). Information criteria for model selection. *WIREs Computational Statistics*, e1607. https://doi.org/10.1002/wics.1607