A Parametricness Index and Consistency with Complexity Penalty for Model Selection

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Wei Liu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Yuhong Yang, Adviser

September 2010

Acknowledgements

I would like to thank my adviser Yuhong Yang for his guidance, support, and encouragement throughout my studies at the University of Minnesota. He is the professor with whom I have taken the most courses at the University of Minnesota. His expertise and deep understanding in various areas have always been inspiring to me both in classes and in my research. I would also like to thank Dennis Cook, Hui Zou, and Wei Pan for serving in my committee and for their valuable suggestions and comments regarding my dissertation research.

Many thanks are due to the faculty, staff, and fellow students in the School of Statistics during my studies here. Your support, help, and friendship have made the past five years in Minnesota a pleasant time for me. In particular, I would like to thank Charles Geyer who is always available for questions and discussions.

I owe my utmost gratitude to my parents who have given me their deepest love throughout my life. I can never overstate what their love means to me. My gratitude also goes to my brother Jun Liu who has always been supportive. My thesis is dedicated to them.

Abstract

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

Traditionally, the consistency property of BIC type of criteria for model selection is derived with a fixed number of predictors. A natural question is: does the consistency property still hold in high dimensional setting? The answer is in the positive direction [18, 69]; however, there are serious limitations of the assumptions in [18, 69]. Specifically, in [18], the size of the true model is assumed to be bounded, which may exclude many applications. In [69], the conditions 2 assumes that the smallest eigenvalue of the covariance matrix of all the predictors is always positive, which could be a little unrealistic due to the correlation among all the predictors, especially when the number of predictors is large. And the condition 4 in [69] assumes that the smallest

coefficient in the true model is higher than a certain order, which is reasonable, but the order could be improved. We provide sufficient conditions on consistency for BIC and similar types of criteria in high dimensional settings and show that these conditions are also necessary in a sense by giving counterexamples. We demonstrate that the results in [18, 69] are special cases of ours. Moreover, our results eliminate the the restriction in [18] on the size of the true model and relax the assumptions in [69] on the true model. We also generalize the concept of consistency and provide similar results to this new concept. A statistical risk bound for the model selected by the BIC type of criterion is also derived.

Contents

Acknowledgements **i**

Abstract **ii**

List of Tables **vii**

List of Figures **viii**

1 Introduction **1**

1.1 A review of model selection methods and the conflict between AIC and BIC 1

1.2 Model selection with high dimensional data 2

1.3 Issues of model selection in this thesis 4

1.4 Organization of the thesis 5

2 Parametric or Nonparametric? A Parametricness Index for Model Selection **6**

2.1 Introduction 6

2.1.1 Model selection criteria and their possibly conflicting properties 8

2.1.2 Model selection: A gap between theory and practice 10

2.2 Setup of the regression problem 14

2.3 Main Theorems 16

2.3.1 Parametric Scenarios 17
2.3.2 Nonparametric Scenarios 19
2.3.3 PI separates parametric and nonparametric scenarios 22
2.3.4 Examples 22
2.3.5 On the choice of λ_n and d 25
2.3.6 Combining strengths of AIC and BIC 25
2.4 PI as a model selection diagnostic measure, i.e., *P*ractical *I*dentifiability of the best model 26
2.5 Simulation Results 31
2.5.1 Single predictor 31
2.5.2 Factors that influence PI 32
2.5.3 Multiple predictors 36
2.5.4 A summary 43
2.6 Real Data Examples 44
2.7 Conclusions 47
2.8 Proofs 50

3 On consistency of model selection with model complexity penalty 60
3.1 Introduction 60
3.2 Problem Setup 63
3.2.1 Model selection criteria 64
3.2.2 Notation 65
3.3 Consistency of GICC when σ^2 is known 66
3.3.1 General conditions on consistency 66
3.3.2 On Constraint ((3.1)) and Conditions $(G1) - (G3)$ 69
3.3.3 Consistency of GICC for subset selection 75
3.3.4 Comparison to existing results 76

3.3.5 Consistency of GICC for selection among a sequence of models 79
3.3.6 Counterexamples 82
3.4 Consistency of GICC when σ^2 is unknown 86
3.4.1 Results for the case with $p_n = o(n)$ 87
3.4.2 Results for arbitrary p_n with additional information on σ^2 . . 89
3.5 Consistency of GICC without parametric assumptions 91
3.5.1 σ^2 known 92
3.5.2 σ^2 unknown 96
3.6 A risk bound for GICC on regression estimation 98
3.7 Conclusions 100
3.8 Proofs 101

4 Summary and future research 120

References 122

List of Tables

2.1 Percentiles of PI for Example 1 32
2.2 Percentiles of PI for Example 2 33
2.3 Proportion of selecting true model 38
2.4 Quartiles of PIs 38
2.5 Reliability of inference for Example 3 40
2.6 Reliability of inference for Example 4 41
2.7 Statistical risks of AIC, BIC, and the Combined procedure 42
2.8 Quartiles of PIs from subsamples of size 400 45
2.9 Quartiles of PIs from subsamples of size 200 45
2.10 The 6 most frequently selected models and their frequencies with a sample size of 400 45
2.11 The 6 most frequently selected models and their frequencies with a sample size of 200 46
2.12 Combining AIC and BIC based on PI with full sample size 46

List of Figures

2.1 Scatterplots For Example 1 32
2.2 Scatterplots For Example 2 33
2.3 Sample size effect for Example 1 34
2.4 Sample size effect for Example 2 34
2.5 Effect of coefficient 35
2.6 Behavior of PI for the special example 39

Chapter 1

Introduction

1.1 A review of model selection methods and the conflict between AIC and BIC

Model/variable selection is one of the most important problems in statistics. A large number of model selection criteria has been proposed in the statistics literature, such as AIC, C_p , BIC, GIC, RIC, FPE, MDL, CV, GCV, LASSO, SCAD, adaptive LASSO, and so on. A tremendous amount of exciting research has been done on theory, computation, and application of various model selection methods. In one direction, pointwise asymptotic results (e.g., [19, 27, 45, 53, 54, 55, 56, 59, 62, 66, 68, 74, 79, 83, 84]) have been established in terms of either selection consistency or an asymptotic optimality. In another direction, results on minimax risk bounds of the model selected by a model selection method with a function class have also been derived (see [3, 4, 5, 6, 11, 15, 16, 23, 24, 47, 76].) In particular, it has been shown that AIC, C_p, FPE, and CV have an asymptotic optimality property which says the accuracy of the estimator based on the selected model is asymptotically the same as the best candidate model when the true model is infinite dimensional. In contrast, BIC and the like are consistent when the true model is finite-dimensional and is among the candidate models. It also has been demonstrated that no model selection method can simul-

taneously enjoy the consistency property of the BIC-type method and the minimax rate optimal property of the AIC-type method [73]. The conflict between AIC and BIC has generated a big debate in the literature regarding model selection methods that is not only statistical but also philosophical, especially about the existence of a true model and the ultimate goal of statistical modeling (see, e.g., [8, 14, 17, 33, 81], and references therein). There also have been efforts on using data-dependent and adaptive penalties to bridge the gap between the AIC-type method and the BIC-type method (see, e.g., [7, 34, 37, 44, 60, 61, 75]). Despite the exciting pointwise asymptotic results on different model selection methods, the finite sample behaviors of model selection methods are quite different. For finite-sample results and effects of model selection on post-model-selection estimators, see e.g., [22, 49]. Particularly, it is argued there that the use of a consistent model selection procedure does not necessarily allows one to act as if the true model were known in advance.

1.2 Model selection with high dimensional data

Traditionally, model selection methods focused on the setting of a finite number of predictors. With the innovations of modern technology, scientists and researchers have been able to collect enormous amounts of data at low costs. This kind of high dimensional data problem has become increasingly frequent and important in many fields, including biology, health sciences, economics, finance, and engineering. It challenges traditional statistical analysis in both theory and computation. Still model/variable selection plays a critical role in knowledge discovery with such massive data. For example, in gene expression studies, microarrays contain information on thousands of genes which may be linked to certain diseases and researchers often need to identify the important ones and construct biochemical models.

Recently, different shrinkage methods have been developed, such as the LASSO

[67] and the SCAD [27], and have been widely studied in high dimensional model selection (see [16, 28, 29, 30, 41, 42, 48, 54, 69, 70, 79, 78, 83, 84, 85].) Fan and Peng [30] established an oracle property and the asymptotic normality of the SCAD estimators under some regularity conditions in situations with a diverging number predictors. Huang et al [42] showed that the adaptive LASSO estimators in sparse, high-dimensional, linear regression models enjoy the same oracle property under appropriate conditions, if a reasonable initial estimator is available. Candes and Tao [16] introduced a new l_1 estimator, the Dantzig selector, and showed the estimator achieves a loss within a logarithmic factor of the ideal mean squared error, even though the sample size may be much smaller than the number of predictors. Meinshausen and Bühlmann [54] proposed a neighborhood selection scheme with LASSO and showed that it is consistent for sparse high-dimensional graphs. In high dimensional linear regressions, Zhang and Huang [79] proved that LASSO is rate consistent under a sparse Riesz condition on the correlation of design variables. Zhao and Yu [83] proved that the " Irrepresentable Condition " is almost necessary and sufficient for LASSO to select the true model both in the classical fixed p setting and in the large p setting as the sample size n gets large. Zhang [78] proposed MC+, a fast, continuous, nearly unbiased and accurate method of penalized variable selection in high dimensional linear regression and proved that the MC+ achieves correct selection, without assuming the strong irrepresentable condition required by the LASSO. Zou and Zhang [85] proposed the adaptive elastic net method and showed that it has the oracle property and handles the collinearity problem better than other oracle-like methods.

In a different direction, Chen and Chen [18] proposed extended BIC for model selection with large model spaces and obtained consistency. Wang et al [69] modified traditional BIC criterion in the case of a diverging number predictors and showed that it is consistent for both unpenalized and penalized estimators.

1.3 Issues of model selection in this thesis

In this thesis, we try to answer two important questions in model selection.

The first is related to the conflict between AIC and BIC. It is well-known that for a typical regression problem with a number of predictors, AIC and BIC tend to choose models of significantly different sizes, which may have serious practical consequences. Therefore, it is important to decide which criterion to apply for a data set at hand. Unfortunately, the current theories on model selection have little to offer to address this issue. Consequently, it is rather common that statisticians/statistical users resort to the "faith" that the true model certainly cannot be finite-dimensional for the choice of AIC, or to the strong preference of parsimony or goal of model identification to defend his/her use of BIC. To us, the question whether or not AIC is more appropriate than BIC for the data at hand should and can be addressed statistically rather than based on one's preferred assumption. Thus the first question we try to answer is that: can we statistically distinguish between parametric and nonparametric scenarios? Chapter 2 deals with this question.

The second question is related to selection consistency of the BIC type criterion with a diverging number of predictors. Although consistency has been obtained [18, 69] for BIC-type of criterion in situations where the number of predictors tends to infinity, general results on similar types of model selection criteria with complexity penalty terms are still missing, despite the fact that BIC and similar criteria have been widely used in applications even in high dimensional situations. Furthermore, there are serious limitations in the technical assumptions in [18, 69], which may exclude many applications. The second question we try to answer is that: what are the sufficient and necessary conditions for consistency to hold for the BIC type or more generalized method? In Chapter 3, we provide detailed discussion on such conditions and compare them to existing results. We also derive a statistical risk bound for the

regression estimator of the selected model based on BIC and similar criteria.

1.4 Organization of the thesis

This thesis is organized as follows. In Chapter 2, we develop a simple parametricness index (PI) for regression based on finite-dimensional models and show how it can be used in application to bridge between AIC and BIC. In Chapter 3, we provide detailed discussion on sufficient and necessary conditions on consistency of the BIC type model selection method and compare them to existing results. We summarize our findings and point out future research directions in Chapter 4.

Chapter 2

Parametric or Nonparametric? A Parametricness Index for Model Selection

2.1 Introduction

When considering parametric models for data analysis, model selection methods have been commonly used for various purposes. If one candidate model describes the data really well (e.g., a physical law), it is obviously desirable to identify it. Consistent model selection rules such as BIC [58] are proposed for this purpose. In contrast, when the candidate models are constructed to progressively approximate an infinite-dimensional truth with a decreasing approximation error, the main interest is usually on estimation and one hopes that the selected model performs optimally in terms of a risk of estimating a target function (e.g., the regression function). AIC [2] has been shown to be the right criterion from an asymptotic efficiency and also a minimax-rate optimality views.

The question if we can statistically distinguish between parametric and nonparametric scenarios motivated our research. In this paper, for regression based on finite-dimensional models, we develop a simple parametricness index (PI) that has the

following properties.

1. With probability going to 1, PI separates typical parametric and nonparametric scenarios.

2. It advises on whether identifying the true or best candidate model is feasible at the given sample size or not by assessing if one of the models stands out as a stable parametric description of the data.

3. It informs us whether interpretation and statistical inference based on the selected model are reasonably reliable or not due to model selection uncertainty.

4. It tells us whether AIC is likely better than BIC or not for the data at hand.

5. It can be used to approximately achieve the better estimation performance of AIC and BIC for both parametric and nonparametric scenarios.

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

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

2.1.1 Model selection criteria and their possibly conflicting properties

To assess performance of model selection criteria, pointwise asymptotic results (e.g., [19, 27, 45, 53, 54, 55, 56, 59, 62, 66, 68, 74, 79, 83, 84]) have been established mostly in terms of either selection consistency or an asymptotic optimality. It is well-known that AIC [2], C_p [52], and FPE [1, 63] have an asymptotic optimality property which says the accuracy of the estimator based on the selected model is asymptotically the same as the best candidate model when the true model is infinite dimensional. In contrast, BIC and the like are consistent when the true model is finite-dimensional and is among the candidate models.

Another direction of model selection theory focuses on oracle risk bounds (also called index of resolvability bounds) that immediately lead to minimax type of results. That is, a risk bound with minimal assumptions is derived for a model selection method, which link the risk of the selected model to the smallest risk among all the candidate models. Given a class of functions to which the target function belong, by maximizing the risk bound over the function class, a bound on the worst-case risk of the model-selection-based estimator is readily obtained. When the candidate models are constructed to work well for target function classes, this yields minimax-rate or near minimax-rate optimality results. Publications of work in this direction include [3, 4, 5, 6, 11, 15, 16, 23, 24, 47, 76], to name a few. In particular, AIC type of model selection methods are minimax-rate optimal for both parametric and nonparametric scenarios (see [5, 73]). A remarkable feature of the works inspired by [6] is that with a complexity penalty (other than one in terms of model dimension) added to deal with a large number of (e.g., exponentially many) models, the resulting risk or loss of the selected model automatically achieves the best trade-off between approximation error, estimation error and the model complexity, which provides tremendous theoretical

flexibility to deal with a fixed countable list of models (e.g., for series expansion based modeling) or a list of models chosen to depend on the sample size (see, e.g., [5, 76, 71]).

While pointwise asymptotic results are certainly interesting, since they describe limiting behaviors when the sample size approaches infinity, similarly to the issue of asymptotic normality of the sample mean from a possibly highly non-Gaussian density, it is not surprising that the limiting behaviors can be very different from the finite-sample reality, especially when model selection is involved. For finite-sample results and effects of model selection on post-model-selection estimators, see e.g., [22, 49]. In particular, it is argued there that the use of a consistent model selection procedure does not necessarily allows one to act as if the true model were known in advance.

The general forms of AIC and BIC make it very clear that they and similar criteria (such as GIC in [57]) cannot simultaneously enjoy the properties of consistency in a parametric scenario and asymptotic optimality in a nonparametric scenario. Efforts have been put on using penalties that are data-dependent and adaptive (see, e.g., [7, 34, 37, 44, 60, 61, 75]). Shen and Ye [61] proposed an adaptive model selection procedure to approximate the best performance of a class of procedures across a variety of situations. Yang [75] showed that the asymptotic optimality of BIC for a parametric scenario (which follows directly from consistency of BIC) and asymptotic optimality of AIC for a nonparametric scenario can be shared by an adaptive model selection criterion. A similar two-stage adaptive model selection rule for time series autoregression has been proposed by Ing [44]. However, Yang [73, 75] proved that no model selection procedure can be both consistent (or pointwise adaptive) and minimax-rate optimal at the same time. More recently, Erven, Grünwald, de Rooij [26] found that if a cumulative risk (i.e., the sum of risks from the sample size 1 to n) is considered instead of the usual risk at sample size n, then the conflict between

consistency in selection and minimax-rate optimality can be resolved by a Bayesian strategy that allows switching between models. As will be seen, if we can properly distinguish between parametric and nonparametric scenarios, a consequent choice of AIC or BIC simultaneously achieves asymptotic efficiency for both parametric and nonparametric situations.

2.1.2 Model selection: A gap between theory and practice

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

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

Model selection is fundamentally important for statistics or even sciences. Indeed, the conflict between AIC and BIC has received a lot of attention not only in the statistics literature but also in fields such as psychology and biology [64]. There has been a lot of debate in the literature regarding model selection procedures that is not only statistical but also philosophical, especially about the existence of a true model and the ultimate goal of statistical modeling (see, e.g., [8, 14, 17, 33, 81], and references therein). Some researchers have no problem with assuming existence of a

true finite-dimensional model while others regard it as a fiction.

We would like to quote a leading statistician here:

"*It does not seem helpful just to say that all models are wrong. The very word model implies simplification and idealization. The idea that complex physical, biological, and sociological systems can be exactly described by a few formulae is patently absurd. The construction of idealized representations that capture important stable aspects of such systems is, however, a vital part of general scientific analysis and statistical models, especially substantive ones (Cox, 1990), do not seem essentially different from other kinds of model.* " (Cox [21])

Fisher in his pathbreaking 1922 paper [32], provided thoughts on the foundations of statistics, including model specification. He stated: "More or less elaborate forms will be suitable according to the volume of the data". Cook [20] discussed Fisher's insights in details.

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

We have mixed feelings towards the concept of consistency (i.e., the property that the true model, assumed to be among the candidates, will be selected with probability going to 1 as the sample size goes to infinity). One indeed should not read too much into it either for interpretation or for estimation because the asymptotic view is often overly optimistic. First, the property of consistency often comes with restrictive assumptions among which is the existence of a finite-dimensional true model, which is strongly objected by many. Second, even if the finite-dimensional true model assumption is justifiable, a consistent model selection method does not necessarily

perform well due to the often high model selection uncertainty and the necessary lack of protection in the worst case of every consistent model selection method as a price paid to aggressively pursue a lower-dimensional alternative to explain the data.

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

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

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

While there are many different performance measures that we can use to assess if one model stands out, following our results on distinguishing between parametric and nonparametric scenarios, we focus on an estimation/prediction accuracy measure. We call it *parametricness index* (PI), which is relative to the list of candidate models and

the sample size. Our theoretical results show that this index converges to infinity for a parametric scenario and converges to 1 for a typical nonparametric scenario. Our suggestion is that when the index is significantly larger than 1, we can treat the selected model as a stably standing out model from the estimation perspective. Otherwise, the selected model is just among a few or more equally well-performing candidates. We call the former case practically parametric and the latter practically nonparametric.

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

be practically nonparametric for others.

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

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

2.2 Setup of the regression problem

Consider the regression model

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

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

To estimate the regression function, a list of linear models are being considered,

from which one is to be selected:

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

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

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

Parametric modeling typically intends to capture the essence of the data by a finite-dimensional model, and nonparametric modeling tries to achieve the best trade-off between approximation error and estimation error for a target infinite-dimensional function. See, e.g., [77] for general relationship between rate of convergence for function estimation and full or sparse approximation based on a linear approximating system.

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

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

Notation and definitions

Let $\mathbf{Y}_n = (Y_1, \cdots, Y_n)^T$ be the response vector and M_k be the projection matrix for model k. Denote $\hat{\mathbf{Y}}_k = M_k \mathbf{Y}_n$. Let $f_n = (f(x_1), \cdots, f(x_n))^T$, $e_n = (\epsilon_1, \cdots, \epsilon_n)^T$,

and I_n be the identity matrix. Let $\|\cdot\|$ denote the Euclidean distance in the R^n space, and let $TSE(k) = \|f_n - \hat{\mathbf{Y}}_k\|^2$ be the total square error of the LS estimator from model k.

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

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

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

when σ is known, and

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

when σ is estimated by $\hat{\sigma}$. A discussion on choice of λ_n and d will be given later in Section Section 2.3.5. We emphasize that our use of $IC_{\lambda_n, d}(k)$ or $IC_{\lambda_n, d}(k, \hat{\sigma}^2)$ is for defining the parametricness index as below and it may not be the one used for model selection.

2.3 Main Theorems

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

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

2. When σ is estimated by $\hat{\sigma}$,

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

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

2.3.1 Parametric Scenarios

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

Conditions:

(P1). There exists $0 < \tau \leq \frac{1}{2}$ such that A_n is of order $n^{\frac{1}{2}+\tau}$ or higher.

(P2). The dimension of the true model does not grow too fast with sample size n in the sense that $r_{k_n^*} \lambda_n \log(n) = o(n^{\frac{1}{2}+\tau})$.

(P3). The selection procedure is consistent: $P(\hat{k}_n = k_n^*) \to 1$ as $n \to \infty$.

Theroem 1

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

(i). With σ^2 known, we have

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

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

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

Remarks:

1. If the number of models of each dimension is of a polynomial order in n, then Condition (P1) can be relaxed. For example, it is sufficient to require A_n of an order higher than $n^{\frac{1}{2}}(\log(n))^{\lambda}$ for some $\lambda > 1$. Note that in a conventional case with Γ fixed, A_n is typically of order n and (P1) certainly holds. The conditions basically eliminates the case that the true model and a sub-model with one fewer term are not distinguishable with the information available in the sample.

2. In our formulation above, we considered comparison of two immediately nested models. One can consider comparing two nested models with size difference m ($m > 1$) and similar results hold.

3. The case $\lambda_n = 1$ corresponds to using BIC in defining the PI. And $\lambda_n = 2/\log(n)$ corresponds to using AIC.

4. When the number of predictors increases with n, Chen and Chen [18] showed that a higher penalty than BIC leads to consistency in all subset selection under some conditions.

2.3.2 Nonparametric Scenarios

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

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

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

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

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

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

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

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

Theroem 2

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

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

Remarks:

1. The first part of Condition (N1) says that the dimension of the selected model lies in a range $[a_n, b_n]$ with probability going to 1, where $a_n \to \infty$. For nonparametric regression, for familiar model selection methods, the order of $r_{\hat{k}_n}$ can be identified (e.g., [44, 77]), sometimes loosing a logarithmic factor. The requirement $\frac{B_{j,n}}{n-j} \to 0$ is easily satisfied in a typical nonparametric scenario.

2. Condition (N2) basically ensures that the number of subset models of each dimension does not grow too fast relative to $B_{j,n}$. When the best model has a slower rate of convergence in regression estimation, more candidate models can be allowed without detrimental selection bias.

3. Roughly speaking, Condition (N3) says that when the model dimension is in a range that contains the selected model with probability approaching 1, the least significant term in the regression function projection is negligible compared to the sum of approximation error, the dimension of the model times $\lambda_n \log(n)$, and the term $dn^{1/2}\log(n)$. This condition is mild.

4. A choice of $d > 0$ can handle situations where the approximation error decays fast, e.g., exponentially fast (see Section 3.4), in which case the stochastic fluctuation of $IC_{\lambda_n,d}$ with $d = 0$ is relatively too large for PI to converge to 1 in probability. In applications, for separating reasonably distinct parametric and nonparametric scenarios, we recommend the choice of $d = 0$.

5. Except for (N1), no specific properties of the model selection rule are assumed for this result.

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

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

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

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

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

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

Theroem 3

Assuming Conditions $(N1)$, (N2′), and (N3′) hold for a nonparametric scenario, then we have

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

Remarks:

1. Conditions (N2′) and (N3′) have similar interpretations as $(N2) - (N3)$ except that $B_{j,n}$ is replaced by $E_{j,n}$. Conditions $(N2')$ and $(N3')$ are stronger than $(N2)$ and $(N3)$ due to estimation of σ^2.

2. For the σ^2 unknown case, we used $\hat{\sigma}_n^2$ from the selected model and hence condition $(N3')$. The conditions $(N2')$ and $(N3')$ may be relaxed if more is known

about $\hat{k}_n$, which is possible under smoothness assumptions on the regression function.

2.3.3 PI separates parametric and nonparametric scenarios

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

Corollary 1

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

2.3.4 Examples

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

(1). All subset selection

Let p_n be the number of terms to be considered.

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

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

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

In this case, both $r_{k_n^*}$ and p_n go to infinity with the sample size n. Since there are more and more terms in the true model, in order for A_n not to be too small, the terms should not be too highly correlated. Otherwise, the true model may not be distinguishable from a sub-model based on the data. An extreme case is that one term in the true model is almost linearly dependent on the others. Then $A_n \approx 0$. To understand Condition (P1) in terms of the coefficients in the true model, under an orthogonal design, Condition (P1) is more or less equivalent to that the square of the smallest coefficient in the true model is of order $n^{\tau-1/2}$ or higher. Since τ can be arbitrarily close to 0, the smallest coefficient should basically be larger than $n^{-\frac{1}{4}}$.

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

(iii). Nonparametric.

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

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

(2). Order selection in series expansion

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

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

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

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

Then a sufficient condition for Condition (N3) is that $d = 0$ and $\lim_{k\to\infty} \frac{\beta_k^2}{\sum_{i=k+1}^{\infty} \beta_i^2} = 0$, which is true if $\beta_k = k^{-\delta}$ for some $\delta > 0$ but not true if $\beta_k = e^{-ck}$ for some $c > 0$. When β_k decays faster so that $\frac{\beta_k^2}{\sum_{i=k+1}^{\infty} \beta_i^2}$ is bounded away from zero and $\sup_{a_n \le k \le b_n} |\beta_k| = o\left(\frac{\sqrt{\log(n)}}{n^{1/4}}\right)$, any choice of $d > 0$ makes Condition (N3) satisfied. An example is the exponential-decay case, i.e., $\beta_k = e^{-ck}$ for some $c > 0$. According to [44], when $\hat{k}_n$ is selected by BIC for order selection, we have that $r_{\hat{k}_n}$ basically falls within a constant from $\frac{1}{2c}\log(n/\log(n))$ in probability. In this case, $\beta_k \approx \frac{\sqrt{\log(n)}}{n^{1/2}}$ for $k \approx \frac{1}{2c}\log(n/\log(n))$. Thus Condition (N3) is satisfied.

2.3.5 On the choice of λ_n and d

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

2.3.6 Combining strengths of AIC and BIC

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

It is well-known that AIC is asymptotically loss (or risk) efficient for nonparametric scenarios and BIC is consistent when there are finite-dimensional correct models (see [18] for a recent result), which implies that BIC is asymptotically loss efficient [59].

Corollary 2

For a given number $c > 1$, let δ be the model selection procedure that chooses either the model selected by AIC or BIC as follows:

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

Under Conditions P1-P3, N1-N3, δ is asymptotically loss efficient in both paramet-

ric and nonparametric scenarios as long as AIC and BIC are loss efficient for the respective scenarios. □

Remarks:

1. In application, for a finite sample, a good cutoff point c needs to be chosen. We will explore this in numerical examples.

2. Previous work on sharing the strengths of AIC and BIC utilized minimum description length criterion in an adaptive fashion ([7, 37]), or flexible priors in a Bayesian framework ([34, 26]). To our knowledge, only Ing [44] and Yang [75] established (independently) simultaneous asymptotic efficiency for both parametric and nonparametric scenarios. Differently from our work here, their main idea to separate parametric and nonparametric scenarios is to compare the models selected by BIC at different sample sizes. In contrast, our method in this paper compares neighboring models at the full sample size.

2.4 PI as a model selection diagnostic measure, i.e., *P*ractical *I*dentifiability of the best model

Based on the theory presented in the previous section, it is natural to use the simple rule for answering the question if we are in a parametric or non-parametric scenario: call it parametric if PI is larger than c for some $c > 1$ and otherwise nonparametric. Theoretically speaking, we will be right with probability going to one.

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

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

From a theoretical perspective, the difference between parametric and nonparametric modeling is quite clear in this context. Indeed, when one is willing to assume that the data come from a member in a parametric family, the focus is then naturally on the estimation of the parameters, and finite-sample and large sample properties (such as UMVUE, BLUE, minimax, Bayes, and asymptotic efficiency) are well understood. For nonparametric estimation, given infinite-dimensional smooth function classes, various approximation systems (such as polynomial, trigonometric and wavelets) have been shown to lead to minimax-rate optimal estimators via various statistical methods (e.g., [9, 24, 43, 65]). In addition, given a function class defined in terms of approximation error decay behavior by an approximating system (or smoothness of the function), rates of convergence of minimax risks have been established (see, e.g., [77]). As is expected, the optimal model size (in rate) based on linear approximation depends on the sample size (and other things) for a nonparametric scenario. In particular, for full and sparse approximation sets of functions, the minimax theory shows that for a typical nonparametric scenario, the optimal model size makes the approximation error (squared bias) roughly equal to estimation error (model dimension over the sample size) [77]. Furthermore, adaptive estimators that are simultaneously optimal for multiple function classes can be obtained by model selection or model combining (see, e.g, [5, 72] for many references).

From a practical perspective, unfortunately, things are much less clear. Consider, for example, the simple case of polynomial regression. In linear regression textbooks,

one often finds data that show obvious linear or quadratic behavior, in which case perhaps most statisticians would be unequivocally happy with a linear or quadratic model (think of Hooke's law for describing elasticity). When the underlying regression function is much more complicated so as to require 4th or 5th power, it becomes difficult to classify the situation as parametric or nonparametric. While few (if any) statisticians would challenge the notion that in both cases, the model is only an approximation to reality, what makes the difference in calling one case parametric quite comfortably but not the other? Perhaps simplicity and stability of the model play key roles as mentioned in Cox [21]. Roughly speaking, when a model is simple and fits the data excellently (e.g, with R^2 close to 1) so that there is little room to significantly improve the fit, the model obviously stands out. In contrast, if we have to use a 10th order polynomial to be able to fit the data with 100 observations, perhaps few would call it a parametric scenario. Most of the situations may be in between.

Differently from the order selection problem, the case of subset selection in regression is substantially more complicated due to the much increased complexity of the list of models. It seems to us that when all subset regression is performed, it is usually automatically treated as a parametric problem in the literature. While this is not surprising, our view is different. When the number of variables is not very small relative to the sample size and the error variance, the issue of model selection does not seem to be too different from order selection for polynomial regression where a high polynomial power is needed. In our view, when analyzing data (in contrast to asymptotic analysis), if one explores over a number of parametric models, it is not necessarily proper to treat the situation as a parametric one, by which we mean the standard practice of reporting the standard errors and confidence intervals for parameters and making interpretations based on the selected model.

Closely related to the above discussion is the issue of model selection uncertainty

(see, e.g., [13, 17]). It is now well recognized that model selection uncertainty should not be conveniently ignored as is still the dominating practice in real world statistical applications. It is an important issue to know when we are in a situation where a relatively simple and reliable model stands out in a proper sense and thus can be used as the "true" model for practical purposes, and when a selected model is just one out of multiple or even many possibilities among the candidates at the given sample size. In the first case, we would be willing to call it parametric (or more formally, practically parametric) and the latter (practically) nonparametric.

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

Based on hypothesis testing theories, obviously, at a given sample size, for any true parametric distribution in one of the candidate families from which the data are generated, one has a nonparametric distribution (i.e., not in any of the candidate

families) that cannot be distinguished from the true distribution. From this perspective, pursuing a rigid finite-sample distinction between parametric and nonparametric scenarios is improper.

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

To summarize, for application,

1. We adopt a pragmatic view on the contrast between parametric and nonparametric scenarios when a number of parametric models are considered to analyze the data. In our view, when a candidate model stably stands out in proper measures, we can call it practically parametric, and in contrast, when the selected model is behaving similarly to some alternative models among the candidates and it is merely providing a sample-size sensitive balancing of the approximation and estimation error, we call it nonparametric to indicate the lack of an outstanding candidate.

2. Note that our assessment of practical parametricness/nonparamreticness depend only on data and the candidate models, but not on assumptions on the true model.

3. PI can be viewed as a model selection diagnostic measure (or practical identifiability of the best model), which tells us, roughly, for the time being and foreseeable future, if the selected model deserves to be called a parametric model based on which parameter estimates and standard errors are to be reported. When PI is close to 1, it indicates that it is not OK to treat the selected model as the "truth", and interpretations based on the selected model, no matter how convenient, should not be treated too seriously.

2.5 Simulation Results

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

2.5.1 Single predictor

Example 1

Compare two different situations:

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

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

BIC is used to select the order of polynomial regression between 1 and 30. The estimated σ from the selected model is used to calculate the PI. Representative scatterplots at $n = 200$ with $\sigma_1 = 3$, $\sigma_2 = 7$ can be found in Figure Figure 2.1. Note that the function estimate based on the selected model by BIC is visually more different from that based on the smaller model with one fewer term for the parametric scenario than the nonparametric one.

Quantiles for the PIs in both scenarios based on 300 replications are presented in Table Table 2.1.

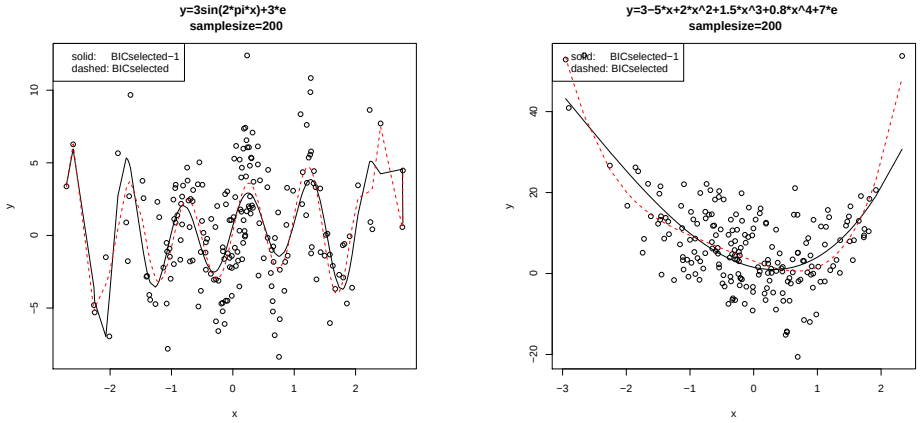
Example 2

Compare the following two situations:

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

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

The two mean functions are the same except the coefficient of the $\sin(2\pi x)$ term.



[h!]

Figure 2.1: Scatterplots For Example 1

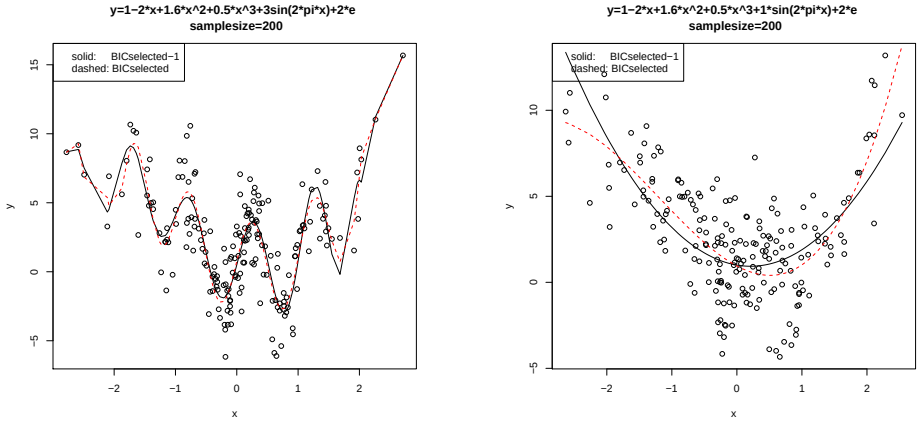
Table 2.1: Percentiles of PI for Example 1

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

Scatterplots and table similar to those for Example 1 are in Figure Figure 2.2 and Table Table 2.2, respectively. As we can see from Table Table 2.2, although both cases are of a nonparametric nature, they have different behaviors in terms of model selection uncertainty and PI values. Case 2 can be called 'practically' parametric and the large PI values provide information in this regard.

2.5.2 Factors that influence PI

As we know, most model selection problems, if not all, are affected by many factors like the regression function itself, the noise level, and the sample size. We expect that these



[h!]

Figure 2.2: Scatterplots For Example 2

Table 2.2: Percentiles of PI for Example 2

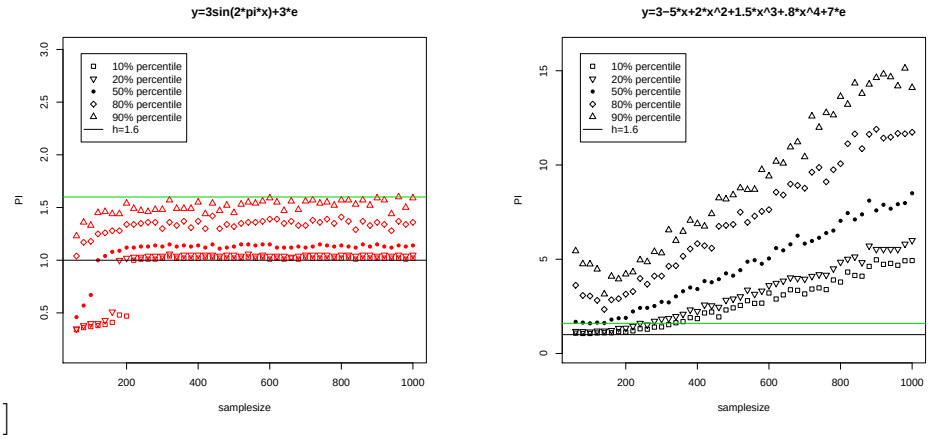
	case 1			**case 2**		
percentile	order selected	PI	$\hat{\sigma}$	order selected	PI	$\hat{\sigma}$
10%	15	1.01	1.87	3	1.75	1.99
20%	15	1.05	1.92	3	2.25	2.03
50%	16	1.14	2.00	3	3.51	2.12
80%	17	1.4	2.11	3	5.33	2.22
90%	18	1.63	2.17	3	6.62	2.26

factors influence the behavior of PI as well. We investigate the effects of these factors on PI and report some representative results below. As we will see, PI, as a diagnostic measure, can tell us whether a problem is 'practically' parametric/nonparametric due to the influences of all the factors that affect model selection.

The effect of sample size

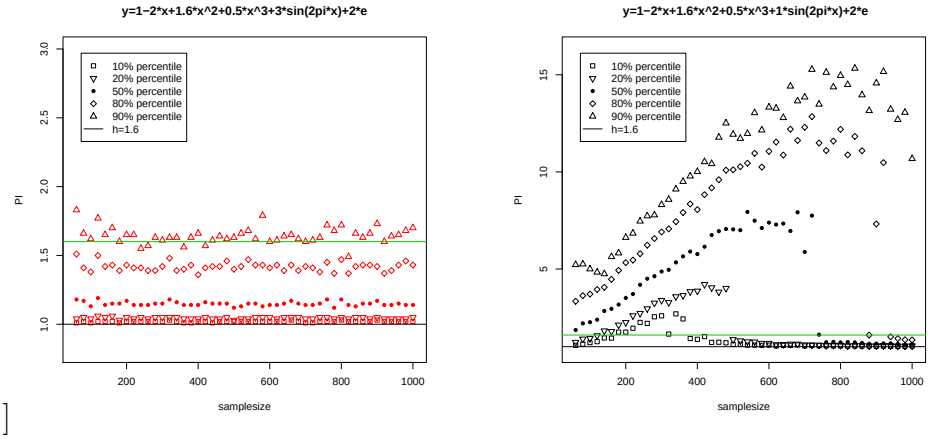
We calculated the PI at different sample sizes with 300 replications for each and report the results for Examples 1 and 2 in Figures Figure 2.3 and Figure 2.4.

For Example 1, from Figure Figure 2.3, we see the PIs in case 1 basically fall in



[h!]

Figure 2.3: Sample size effect for Example 1

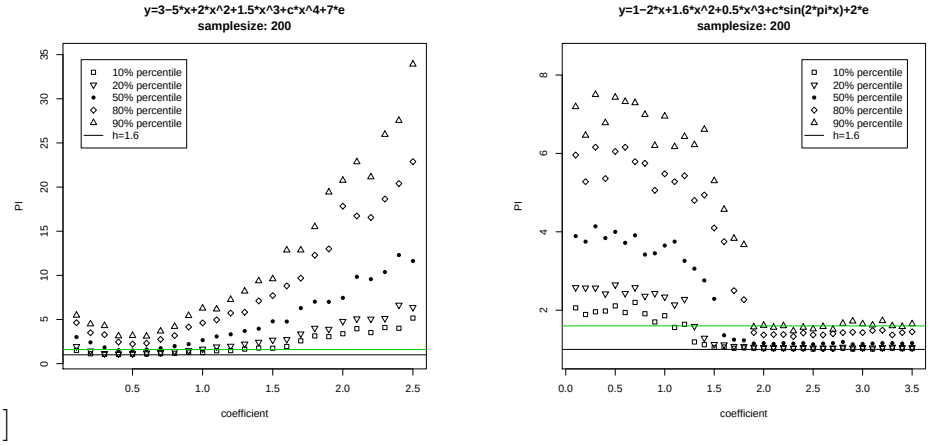


[h!]

Figure 2.4: Sample size effect for Example 2

between 1 and 1.6, whereas the PIs in case 2 become larger as sample size increases.

From Figure Figure 2.4, in case 2 of Example 2, the PIs first increase and then drop down as the sample size increases. This is due to the fact that in the beginning, the sine term is better to be ignored due to lack of information, and when the sample size is bigger, say 300-400, the PI indicates a strong parametric scenario. With a sample size in this range, the problem is 'practically' parametric. With more and more data we are then gradually able to detect the signal of the $\sin(2\pi x)$ term, thus capturing the nonparametric nature of the mean function.



[h!]

Figure 2.5: Effect of coefficient

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

The examples show that given the regression function and the noise level, the value of PI indicates whether the problem is 'practically' parametric/nonparametric at the current sample size.

The effect of coefficient

We study the PIs for different values of the coefficient of the last term in case 2 of Example 1 and Example 2, respectively. The results are reported in Figure Figure 2.5.

For Example 1, the values of PI first decrease and then increase as the coefficient of x^4 increases. This is because when the coefficient for the term of x^4 is small (less than .5), the true mean function behaves just like a polynomial of order 3 at the current sample size. As the coefficient gets slightly larger, there is no clear distinction between a polynomial of order 3 and a polynomial of order 4 at the current sample size. That is why we see the PIs drop a little in the beginning. However, when the

coefficient gets bigger than .5 or .6, then we can detect the term of x^4 and the PIs increase with the coefficient. Overall, the PI values are mostly larger than 1.6 in this example.

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

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

2.5.3 Multiple predictors

Now we study several examples with multiple predictors. The first two examples were used in the original lasso paper [67].

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

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

where $\mathbf{x}$ is generated from a multivariate normal distribution with mean 0, variance 1, and correlation structure given in each example. For each generated data set, we apply the Branch and Bound algorithm [36] to do all subset selection by BIC and then calculate the PI value (part of our code is modified from the aster package of Geyer [35]). Unless otherwise stated, in these examples, the sample size is 200 and

we replicate 300 times.

Example 3

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

Example 4

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

Example 5

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

Example 6

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

Example 7

This example is the same as example 3 except that we add a nonlinear component in the mean function and $\sigma = 3$, i.e., $Y = \beta^T \mathbf{x} + \phi(u) + \sigma\epsilon$, where $u \sim uniform(-4, 4)$ and $\phi(u) = 3(1 - 0.5u + 2u^2)e^{-u^2/4}$. All subset selection is carried out with predictors $x_1, \cdots, x_8, u, \cdots, u^8$ which are coded as 1-8 and A-G in the Table Table 2.3.

The selection behaviors and PI values are reported in Table Table 2.3 and Table Table 2.4, respectively. From those results, we see that the PIs are large for Example 3 and small for Example 4. Note that in Example 3 we have 82% chance selecting the true model, while in Example 4 the chance is only 12%. Although both Example

Table 2.3: Proportion of selecting true model

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

Table 2.4: Quartiles of PIs

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

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

Example 7 is quite interesting. Previously, without the $\phi(u)$ component, even at $\sigma = 5$, large values of PI are seen. Now with the nonparametric component present, the PI values are close to 1. (The asterisk mark (*) in Table Table 2.3 indicates the model is the most frequently selected one instead of being the true model.)

An illuminating example

We now look at a special example. We still generate data from a linear model with $\beta = (2, 2, 0.3, 0.3, 0.1, 0.1, 0, 0, 0, 0)^T$ and $\sigma = 2$. The pairwise correlation among the predictors is 0.5. For this example we do all-subset selection by BIC at different

sample sizes. Our thinking is that since some of the coefficients are large and others are small, BIC is going to pick up the significant predictors gradually as the sample size increases. We expected to see both big and small PI values alternating to some degree when the sample size changes. In this example, we replicate 500 times for each sample size.

The results of median PIs at different sample sizes are shown in figure Figure 2.6. From the plot we see PI first increases with the sample size, then decreases, then increases and decreases again, and finally increases. This is because when the sample size is small, most of the time BIC only picks up x_1 and x_2 and the PI increases with the sample size. As the sample size further increases, BIC finds the predictors x_3 and x_4 relevant and the PI then decreases since the coefficients for x_3 and x_4 are small (but not too small) so that BIC is not quite sure about the best model. When the sample size gets big enough so that most of the times BIC chooses x_1, x_2, x_3, and x_4, the PI increases again with sample size. A similar story repeats for the predictors x_5, and x_6. If we choose 1.2 as a cutoff point, we would see (practically) parametric and (practically) nonparametric scenarios alternating as the sample size changes.

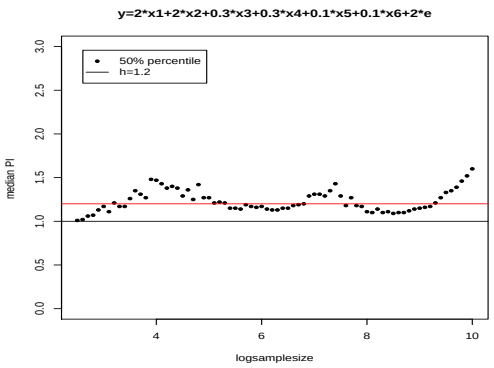


Figure 2.6: Behavior of PI for the special example

Inference after model selection (PI as practical identifiability)

With Examples 3 and 4, we assess the accuracy of statistical inferences after model selection. We first generate an evaluation set of predictors from the same distribu-

tion as that for observations. Then for each replication, we generate data and do subset selection by BIC. After selecting a model, we get the resulting predicted value (estimated regression function at the given point), the standard error and the 95% confidence interval for the regression estimate at each of the new design points in the evaluation set. We replicate 500 times. Then at each new design point we calculate the actual standard deviation (which is called se.fit.t in the output table) of the 500 regression estimates and compare that to quantiles of the 500 standard errors that are obtained based on the selected model. We also take a look at the actual coverage of the 95% CIs for the true mean of the regression function at each new design point. Results at 10 randomly chosen new design points are reported in Table Table 2.5 and Table Table 2.6 for the two examples.

Table 2.5: Reliability of inference for Example 3

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

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

Table 2.6: Reliability of inference for Example 4

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

several evaluation data sets with different sizes, the results are similar.

It is now well known that model selection has an impact on subsequent statistical inferences (see, e.g., [80, 40, 31, 46]). For observational data, typically one cannot avoid making various modeling choices (such as which type of statistical analysis to pursue, which kind of models to consider) after seeing the data, but their effects are very difficult to quantify. Thus it can be very helpful to know when the choices have limited impact on the final results. The above results together with Tables Table 2.3 and Table 2.4 show that the value of PI can provide valuable information on the parametricness of the underlying regression function and hence on how confident we are on the accuracy of subsequent inferences.

Combining strengths of AIC and BIC based on PI

Still with Examples 3-7, we investigate the performance of an adaptive choice between AIC and BIC based on the PI value. Again we first generate an evaluation data set with 500 new design points from the same distribution as the one for observations. Then for each replication, we use both AIC and BIC to select a model. The combined

procedure is BIC if the PI value is larger than a cutoff point (chosen as 1.2 in these examples) and AIC otherwise. Then for each procedure (AIC, BIC, and the combined) in each replication, we calculate the average squared error (which is the average squared difference between the true regression mean and the fitted value based on the selected model) at the new design points in the evaluation data. We replicate 500 times and the statistical risk is estimated to be the average of the 500 average squared errors. The risk ratios are reported in Table Table 2.7 with BIC as the reference.

Table 2.7: Statistical risks of AIC, BIC, and the Combined procedure

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

From the results we see in all these examples the combined procedure shows capability of adaptation between AIC and BIC in terms of the statistical risk. We also see from Tables Table 2.7, Table 2.3 and Table 2.4 that in Examples 5 and 7, the PIs are roughly around 1.2 and AIC and BIC have similar performance in terms of statistical risks, while in the other examples the PIs are either large or small and correspondingly, either BIC or AIC has a smaller statistical risk. These results show that PI provides helpful information regarding whether AIC or BIC works better or they have similar performances in statistical risks. Therefore, PI can be viewed as a **P**erformance **I**ndicator of AIC versus BIC.

2.5.4 A summary

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

1. Factors other than the nature of the regression function also influence the value of PI, including the sample size and the noise level. From a practical point of view, PI, as a diagnostic measure, indicates whether a specific problem is 'practically' parametric/nonparametric with the influences of all those factors.

2. Model selection effect on subsequent statistical inferences may or may not be reasonably ignored, and the value of PI provides useful information in that regard.

3. For Examples 3 and 4, both being parametric, one is practically parametric and the other practically nonparametric for $n = 200$. Correspondingly, BIC works better for the former and AIC for the latter in terms of risk for estimating the regression function. This phenomenon will be seen again in the next section.

4. Combining AIC and BIC based on the PI value shows adaptation capability in terms of statistical risk. That is, the composite rule yields a risk close to the better one of AIC and BIC.

5. In nested model problems (like order selection of series expansion), a cutoff point of $c = 1.6$ seems to be good. In subset selection problems, we expect the cutoff point to be smaller since the infimum is taken over many models. The choice of 1.2 seems to be reasonably good based on our numerical investigations, which is also supported by the observation that when PI is around 1.2, AIC and BIC perform similarly.

2.6 Real Data Examples

In this section, we study three data sets: the Ozone data (e.g. [12]), the Boston housing data (e.g. [38]), and the Diabetes data (e.g. [25]).

In these examples, we conduct all subset selection by BIC using the Branch and Bound algorithm. Besides finding the PI values for the full data, we also do the same with sub-samples from the original data at different sample sizes. In addition, we carry out a parametric bootstrap from the model selected by BIC based on the original data to assess the stability of model selection. (The design points of the predictors are randomly selected with replacement from the original data.) Like in the multiple-predictor simulation study, we also combine AIC and BIC based on the PI value when doing parametric bootstrap. Unless otherwise stated, the subsampling and the bootstrap are both replicated 500 times at each sample size. (In the results, the predictors are coded to be a single digit between 1 and 9 and then a single capital letter between 'A' and 'Z', i.e, letter 'A' stands for the 10th predictor, 'B' for the 11th, and so on.)

Ozone Data

There are 9 variables with 8 predictors and 330 observations. We followed the transformations of the predictors and the response suggested by Hawkins [39]. (In that paper a ninth predictor, day of the year, was also included. We left this predictor out as many others did. See [12].) After the transformations, we have 10 predictors with quadratic terms of two predictors added.

Boston Housing Data

The data consists of 14 variables (1 response and 13 predictors). There are 506 observations. We followed the transformations of the variables in Harrison and Rubinfeld's

paper [38].

Diabetes Data

There are 11 variables with 10 predictors and 442 observations.

The PIs from the original data for these three examples are: 1.277 (ozone), 1.028 (Boston housing), and 1.298 (diabetes). The results of subsampling and bootstrap are reported in Tables Table 2.8-Table 2.9 and Tables Table 2.10-Table 2.11, respectively.

Table 2.8: Quartiles of PIs from subsamples of size 400

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

Table 2.9: Quartiles of PIs from subsamples of size 200

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

Table 2.10: The 6 most frequently selected models and their frequencies with a sample size of 400

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

From the tables, we see the PIs for the ozone data are mostly larger than 1.2, while those for the Boston housing data are smaller than 1.2. Moreover, the parametric bootstrap suggests that for the Ozone data, the model selected from the full data

Table 2.11: The 6 most frequently selected models and their frequencies with a sample size of 200

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

still reasonably stands out even when the sample size is reduced to about 200 and noises are added (not all shown due to space limitation). For the Boston housing data, however, even at a sample size of 400 we only have 28% chance selecting the same model as the one selected with the full data. Interestingly, the diabetes data exhibit a parametric behavior when $n = 400$, but with the sample size reduced by half, it looks more like a nonparametric scenario.

Combining AIC and BIC based on PI

Table 2.12: Combining AIC and BIC based on PI with full sample size

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

Similar to the simulation results in Section 5, by parametric bootstrap at the original sample size from the selected model, in these data examples, combining AIC and BIC based on PI shows good overall performance in terms of statistical risk

(Table Table 2.12). The combined procedure has a statistical risk close to the better one of AIC and BIC in each case.

2.7 Conclusions

Parametric models have been commonly used to estimate a finite-dimensional or infinite-dimensional function. While there have been serious debates on which model selection criterion to use to choose a candidate and there has been some work on combining the strengths of very distinct model selection methods, there is a major lack of understanding on statistically distinguishing between scenarios that favor one method (say AIC) and those that favor another (say BIC). To address this issue, we have derived a parametricness index (PI) that has the desired theoretical property: PI converges in probability to infinity for parametric scenarios and to 1 for nonparametric ones. The use of a consistent model selection rule in constructing PI effectively prevents overfitting when we are in a parametric scenario. The comparison of the selected model with a subset model separates parametric and nonparametric scenarios through the distinct behaviors of the approximation errors of these models in the two different situations.

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

We advocate a practical view on parametricness/nonparametricness. In our view,

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

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

Identifying a stable and strong message in data as is expressed by a meaningful parametric model, if existing, is obviously important. In biological and social sciences, especially observational studies, a strikingly reliable parametric model is often too much to ask for. Thus, to us, separating scenarios where one model is reasonably standing out and is expected to shine over other models for sample sizes not too much larger than the current one from those where the selected model is simply the lucky one to be chosen among multiple equally performing candidates is an important step beyond simply choosing a model based on one's favorite selection rule or, in the opposite direction, not trusting any post model selection interpretation due to existence of model selection uncertainty.

For the other goal of regression function estimation, in application, one typically applies a model selection method, or considers estimates from two (or more) model

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

It has been suggested that AIC performs better for a nonparametric scenario and BIC better for a parametric one (see [75] for a study on the issue in a simple setting). This is asymptotically justified but certainly not quite true in reality. Our numerical results have demonstrated that for some parametric regression functions, AIC is much better. On the other hand, for an infinite-dimensional regression function, BIC can give a much more accurate estimate. Regarding this discrepancy between asymptotics and finite-sample reality, one typically explains that of course finite-sample behaviors can be totally different from asymptotic ones. Our numerical results tend to suggest that a much more helpful statement is: when PI is high and thus we are in a practical parametric scenario (whether the true regression function is finite-dimensional or not), BIC tends to be better for regression estimation; when PI is close to 1 and thus we are in a practical nonparametric scenario, AIC tends to be better. We feel that the use of PI as a suitable indicator of the relative performance of AIC and BIC is a positive step forward towards a data-driven sound choice of a model selection method.

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

2.8 Proofs

The following two facts will be used in our proofs (see [71]).

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

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

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

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

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

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

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

$$P_{k^{(s)},k} = \alpha_{New} \cdot \alpha^T_{New} / \|\alpha_{New}\|^2. \tag{2.1}$$

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

For the ease of notation, we denote $P_{k^{(s)},k}$ by P, $rem_1(k) = e_n^T(f_n - M_k f_n)$, and

$rem_2(k) = \|(I_n - M_k)e_n\|^2/\sigma^2 - n$ in the proofs . Then

$$\begin{aligned} \|(I_n - M_{k^{(s)}})e_n\|^2 &= \|(I_n - M_k)e_n\|^2 + \|Pe_n\|^2 \quad (2.2)\\ \|(I_n - M_{k^{(s)}})f_n\|^2 &= \|(I_n - M_k)f_n\|^2 + \|Pf_n\|^2 \quad (2.3)\\ rem_1(k^{(s)}) &= rem_1(k) + e_n^T P f_n \quad (2.4) \end{aligned}$$

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

Proof 2.1 (Proof of Theorem 1 (parametric, σ known))
Under the assumption that $P(\hat{k}_n = k_n^*) \to 1$, we have $\forall \epsilon > 0$, $\exists\ n_1$ such that $P(\hat{k}_n = k_n^*) > 1 - \epsilon$ for $n > n_1$.

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

Observe

$$\|\mathbf{Y}_n - \hat{\mathbf{Y}}_k\|^2 = \|(I_n - M_k)f_n\|^2 + \|(I_n - M_k)e_n\|^2 + 2rem_1(k). \quad (2.5)$$

Then

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

By Fact 2,

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

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

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

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

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

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

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

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

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

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

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

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

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

that $IC(k_n^{*(s)})$ is positive for $n > n_3$ with probability higher than $1 - e^{-\frac{1}{8}n^{2\tau}h_n^2} - e^{-c^2\|(I-M_{k_n^{*(s)}})f_n\|^2/2}$.

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

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

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

And for $n > n_3$,

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

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

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

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

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

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

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

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

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

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

Proof 2.2 (Proof of Theorem 2 (nonparametric, σ known))
Similar to the proof of Theorem 1, consider $\frac{IC(\hat{k}_n^{(s)})}{IC(\hat{k}_n)}$ for any $\hat{k}_n^{(s)}$ being a sub-model of $\hat{k}_n$ with one fewer term, and we have

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

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

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

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

Similarly, $\forall c_2 > 0$,

$$P\left(\frac{|e_n^T Pf_n|}{B_{r_{k_n}}} \geq c_2\right) \leq e^{-\frac{c_2^2 B_{r_{k_n}}^2}{2\|Pf_n\|^2}} \leq e^{-c_2^2 B_{r_{k_n}}/2} \quad (\text{if } \|Pf_n\|^2 \leq B_{r_{k_n}}), \tag{2.6}$$

$$P\left(\frac{|e_n^T Pf_n|}{\|Pf_n\|^2} \geq c_2\right) \leq e^{-\frac{c_2^2 \|Pf_n\|^2}{2}} \leq e^{-c_2^2 B_{r_{k_n}}/2} \quad (\text{if } \|Pf_n\|^2 > B_{r_{k_n}}). \tag{2.7}$$

By Fact 2,

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

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

$$P\left(\|(I_n - M_{k_n})e_n\|^2 - (n - r_{k_n}) \leq -\gamma_n B_{r_{k_n}}\right) \leq e^{-\frac{\gamma^2 B_{r_{k_n}}^2}{4(n-r_{k_n})}}. \tag{2.8}$$

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

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

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

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

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

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

$$\sum_{j=a_n}^{b_n} \left\{ N_j \cdot \left(e^{-\frac{\gamma^2 B_{j,n}^2}{4(n-j)}} + e^{-c_1^2 B_{j,n}/2} \right) + N_j \cdot L_j \cdot \left(e^{-D_n/2} + e^{-c_2^2 B_{j,n}/2} \right) \right\} \leq \epsilon. \tag{2.9}$$

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

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

Note that

$$PI_n = 1 + \inf_{\hat{k}_n^{(s)}} \frac{\|Pf_n\|^2 + \|Pe_n\|^2 + e_n^T P f_n - \lambda_n \log(n)}{\|(I_n - M_{\hat{k}_n})f_n\|^2 + rem_2(\hat{k}_n) + 2rem_1(\hat{k}_n) + \lambda_n \log(n) r_{\hat{k}_n} + dn^{1/2} \log(n)}. \tag{2.10}$$

Also with probability higher than $1 - 2\epsilon$, the denominator in equation ((2.10)) is bigger than $(1 - 2c_1)\left[\|(I_n - M_{\hat{k}_n})f_n\|^2 + (\lambda_n \log(n) - 1)r_{\hat{k}_n} + dn^{1/2} \log(n)\right] - \gamma B_{r_{\hat{k}_n}}$. Thus when $2c_1 + \gamma < 1$, the denominator in ((2.10)) is positive.

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

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

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

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

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

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

Similarly,

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

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

To complete the proof, we just need to check the claim of ((2.9)). By Condition (N2), $\forall \epsilon > 0,\ \exists\ n_\epsilon$ such that for $n \geq n_\epsilon$, $\sum_{j=a_n}^{b_n} c_0 \cdot e^{-\frac{B_{j,n}^2}{10(n-j)}} < \epsilon/4$. Then for $n > n_\epsilon$,

$$\begin{aligned}
\sum_{j=a_n}^{b_n} N_j \cdot e^{-\frac{\gamma^2 B_{j,n}^2}{4(n-j)}} & \leq \sum_{j=a_n}^{b_n} c_0 \cdot e^{\frac{B_{j,n}^2}{10(n-j)}} \cdot e^{-\frac{\gamma^2 B_{j,n}^2}{4(n-j)}} \leq \sum_{j=a_n}^{b_n} c_0 \cdot e^{-\frac{B_{j,n}^2}{10(n-j)}} < \epsilon/4 \\
\sum_{j=a_n}^{b_n} N_j \cdot L_j \cdot e^{-D_n/2} & = \sum_{j=a_n}^{b_n} N_j \cdot L_j \cdot e^{-2\zeta_n^{''} B_{j,n}} \leq \sum_{j=a_n}^{b_n} c_0 \cdot e^{-\zeta_n^{''} B_{j,n}} < \frac{\epsilon}{4}.
\end{aligned}$$

Similarly,

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

Thus claim ((2.9)) holds and this completes the proof. □

Proof 2.3 (Proof of Theorem 1 (parametric, σ unknown))

The proof of the case of unknown σ is almost the same as the one for the case where σ is known.

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

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

□

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

Proof 2.4 (Proof of Theorem 3 (nonparametric, $\boldsymbol{\sigma}$ unknown))

The proof of Theorem 3 is similar to that of Theorem 2. We only point out the major difference. Note that

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

Then

$$
\begin{aligned}
\frac{IC(\hat{k}_n^{(s)})}{IC(\hat{k}_n)} &= \frac{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{\hat{k}_n^{(s)}}\|^2 + \lambda_n \log(n) r_{\hat{k}_n^{(s)}}\hat{\sigma}_n^2 - n\hat{\sigma}_n^2 + dn^{\frac{1}{2}}\log(n)\hat{\sigma}^2}{\|\mathbf{Y}_n - \hat{\mathbf{Y}}_{\hat{k}_n}\|^2 + \lambda_n \log(n) r_{\hat{k}_n}\hat{\sigma}_n^2 - n\hat{\sigma}_n^2 + dn^{\frac{1}{2}}\log(n)\hat{\sigma}^2} \\
&= 1 + \frac{\|Pf_n\|^2 + \|Pe_n\|^2 + e_n^T P f_n - \lambda_n \log(n)\hat{\sigma}_n^2}{[(\lambda_n \log(n) - 1)r_{\hat{k}_n} + dn^{\frac{1}{2}}\log(n)]\hat{\sigma}_n^2},
\end{aligned}
$$

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

The rest of the proof follows similarly. □

Chapter 3

On consistency of model selection with model complexity penalty

3.1 Introduction

Identifying the most appropriate model among a list of candidates for an understanding of the nature of the data generating process is a fundamentally important issue in statistics. In this direction, model selection consistency has been widely studied. For classical model selection procedures (such as BIC), the consistency property was first derived in case of finite numbers of predictors (see, e.g., [59] for references). In recent years, to handle data with a large number of predictors (possibly much larger than the sample size) that result from modern technology of data collection, consistency results in the context of a diverging number of predictors have been established for various model selection methods such as adaptive LASSO and SCAD, see [18, 28, 30, 41, 54, 69, 70, 78, 83, 85].

In particular, [18, 69] proved that with suitable modifications, BIC-type of criteria continue to be consistent in the new setting with a large number of predictors. Chen and Chen [18] extended BIC by adding a penalty term with regard to the size of model space of each dimension and showed its consistency under the condition that the size of the true model does not grow in the sample size. Wang *et al* [69] modified BIC by

multiplying the BIC penalty term with a factor that goes to infinity with the sample size. They showed that the modified BIC is consistent for both unpenalized estimators and penalized estimators such as LASSO and SCAD under an assumption on the correlation among the predictors. These results are very interesting and encouraging as they enlarged the scope of applicability of the traditional BIC type criteria.

With the above background, our main goal in this work is to establish model selection consistency of a class of extended information criteria with a complexity penalty. The criteria, called generalized information criterion with model complexity (GICC), has the form $\frac{RSS_k}{\sigma^2} + m_k\eta_n + \lambda C_k$ when σ^2 is known and the form $n\log(\hat{\sigma}_k^2) + m_k\eta_n + \lambda C_k$ when σ^2 is unknown, where RSS_k is the residual sum of squares of model k based on least squares fit, m_k is the model dimension, η_n is a deterministic sequence approaching ∞, C_k is the complexity of model k in the list that satisfies the condition $\sum_k e^{-C_k} \leq 1$, λ is a positive constant and $\hat{\sigma}_k^2$ is the usual variance estimate from model k. From a Bayesian perspective, e^{-C_k} reflects the prior probability assignment on the candidate models. For more discussion, see [18, 71].

Clearly, without the model complexity term, the above criteria are the familiar GIC and $\eta_n \to \infty$ is needed for model selection consistency in the traditional setting of fixed truth and fixed set of models [59]. It has been well know that when exponentially many or mode models are present, GIC encounters severe selection bias and one effective way out is the addition of a complexity penalty that characterizes the complexity of the model list. Inspired by the pioneering work of Barron and Cover [6] that makes effective use of the complexity penalty term λC_k from an information theoretical point of view (C_k is a multiple of the length of code to describe the index of model k in the list), optimal risk bounds have been derived for criteria of the general form $-$loglikelihood$+m_k\eta_n + \lambda C_k$ or the like (typically with η_n bounded for achieving minimax-rate optimality in estimation) for density estimation, regression and related problems (see, e.g., [3, 5, 76], to name a few). It remains unknown, however, if this

approach leads to a satisfactory model selection consistency theory of the information criteria.

We intend to establish consistency of GICC under weak general conditions. We will show that the results in [18, 69] are special cases of ours. Specifically, we will allow the true model size to increase in the sample size and try to avoid restrictive assumptions on the correlations of the predictors. Besides giving sufficient conditions to guarantee model selection consistency, counter-examples are provided to show that these conditions are needed and cannot be generally removed. The general results will be applied in the contexts of all subset selection and also order selection. A risk bound for regression estimation will also be given for GICC.

A criticism of the concept of model selection consistency is that it is unrealistic to model the real world by a finite-dimensional parametric model. In an attempt to address this criticism, we slightly generalize the concept of model selection consistency to allow cases where none of the candidate model is the true one.

We need to point out that GICC is not computationally feasible for all subset selection with a large number of predictors. Nonetheless, the understanding of its consistency property is not only of theoretical interest and importance on its own, but also provides a benchmark when deriving theoretical properties for computationally fast model selection algorithms.

The rest of the paper is organized as follows. After introducing the setup of the problem and some notation in section 2, we then in section 3 give general sufficient conditions on selection consistency for $GICC$ for the cases when σ^2 is known. In this section, we discuss those conditions in details and provide simplified special cases in the context of all subset selection and order selection. We also compare our results to existing ones in the literature and give counterexamples to show that those general conditions are necessary in some sense. In section 4, we provide similar conditions for $GICC$ in the cases when σ^2 is unknown. In section 5, we generalize the concept

of consistency to the situations without the parametric assumption of a true model and extend the results in sections 3 and 4 to this new concept of consistency. In section 6, we derive a statistical risk bound for regression estimators of the selected model based on $GICC$. Conclusions are given in section 7 and proofs are presented in section 8.

3.2 Problem Setup

Consider model selection criteria in the context of linear regression problems:

$$Y_i = \mathbf{X}_i^T \beta + \epsilon_i, \quad i = 1, \cdots, n,$$

where $Y_i \in \mathbb{R}^1$ is the response variable, $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip_n})^T \in \mathbb{R}^{p_n}$ is the vector of predictors, and ϵ_i are assumed to be independent and normally distributed with mean zero and variance σ^2.

Consider linear models that have terms as functions of the original predictors. We denote a candidate model which consists of a subset of the p_n predictors and or functions/transformations of the predictors by model index k. For instance, we may consider adding interaction terms between certain original predictors, or require some variables to go together in models. Thus, clearly, our model selection problem cannot be reduced to that of all subset selection with an enlarged list of candidate terms.

Let Γ be the collection of indices of all the candidate models and let m_k be the model dimension, the number of terms in model k. Denote the projection matrix of the model k by M_k and its rank by r_k. Note that $m_k \geq r_k$. We point out that the candidate list Γ is arbitrary unless stated otherwise. Results for the important special case of all subset selection from the p_n predictors will be provided.

Let I_n be the identity matrix of size n and $\mu_n = E(\mathbf{Y}_n|\mathbf{X}) = (\mathbf{X}_1^T\beta, \cdots, \mathbf{X}_n^T\beta)^T$

be the mean vector, where $\mathbf{Y}_n = (Y_1, \cdots, Y_n)^T$ is the response vector and $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$.

3.2.1 Model selection criteria

Since our main interest in this work is to gain a general theoretical understanding on model selection consistency when dealing with a large number of linear models which is allowed to be flexible for potentially more applications, we resort to traditional information criteria that can be used for an arbitrary list of candidate linear models. Because there are exponentially many or more models in the sample size, one must deal with the complexity of the model list. We follow the idea of [6] to add a complexity penalty term to the information criteria (see, e.g., [71]).

In the case where σ^2 is known, we consider sufficient and necessary conditions on consistency for model selection criterion:

$$GICC_\lambda(k) = \frac{RSS_k}{\sigma^2} + m_k\eta_n + \lambda C_k$$

with $0 < \eta_n \to \infty$, where $RSS_k = \|\mathbf{Y}_n - \hat{\mathbf{Y}}_n\|^2 = \|(I_n - M_k)\mathbf{Y}_n\|^2$ is the residual sum of squares, C_k is the complexity penalty associated with model k, and $\lambda \geq 0$ is a constant.

In the case where σ^2 is unknown, we consider sufficient and necessary conditions on consistency for model selection criterion:

$$GICC_\lambda^{'}(k) = n\log(\hat{\sigma}_k^2) + m_k\eta_n + \lambda C_k$$

with $0 < \eta_n \to \infty$, where $\hat{\sigma}_k^2 = \frac{RSS_k}{n}$.

The complexity penalty terms C_k satisfy $C_k > 0$ and $\sum_{k\in\Gamma} e^{-C_k} \leq 1$.Thus λC_k is minus logarithm of the prior probability of model k. Alternatively, the penalty term

can be viewed as a multiple of number of bits needed to describe the model index k from a coding viewpoint. Typically, since Γ depends on n, so does C_k. In fact, many selection procedures can be reduced to an equivalent one that satisfies the above constraint. For instance, denote $\sum_{k\in\Gamma} e^{-C_k} = s_n$. Note that $\sum_{k\in\Gamma} e^{-(C_k+\log(s_n))} = 1$ and the model selection criterion $GICC_\lambda(k)$ is equivalent to $\frac{RSS_k}{\sigma^2} + m_k\eta_n + \lambda(C_k + \log(s_n))$ since the term $\log(s_n)$ is common to every candidate model k.

3.2.2 Notation

We assume that there exists a model in Γ, k_n^*, such that $\|(I_n - M_{k_n^*})\mu_n\|^2 = 0$ and that for any model k satisfying $\|(I_n - M_k)\mu_n\|^2 = 0$, we have $m_{k_n^*} < m_k$. The model k_n^* is called the true model at sample size n. In other words, we assume the existence and uniqueness of the true model. Additionally, we assume $m_{k_n^*} < n$.

We say model k is a sub-model of k', denoted $k \subset k'$, if all the terms in model k are also in model k'. Denote $\Gamma_{sup} = \{k : k \in \Gamma,\ k_n^* \subset k, \text{ and } k \neq k_n^*\}$, $\Gamma_w = \{k : k \in \Gamma,\ k_n^* \not\subset k\}$, and $\Gamma_{sub} = \{k : k \in \Gamma,\ k \subset k_n^*, \text{ and } k \neq k_n^*\}$. For $j \in \mathbb{Z}^+$, let $\Gamma_{sup}(j) = \{k : k \in \Gamma_{sup}, m_k = m_{k_n^*} + j\}$.

Let $\tilde{m}_k$ be the number of common terms between model k and the true model k_n^* and S_k be the linear space spanned by these common predictors. Denote the projection matrix onto the space S_k by $\tilde{M}_k$ and its rank by $\tilde{r}_k$. Still we have $\tilde{m}_k \geq \tilde{r}_k$.

Denote $e_n = (\epsilon_1, \cdots, \epsilon_n)^T$. Let $B_k \geq 0$ (to be chosen) be a quantity associated with model k satisfying the following constraint:

$$\sum_{k\in\Gamma_w} P\left(e_n^T(M_k - \tilde{M}_k)e_n/\sigma^2 > B_k\right) \to 0 \qquad \text{as } n \to \infty. \tag{3.1}$$

The term B_k is to bound the random error terms for models in Γ_w that include nuisance predictors. The constraint ((3.1)) says we need a bound B_k on these terms so that the probability that any random error term exceeds the bound goes to zero

as the sample size increases. With this bound, we then can determine the conditions for consistency, as will be seen in the next section.

Define $T_k = \|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) - B_k$. Denote $rem_1(k) = \mu_n^T(I_n - M_k)e_n$.

3.3 Consistency of GICC when σ^2 is known

In this section, we provide sufficient conditions for the model selection criterion $GICC$ and compare them to existing results in the literature. We also demonstrate that those conditions are necessary in some sense by giving some counterexamples. When σ^2 is known, the central issue of the influence of the complexity of the model list is more clearly seen without the additional technicality in estimating σ^2.

3.3.1 General conditions on consistency

Note that

$$\begin{aligned} GICC_\lambda(k) - GICC_\lambda(k_n^*) &= \left[\|(I_n - M_k)\mu_n\|^2 + 2\mu_n^T(I_n - M_k)e_n + e_n^T(M_{k_n^*} - M_k)e_n\right]/\sigma^2 \\ &\quad +(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}). \end{aligned} \tag{3.2}$$

In order to have consistency, we must have the overfitting and underfitting probabilities approach 0 as the sample size increases. When the number of models and the true model are fixed, we just need the true model to beat each of the candidate models with probability going to 1. But when the candidate models enlarge in n, that is no longer sufficient. We need some conditions to control the overall overfitting and underfitting probabilities, as is presented below.

Conditions:

(G1). For $k \in \Gamma_{sup}$, $\lambda(C_k - C_{k_n^*}) + (m_k - m_{k_n^*})\eta_n > (r_k - r_{k_n^*})\log(\eta_n)$.

(G2). There exist $0 < \alpha < 1$ and $0 \leq \zeta_n \to 0$ such that for all j and n,

$$\sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*}) \cdot \frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} \leq e^{\zeta_n \eta_n j}.$$

(G3). For $k \in \Gamma_w$, $T_k \geq c\|(I_n - M_k)\mu_n\|^2/\sigma^2$ for some constant $0 < c < 1$, and $\sum_{k\in\Gamma_w} e^{-cT_k/8} \to 0$ as $n \to \infty$.

Theroem 4

For choices of B_k satisfying the constraint ((3.1)), under Conditions (G1)-(G3), we have

$$P\left(\min_{\{k\in\Gamma, k\neq k_n^*\}} GICC_\lambda(k) > GICC_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{3.3}$$

□

Remarks:

1. Theorem 4 is a general result and it holds regardless what the candidate list Γ is.

2. It can be seen from the proof that the term $\log(\eta_n)$ in Condition $(G1)$ can be replaced by any positive sequence that goes to infinity with sample size n. Note that for $k \in \Gamma_{sup}$, $r_k - r_{k_n^*} \leq m_k - m_{k_n^*}$. Thus if $r_k - r_{k_n^*} \neq 0$, Condition (G1) can be written as: $\frac{\lambda(C_k - C_{k_n^*})}{r_k - r_{k_n^*}} + \eta_n \geq \log(\eta_n)$.

3. There is no requirement on how fast ζ_n approaches zero in Condition $(G2)$.

4. Condition $(G2)$ deals with Γ_{sup} and controls the overfitting probability. Basically, what this condition says is that the difference of penalty terms between an overfitting model and the true model should not be too small in order for the overall overfitting probability to go to zero.

5. Note that in Condition $(G2)$, the exponent on the left hand side is negative, while the one on the right hand side is positive. This reflects the fact that when p_n goes to infinity, even if the probability of selecting every single overfitting model goes to zero, the overall overfitting probability could still be large.

6. Condition $(G3)$ deals with Γ_w and controls the underfitting probability. Basically, what Condition $(G3)$ says is that for underfitting models, models in Γ_w, the approximation error $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ should not be too small relative to the difference of the penalty term and the bound of the error term B_k. More discussion about this condition can be found in the next subsection.

7. There are many choices of B_k satisfying the constraint ((3.1)). The more accurate the bound B_k is to the error term, the closer the Condition $(G3)$ is to the optimal one.

8. Intuitively, the complexity penalty C_k helps prevent overfitting since p_n is large. But for underfitting models with $m_k < m_{k_n^*}$, it is often the case that $\lambda(C_k - C_{k_n^*}) + \eta_n(m_k - m_{k_n^*}) - B_k < 0$ and thus the approximation error $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ has to be larger than the absolute value of $\lambda(C_k - C_{k_n^*}) + \eta_n(m_k - m_{k_n^*}) - B_k$ in order to have consistency. So although Conditions $(G1) - (G3)$ still guarantees selection consistency, the intrinsic requirement of Condition $(G3)$ on $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ becomes more stringent if the penalty term $(C_k - C_{k_n^*})$ is too small in the negative direction. This will be seen again in the remarks after Lemma 4.

9. There is no requirement on how large p_n could be, but like the role of the complexity penalty C_k, the order of p_n has an intrinsic impact on the term $\|(I_n - M_k)\mu_n\|^2/\sigma^2$, as will be seen from the remarks after Lemma 5.

3.3.2 On Constraint ((3.1)) and Conditions $(G1)-(G3)$

Having presented the main result on consistency, we now take a closer look at the constraint ((3.1)) and the Conditions $(G1)-(G3)$ to gain a better understanding. We provide some sufficient conditions for each of them to be satisfied. The discussions in this subsection apply to the settings where Γ is the collection of all subset models from the p_n predictors or a proportion of the collection.

Lemma 1

For a given constant $c > 2$, take $B_k = c\log(\nu_k)$, where $\nu_k = m_{k_n^*}^{(m_{k_n^*}-\tilde{m}_k)} \cdot p_n^{(m_k-\tilde{m}_k)}$, then the constraint ((3.1)) on B_k is satisfied. □

Remarks:

1. Recall that $\tilde{m}_k$ is the number of common predictors between model k and the true model k_n^*. Then $(m_{k_n^*} - \tilde{m}_k)$ is the number of predictors in the true model that model k misses and $(m_k - \tilde{m}_k)$ is the number of nuisance predictors in model k.

2. The bound B_k of the error term $e_n^T(M_k - \tilde{M}_k)e_n$ here is the same for models that have the same $(m_{k_n^*} - \tilde{m}_k)$ and $(m_k - \tilde{m}_k)$, or equivalently the same m_k and $\tilde{m}_k$.

3. Let $\nu_k^{'}$ be the number of models that have the same m_k and $\tilde{m}_k$. Note that ν_k is an upper bound for $\nu_k^{'}$ since $\nu_k^{'} < \begin{pmatrix} m_{k_n^*} \\ \tilde{m}_k \end{pmatrix} \cdot \begin{pmatrix} p_n \\ m_k - \tilde{m}_k \end{pmatrix}$.

4. There are other choices of B_k for constraint ((3.1)) to be met. For instance, we can also take $B_k = cm_k\log(p_n)$ for some $c > 2$, which is the same for models of the same dimension. Different choices of $B_k$will reflect different requirement on the true model. This will be seen in the remarks after Lemma 4 and Lemma 5.

Lemma 2

For the case $C_k = m_k \log(p_n)$, if $\lambda > 2$, then conditions $(G1)$ and $(G2)$ are satisfied.□

Remarks:

1. There is no restriction on p_n or η_n.

In a special case of p_n and η_n, we have the following.

Lemma 3

For the case $\eta_n = \log(n)$, $C_k = m_k \log(p_n)$, and $p_n = O(n^\kappa)$ for some $\kappa > 0$, if $\lambda > \max\{2(1 - \frac{1}{2\kappa}), 0\}$, then conditions $(G1)$ and $(G2)$ are satisfied. □

Remarks:

1. The condition $\lambda > \max\{2(1 - \frac{1}{2\kappa}), 0\}$ is exactly the same as the one in Chen and Chen [18], where $\lambda = 2\gamma$ and they require that $\gamma > 1 - \frac{1}{2\kappa}$ and $\gamma \geq 0$.

2. To incorporate the case $\kappa = 0$, which is the case of fixed p_n, we can replace $\lambda > \max\{2(1 - \frac{1}{2\kappa}), 0\}$ with $(\lambda \cdot \frac{1+\alpha}{2} - 2)\kappa + \alpha > 0$ and in this case Condition $(G2)$ is automatically satisfied.

Lemma 4

If there exist constants $0 < c < 1$ and a sequence $0 < \delta_n \to \infty$ such that for all $k \in \Gamma_w$,

$$\begin{aligned} & (1-c)\|(I_n - M_k)\mu_n\|^2/\sigma^2 + \eta_n(m_k - m_{k_n^*}) + \lambda(C_k - C_{k_n^*}) - B_k \\ \geq \; & \frac{8}{c}\left[\log\binom{m_{k_n^*}}{m_{k_n^*} - \tilde{m}_k} + \log\binom{p_n}{m_k - \tilde{m}_k} + \log(m_{k_n^*}) + \log(p_n) + \delta_n\right], \end{aligned}$$

then conditions $(G3)$ is satisfied. □

Remarks:

1. As is mentioned in the remarks after Theorem 4, for models in Γ_w with $m_k < m_{k_n^*}$, the difference of the complexity penalty term $C_k - C_{k_n^*}$, which is typically negative, the term $\eta_n(m_k - m_{k_n^*})$, and the bound of the error term B_k determine the requirement on the order of $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ which is related to the magnitude of the coefficients in the true model and the correlation structure among the predictors.

2. In [69], the complexity penalty $C_k = (D_n - 1)m_k \log(n)$ with $D_n \to \infty$. (In their notation, they used C_n. To avoid confusion with the complexity term C_k, we use D_n.) In this setting, when we choose $B_k = 3m_k \log(p_n)$, then the inequality in Lemma 4 implies that $\|(I_n - M_k)\mu_n\|^2$ needs to larger than $D_n(m_k - m_{k_n^*}) \log(n)$ and $3m_k \log(p_n)$. The condition 4 in [69] assumes that the smallest coefficient in the true model is of an order higher than $\sqrt{\frac{D_n p_n \log(n)}{n}}$ and the condition 2 in [69] assumes that the smallest eigenvalue of the covariance matrix of all the predictors is positive. These two conditions together are more or less saying that $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ is of an order higher than $D_n p_n \log(n)$, which makes the above inequality well satisfied. To us, these conditions can be improved with different choices of B_k and C_k. See more on this in the remarks after Lemma 5.

3. There is no requirement on the rate at which δ_n goes to infinity.

When we take B_k as the one in Lemma 1, then Lemma 4 becomes the following.

Lemma 5

If there exist constants $0 < c_1 < 1$, $c_2 > 2$, and a sequence $0 < \delta_n \to \infty$ such that for all $k \in \Gamma_w$,

$$\begin{aligned} & (1 - c_1)\|(I_n - M_k)\mu_n\|^2/\sigma^2 + \lambda(C_k - C_{k_n^*}) + (m_k - \tilde{m}_k)\left[\eta_n - (c_2 + \frac{8}{c_1})\log(p_n)\right] \\ \geq\ & (m_{k_n^*} - \tilde{m}_k)\left[\eta_n + (c_2 + \frac{8}{c_1})\log(m_{k_n^*})\right] + \frac{8}{c_1}\left[\log(m_{k_n^*}) + \log(p_n) + \delta_n\right], \end{aligned}$$

then conditions $(G3)$ is satisfied. In particular, with $C_k = m_k \log(p_n)$, the above inequality reduces to:

$$\begin{aligned} & (1-c_1)\|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - \tilde{m}_k)\left[\eta_n + (\lambda - c_2 - \frac{8}{c_1})\log(p_n)\right] \\ \geq \; & (m_{k_n^*} - \tilde{m}_k)\left[\eta_n + \lambda\log(p_n) + (c_2 + \frac{8}{c_1})\log(m_{k_n^*})\right] + \frac{8}{c_1}\left[\log(m_{k_n^*}) + \log(p_n) + \delta_n\right] \end{aligned}$$ □

Remarks: The implication of the above inequality is as follows. For simplicity, we present the implication of the inequality with $\eta_n = \log(n)$.

1. First note that for models in Γ_{sub}, $(m_k - \tilde{m}_k) = 0$ and the above inequality says that $\inf_{k\in\Gamma_{sub}} \frac{\|(I_n - M_k)\mu_n\|^2/\sigma^2}{(m_{k_n^*} - \tilde{m}_k)[\log(n) + \lambda\log(p_n) + (c_2 + \frac{8}{c_1})\log(m_{k_n^*})]}$ needs to be larger than a certain constant. Also note that in the case of orthogonal design matrix, for models in Γ_{sub}, $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ is roughly of an order $(m_k - \tilde{m}_k)n\beta_{min}^2$, where β_{min} is the coefficient in the true model that has the smallest absolute value. Then the above inequality is more or less saying that β_{min}^2 is of an order $\frac{\log(n)+\log(p_n)}{n}$ or higher, which is significantly weaker than the condition 4 in [69] where β_{min}^2 is of an order higher than $\frac{D_n p_n \log(n)}{n}$ for some $0 < D_n \to \infty$.

2. For models in $\Gamma_w - \Gamma_{sub}$, the approximation error $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ may not necessarily be large enough to dominate the term $(m_{k_n^*} - \tilde{m}_k)[\log(n) + \lambda\log(p_n) + (c_2 + \frac{8}{c_1})\log(m_{k_n^*})]$ due to the correlation among all the predictors unless we make strong assumptions on the magnitude of the coefficients in the true model and the correlation among all the predictors. Nonetheless, what we need is that $\inf_{\{k\in\Gamma_w\}} \frac{\|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - \tilde{m}_k)(\log(n) + (\lambda - c_2 - \frac{8}{c_1})\log(p_n))}{(m_{k_n^*} - \tilde{m}_k)[\log(n) + \lambda\log(p_n) + (c_2 + \frac{8}{c_1})\log(m_{k_n^*})]}$ is large enough, especially when λ is large and the term $\lambda - c_2 - \frac{8}{c_1}$ is positive.

3. Suppose $\inf_{\{k\in\Gamma_w\}} \frac{\|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - \tilde{m}_n)(\log(n) + (\lambda - c_2 - \frac{8}{c_1})\log(p_n))}{(m_{k_n^*} - \tilde{m}_k)[\log(n) + \lambda\log(p_n) + (c_2 + \frac{8}{c_1})\log(m_{k_n^*})]} := a_n$. Basically, the inequality requires that a_n to large than a certain constant. This is a generalization of the "asymptotic identifiability" condition in [18]. In fact, with $p_n =$

$O(n^\tau)$ and for those models in Γ_{sub}, we basically have $a_n = \inf_{\{k\in\Gamma_{sub}\}} \frac{\|(I_n-M_k)\mu_n\|^2/\sigma^2}{(m_{k_n^*}-\tilde{m}_k)\log(n)}$. When $m_{k_n^*}$ is bounded and $a_n \to \infty$, this becomes exactly the "asymptotic identifiability" condition.

4. When λ is large enough so that $\lambda - c_2 - \frac{8}{c_1} > 0$, then the above inequality is saying that the reduction in $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ by introducing an additional nuisance predictor into model k is less than or equal to $\frac{a_n}{1-c_1}[\log(n) + (\lambda - c_2 - \frac{8}{c_1})\log(p_n)]$.

5. When λ is not very large so that $\lambda - c_2 - \frac{8}{c_1} < 0$ but p_n is not too large so that $[\log(n) + (\lambda - c_2 - \frac{8}{c_1})\log(p_n)] > 0$, the inequality has the same implication as the one above.

6. When λ is not very large so that $\lambda - c_2 - \frac{8}{c_1} < 0$ and p_n is large so that $[\log(n) + (\lambda - c_2 - \frac{8}{c_1})\log(p_n)] < 0$, then it becomes hard for Condition $(G3)$ to be satisfied unless we make strong assumptions on $\|(I_n - M_k)\mu_n\|^2/\sigma^2$. But in this case, we can consider only those models up to a certain size, say s_n, and require that the above inequality holds for all $k \in \Gamma_w$ with $m_k \leq s_n$. Then we still have selection consistency among models with $m_k \leq s_n$ provided that $m_{k_n^*} < s_n$. Since $m_{k_n^*}$ is unknown, it must be estimated with prior information to get information on s_n. Chen and Chen [18] assumed that $m_{k_n^*}$ is bounded by a constant and obtained selection consistency among models with sizes up to a constant.

Corollary 3

For the case $\eta_n = \log(n)$ and $C_k = m_k \log(p_n)$, take $B_k = 3\log(\nu_k)$, where $\nu_k = m_{k_n^*}^{(m_{k_n^*}-\tilde{m}_k)} \cdot p_n^{(m_k-\tilde{m}_k)}$, if $\lambda > 2$, then we have consistency ((3.3)) under Condition $(G3)$. □

Remarks:

1. According to Lemma 1, instead of taking $B_k = 3\log(\nu_k)$, we can take $B_k = c\log(\nu_k)$ for any $c > 2$. With the setting and the choice of B_k in Corollary 3, we have

$$\begin{aligned} T_k &= \|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - m_{k_n^*})\log(n) + \lambda(m_k - m_{k_n^*})\log(p_n) \\ &\quad -3[(m_k - \tilde{m}_k)\log(p_n) + (m_{k_n^*} - \tilde{m}_k)\log(m_{k_n^*})] \\ &= \|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - \tilde{m}_k)\log(n) - (m_{k_n^*} - \tilde{m}_k)\log(n) + \lambda(m_k - \tilde{m}_k)\log(p_n) \\ &\quad -\lambda(m_{k_n^*} - \tilde{m}_k)\log(p_n) - 3[(m_k - \tilde{m}_k)\log(p_n) + (m_{k_n^*} - \tilde{m}_k)\log(m_{k_n^*})] \\ &\geq \|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - \tilde{m}_k)[\log(n) + (\lambda - 3)\log(p_n)] \\ &\quad -(m_{k_n^*} - \tilde{m}_k)[\log(n) + (\lambda + 3)\log(p_n)]. \end{aligned}$$

2. For the interpretation of Condition $(G3)$ with such a choice of B_k, see the remarks after Lemma 5.

3. The corollary eliminates Chen and Chen's [18] constraint that the size of the true model is bounded.

4. The corollary also relaxes Wang's et al [69] assumption on the order of the smallest coefficient in the true model. See the remark 1 after Lemma 5.

5. In our settings, there is no restriction on the order of p_n. In the case that $p_n = O(n^\kappa)$ for some $\kappa > 0$, which is the setting in [18], according to Lemma 3, we still have consistency for $\lambda > \max\{2(1 - \frac{1}{2\kappa}), 0\}$. This is the same as the condition in [18]. See the remarks after Lemma 3. Although there is no restriction on p_n, when p_n diverges extremely fast, say exponentially, then the Conditions $(G1) - (G3)$ become more difficult to be satisfied.

3.3.3 Consistency of GICC for subset selection

Having provided detailed discussions on the Conditions (G1)-(G3), we now give simplified results in the context all subset selection. The conditions in Theorem 5 below are special cases of the ones in previous lemmas.

In the all subset selection setting, suppose we have p_n predictors and the candidate list Γ consists of all the possible subsets of the p_n predictors.

Theroem 5

In the subset selection setting, choose $C_k = \lambda m_k \log(p_n)$, if $\eta_n \to \infty$ and there exists $0 < c < 1$ such that $(1-c)\|(I_n - M_k)\mu_n\|^2/\sigma^2 \geq m_{k_n^*}\left[\eta_n + \lambda \log(p_n)\right] + \frac{8}{c}\log(p_n)$ for $k \in \Gamma_w$ and $\lambda > 2 + \frac{8}{c}$, then

$$P\left(\min_{\{k\in\Gamma, k\neq k_n^*\}} GICC_\lambda(k) > GICC_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \qquad \square$$

Remarks:

1. In the all subset selection setting, the candidate list Γ is much more complicated than that of the order selection setting. Additional complexity penalty C_k is needed in order to prevent overfitting since p_n could go to infinity.

2. The conditions in Theorem 5 essentially say that for models in Γ_w, their approximation errors $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ need to larger than a multiple of $m_{k_n^*}\left[\eta_n + \lambda \log(p_n)\right]$ and λ is larger than a certain constant. These conditions are special cases of the Lemma 2 and Lemma 4.

3. Translating the conditions in Theorem 5 into the language of the order of coefficients in the true model, the conditions are more or less saying that the smallest coefficient in the true model is of an order $\sqrt{\frac{m_{k_n^*}[\eta_n + \lambda \log(p_n)]}{n}}$ or higher, in orthogonal design cases, for example. Note that this order, not optimal though,

is still better than the one in [69], $\sqrt{\frac{D_n p_n[\eta_n+\lambda\log(p_n)]}{n}}$ for some $D_n \to \infty$. See the remarks after Lemma 4 and Lemma 5.

4. There is no restriction on the rate at which p_n goes to infinity.

3.3.4 Comparison to existing results

In this subsection, we point out the connection between Theorem 4 and existing results in the literature on consistency of BIC-type model selection criterion. In particular, we compare the conditions in Theorem 4 to those in [18] and [69] , which are in the subset selection setting.

Chen and Chen [18] considered the case where the size of the true model is bounded and gave consistency results for models with sizes up to a fixed constant. In fact, their result is a special case of Theorem 4. Specifically, in their scenario, $\eta_n = \log(n)$, $p_n = O(n^\tau)$, $C_k = \log\begin{pmatrix} p_n \\ m_k \end{pmatrix}$, and only those models with $m_k \le K$ are considered, where K is a constant larger than the size of the true model. They assume that $\min_{\{k:\ m_k \le K\}} \frac{\|(I_n - M_k)\mu_n\|^2/\sigma^2}{\log(n)} \to \infty$, which is the "Asymptotic identifiability" condition in the paper. In this scenario, it is easy to check that there exist choices of B_k such that the constraint ((3.1)) and Conditions $(G1) - (G3)$ are all satisfied. First of all, according to the remark of Lemma 1, constraint ((3.1)) is met for $B_k = 3m_k\log(p_n)$. Second, according to Lemma 3, Conditions $(G1)$ and $(G2)$ are satisfied. Note that for $m_k \le K$, we have $B_k = 3m_k\log(p_n) = O(\log(n))$ and $C_k = \log\begin{pmatrix} p_n \\ m_k \end{pmatrix} = O(\log(n))$. Thus under their "Asymptotic identifiability" condition, we have that $T_k \ge a_n\log(n)$ for some $0 < a_n \to \infty$ and $\sum_{\{k\in\Gamma_w, m_k\le K\}} e^{-cT_k/8} < K \cdot p_n^K \cdot e^{-ca_n\log(n)/8} = K \cdot n^{\tau\cdot K} \cdot n^{-ca_n/8} \to 0$. Thus Condition $(G3)$ is satisfied. Therefore, according to Theorem 4, we have selection consistency.

Wang et al [69] modified BIC by multiplying the penalty term by a factor $D_n \to \infty$

(In their notation, they used C_n. To avoid confusion with the complexity term C_k, we use D_n.) and obtained consistency results under some conditions. In a sense, their conditions are also a special case of Theorem 4. For simplicity, we consider the case where the columns of the design matrix are orthogonal. We show that there exist choices of B_k such that the constraint ((3.1)) is met and Conditions $(G1) - (G3)$ are all satisfied under the assumptions in [69]. In their scenario, $\eta_n = \log(n)$, $p_n < n$, $C_k = m_k \log(n)$, and $\lambda = D_n - 1 \to \infty$. Still by Lemma 1, constraint ((3.1)) is met for $B_k = 3m_k \log(p_n)$. According to Lemma 2, Conditions $(G1) - (G2)$ are satisfied since $\lambda = D_n - 1 \to \infty$. In [69], they assume that the smallest coefficient in the true model is of an order higher than $\sqrt{\frac{p_n D_n \log(n)}{n}}$. Since the design matrix is orthogonal, then $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ is roughly of an order higher than $p_n D_n \log(n)$. Note that in their scenario, $T_k = \|(I_n - M_k)\mu_n\|^2/\sigma^2 + (m_k - m_{k_n^*})D_n \log(n) - B_k$. With $B_k = 3m_k \log(p_n)$, then T_k is of an order higher than $p_n D_n \log(n)$. Thus $\sum_{k\in\Gamma_w} e^{-cT_k/8} < 2^{p_n} \cdot e^{-D_n p_n \log(n)} = 2^{p_n} \cdot n^{-D_n p_n} \to 0$. That is, Condition $(G3)$ is satisfied. Again according to Theorem 4, we have selection consistency.

From the above discussion, we see that the Theorem 4 is a more general result than those in [18, 69].

We also examined the assumptions on the correlation among all the predictors in [16, 30, 42, 69, 78] and report in the following.

1. Fan and Peng [30] assumed regularity conditions on the likelihood function. In their notation, Condition (F) assumes that the eigenvalues of the Fisher information matrix are bounded away from zero and infinity, i.e.,

$$I_n(\beta_n) = -E_{\beta_n}\left\{\frac{\partial^2 \log(f_n(V_n;\beta_n))}{\partial\beta_{nj}\partial\beta_{nk}}\right\}$$

satisfies that

$$0 < C_1 < \lambda_{min}\{I_n(\beta_n)\} \leq \lambda_{max}\{I_n(\beta_n)\} < C_2 < \infty \quad \text{for all } n.$$

In the normal regression setting, it is easy to check that $\frac{\partial^2 \log(f_n(V_n;\beta_n))}{\partial\beta_{nj}\partial\beta_{nk}}$ is basically $n^{-1}X^TX$, where X is the design matrix.

2. Candes and Tao [16] assumed the uniform uncertainty principle, which states that the design matrix X obeys the "restricted isometry hypothesis". Specifically, let X_T, $T \subset \{1, ..., p\}$ be the $nx|T|$ submatrix obtained by extracting the columns of X corresponding to the indices in T; then Candes and Tao (2005) define the S-restricted isometry constant δ_S of X which is the smallest quantity such that

$$(1 - \delta_S)\|c\|_{l_2}^2 \leq \|X_T c\|_{l_2}^2 \leq (1 + \delta_S)\|c\|_{l_2}^2$$

for all subsets T with $|T| \leq S$ and coefficient sequences $(c_j)_{j\in T}$. This property essentially requires that every set of columns with cardinality less than S approximately behaves like an orthonormal system.

An implication of the "restricted isometry hypothesis" is that the eigenvalues of X_T are within the range $[1 - \delta_S, 1 + \delta_S]$, which is bounded away from zero and infinity. In their discussions of the isometry condition, they mentioned that the condition holds for $S \approx n/\log(p)$.

3. Huang et al [42] imposed conditions on the correlation among predictors. In their notation, Condition (A5) assumes that the eigenvalues of the Σ_{n11} are bounded away from zero, where $\Sigma_{n11} = n^{-1}\mathbf{X}_1^T\mathbf{X}_1$ and $\mathbf{X}_1$ is the design matrix restricted to the columns corresponding to the predictors in the true model. Condition (B2) (called partial orthogonality in the paper) assumes that the the covariates with zero coefficients and those with nonzero coefficients are weakly

correlated.

4. Wang et al [69] assumed that the covariance matrix of all the predictors are always bounded away from zero.

5. Zhang [78] assumed the sparse Riesz condition (SRC) on the design matrix X: for suitable $0 < c_* \leq c^* < \infty$ and rank d^*,

$$c_* \leq \min_{|A| \leq d^*} c_{min}(\Sigma_A) \leq \max_{|A| \leq d^*} c_{max}(\Sigma_A) \leq c^*,$$

where $A \subset \{1, 2, \cdots, p\}$, $X_A = (\mathrm{x}_j, j \in A)_{n \times |A|}$, $\Sigma_A = X_A^T X_A / n$, and $c_{min}(\Sigma_A)$ and $c_{max}(\Sigma_A)$ are the smallest and largest eigenvalues of Σ_A, respectively.

So the assumptions in all the above papers are essentially the same, i.e., the design matrix X is basically close to orthogonal. In such cases, the requirement of Condition (G3) in our paper reduces to that the smallest coefficient in the true model is of an order $\sqrt{\frac{\log(n)+\log(p_n)}{n}}$ or higher, as is discussed in the remarks after Lemma 5 and Theorem 5. Note that in [30, 42, 69, 78], they all explicitly assumed that the smallest coefficient needs to be larger than a certain rate. In [16], when evaluating the oracle inequality, they more or less assumed that the smallest coefficient is larger than a certain order.

3.3.5 Consistency of GICC for selection among a sequence of models

In this subsection, we first provide simplified sufficient conditions for consistency of GICC in the context of order selection and then give a more applicable result. These results are obtained from the earlier tools we have already presented.

Order selection

For order selection, suppose we have an ordered list of predictors $x_1, \cdots, x_{p_n}$ and the true model k_n^* consists of predictors $x_1, \cdots, x_{m_{k_n^*}}$. Consider the candidate list Γ to be the collection of all the models following the ordered sequence of the predictors. In other words, in this setting, there is only one model in Γ of each dimension.

Theroem 6

In the above order selection setting, choose $\lambda = 0$, if $\eta_n \to \infty$ and there exist $0 < c < 1$ and $\delta_n \to \infty$ such that

$$\frac{c}{8}\left[(1-c)\|(I_n - M_k)\mu_n\|^2/\sigma^2 - (m_{k_n^*} - m_k)\eta_n\right] \geq \log(m_{k_n^*}) + \delta_n$$

for $k \in \Gamma_w$, then

$$P\left(\min_{\{k\in\Gamma, k\neq k_n^*\}} GICC_\lambda(k) > GICC_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \qquad \square$$

Remarks:

1. In above the order selection setting, the candidate list Γ is simple and consists of one model of each dimension. Thus we do not need additional complexity penalty C_k to prevent overfitting. As long as $\eta_n \to \infty$, the probability of overfitting would go to zero as the sample size goes to infinity.

2. Note that models in Γ_w are all sub-models of the true model k_n^*. The condition in Theorem 5 says that the approximation error $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ of a underfitting model just needs to be larger than the number of predictors it misses multiplied by η_n plus a multiple of $\log(m_{k_n^*}) + \delta_n$.

3. δ_n can diverge to infinity at any rate.

Selection along a solution path

As mentioned earlier, all subset selection by traditional information criteria is computationally infeasible when p_n is large. Fast computing path algorithms are now available. It has been suggested to use BIC and the like to consistently select a model along the path (see, e.g., [69, 68]). We give such a result here, assuming the solution path of a model selection method contains the true model with probability tending to 1.

Suppose we have two data sets from the same regression setting. Let $\Gamma(\Delta)$ be the collection of the models on the solution path of a path-algorithm based model selection method Δ with the first data set. We then apply GIC (i.e., GICC with $\lambda = 0$) with $\eta_n \to \infty$ to choose a model on the path with the second one. Note that we could split a data set into two parts, with the first one used to find the solution path and second one used to select the best model on the path.

Corollary 4

Assume there exist $0 < c < 1$ and $\delta_n \to \infty$ such that

$$\|(I_n - M_k)\mu_n\|^2/\sigma^2 - (m_{k_n^*} - m_k)\eta_n - 2(r_k - \tilde{r}_k) \geq \frac{8}{c}\left[\log(m_{k_n^*}) + \delta_n\right] + \frac{2}{1 - \log(2)}\log(p_n)$$

for $k \in \Gamma_w$, where $\Gamma_w = \{k : k \in \Gamma,\ k_n^* \not\subset k\}$. Then if $P\left(k_n^* \in \Gamma(\Delta)\right) \to 1$, we have

$$P\left(\min_{\{k \in \Gamma(\Delta),\, k \neq k_n^*\}} GIC(k) > GIC(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \qquad \square$$

Remarks:

1. The shrinkage type of model selection methods such as SCAD and LASSO has been widely studied in the literature, see [16, 28, 30, 41, 42, 54, 69, 78, 83, 84, 85].

It has been shown that they are consistent under some regularity conditions. In other words, their solution paths contain the true model with probability going to 1.

2. Wang et al [68] proved that BIC tuning parameter selector with the SCAD can identify the true model consistently. Wang et al [69] showed that modified BIC is consistent for both penalized and unpenalized estimators with a diverging number of predictors.

3. The condition in Corollary 4 on $\|(I_n - M_k)\mu_n\|^2/\sigma^2$ for $k \in \Gamma_w$ is a little different than that in Theorem 6 because the models on the solution paths may not necessarily be nested.

4. For model $k \in \Gamma_w$, the term $(r_k - \tilde{r}_k)$ is the number of nuisance predictors in model k that provide no contribution at all in approximating the true model. In nested situations, $(r_k - \tilde{r}_k) = 0$.

5. Compared to the condition in Theorem 6, there is an additional term $\log(p_n)$ in the condition of Corollary 4. This is still due to the fact that a model $k \in \Gamma_w$ on the solution path may not be a sub-model of k_n^* and as a result it could be the case that $m_k \geq m_{k_n^*}$.

3.3.6 Counterexamples

In this sub-section, we give some examples to demonstrate that the above conditions are necessary in some sense. Without loss of generality, we assume $\sigma^2 = 1$.

Necessity of Condition $(G2)$

As has been pointed out in many papers such as [18, 69], when the number of predictors is diverging, traditional model selection criteria of BIC-type need to be modified

in order to maintain the consistency property. In the following, we take BIC as an example. In the case p_n goes to infinity, as will be seen in the following example, since the number of overfitting models increases dramatically with the sample size, an additional penalty is needed to prevent overfitting.

Denote $A_k = \{GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\}$ and recall that $\Gamma_{sup}(1) = \{k : k \in \Gamma_{sup} \text{ and } m_k = m_{k_n^*} + 1\}$. In order for the criterion to be consistent, we must have

$$P\left(\cup_{k\in\Gamma_{sup}(1)} A_k\right) \to 0 \quad \text{as } n \to \infty. \tag{3.7}$$

We now give an example in which Condition $(G2)$ is violated and ((3.7)) is not satisfied.

According to equation (3.2), for model $k \in \Gamma_{sup}(1)$, we have

$$\begin{aligned} GICC_\lambda(k) - GICC_\lambda(k_n^*) &= -e_n^T(M_k - M_{k_n^*})e_n + \eta_n + \lambda(C_k - C_{k_n^*}) \\ &= -\chi_1^2 + \eta_n + \lambda(C_k - C_{k_n^*}). \end{aligned}$$

Suppose we have a linear regression problem with the number of predictors and the size of the true model satisfying $p_n - m_{k_n^*} = \left\lceil \sqrt{n\log(n)} \right\rceil$. For simplicity, we assume that the predictors are orthogonal to each other. In this case, the events A_k are independent for all $k \in \Gamma_{sup}(1)$. Let $a_k = \eta_n + \lambda(C_k - C_{k_n^*})$. Note that

$$\frac{2}{\sqrt{2\pi}} \cdot \frac{\sqrt{a_k}}{1+a_k} \cdot e^{-\frac{a_k}{2}} \leq P(A_k) = P(\chi_1^2 \geq a_k) \leq \frac{2}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{a_k}} \cdot e^{-\frac{a_k}{2}}$$

According to Bonferroni's inequality,

$$\begin{aligned} P\left(\cup_{k\in\Gamma_{sup}(1)}A_k\right) &\geq \sum_{k\in\Gamma_{sup}(1)} P(A_k) - \sum_{k1<k2\ \in\Gamma_{sup}(1)} P(A_{k1}\cap A_{k2}) \\ &= \sum_{k\in\Gamma_{sup}(1)} P(A_k) - \sum_{k1<k2\ \in\Gamma_{sup}(1)} P(A_{k1})\cdot P(A_{k2}) \end{aligned}$$

For traditional BIC, $a_k = \log(n)$. Then

$$\begin{aligned} P\left(\cup_{k\in\Gamma_{sup}(1)}A_k\right) &\geq \sum_{k\in\Gamma_{sup}(1)} P(A_k) - \sum_{k1<k2\ \in\Gamma_{sup}(1)} P(A_{k1})\cdot P(A_{k2}) \\ &= (p_n - m_{k_n^*})\cdot\frac{2}{\sqrt{2\pi}}\cdot\frac{\sqrt{\log(n)}}{1+\log(n)}\cdot e^{-\frac{\log(n)}{2}} \\ &\quad -\frac{(p_n - m_{k_n^*})(p_n - m_{k_n^*}-1)}{2}\cdot\frac{4}{2\pi}\cdot\frac{1}{\log(n)}\cdot e^{-\log(n)} \\ &= \frac{2(p_n - m_{k_n^*})}{\sqrt{2\pi\log(n)}}\cdot e^{-\frac{\log(n)}{2}}\left[\frac{\log(n)}{1+\log(n)} - \frac{p_n - m_{k_n^*}-1}{\sqrt{2\pi\log(n)}}\cdot e^{-\log(n)/2}\right] \\ &= \frac{2(p_n - m_{k_n^*})}{\sqrt{2\pi n\log(n)}}\left[\frac{\log(n)}{1+\log(n)} - \frac{p_n - m_{k_n^*}-1}{\sqrt{2\pi n\log(n)}}\right] \end{aligned}$$

Since $p_n - m_{k_n^*} = \left\lceil\sqrt{n\log(n)}\right\rceil$, then $P\left(\cup_{k\in\Gamma_{sup}(1)}A_k\right) \geq \frac{2}{\sqrt{2\pi}}\left(1-\frac{1}{\sqrt{2\pi}}\right) > 0$ as $n\to\infty$.

Now let us check Condition $(G2)$. Clearly, for traditional BIC, $\lambda = 0$ and for any $0\leq\alpha<1$,

$$\begin{aligned} \sum_{k\in\Gamma_{sup}(1)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*})\cdot\frac{1+\alpha}{2} + \alpha\log(n)(m_k - m_{k_n^*})\right]} &= (p_n - m_{k_n^*})\cdot e^{-\alpha\log(n)/2} \\ &\geq (p_n - m_{k_n^*})\cdot e^{-\log(n)/2} \\ &= (p_n - m_{k_n^*})\cdot\frac{1}{\sqrt{n}} \\ &= O\left(\sqrt{\log(n)}\right) \to\infty. \end{aligned}$$

In sum, we have shown in this example that Condition $(G2)$ is violated and the criterion (BIC) is not consistent.

Necessity of Condition $(G3)$

An immediate counterexample of Condition $(G3)$ is that the additional penalty term C_k is too heavy that the term $\tilde{T}_k := \|(I_n - M_k)\mu\|^2 - \eta_n(m_k - m_{k_n^*}) + \lambda(C_k - C_{k_n^*})$ becomes negative for some underfitting models. Let $\Gamma_{sub}(1) := \{k : k \in \Gamma_{sub} \text{ and } m_k = m_{k_n^*} - 1\}$. For model $k \in \Gamma_{sub}(1)$, the term $\tilde{T}_k = \|(I_n - M_k)\mu\|^2 - \eta_n + \lambda(C_k - C_{k_n^*})$. According to equation ((3.2)),

$$\begin{aligned} GICC_\lambda(k) - GICC_\lambda(k_n^*) &= 2rem_1(k) + e_n^T(M_{k_n^*} - M_k)e_n + \tilde{T}_k \\ &= 2rem_1(k) + \chi_1^2 + \tilde{T}_k \end{aligned}$$

We claim that if there exists a model $k \in \Gamma_{sub}(1)$ with $\tilde{T}_k \le 0$, then $P(GICC_\lambda(k) - GICC_\lambda(k_n^*) \le 0) \not\to 0$. That is, the probability of selecting an underfitting model is not vanishing as $n \to \infty$, as is proved below.

We know for any $0 < \epsilon < 1/2$, there exists a constant $a_\epsilon > 0$ such that $P(\chi_1^2 \ge a_\epsilon) \le \epsilon$. Then

$$\begin{aligned} P(A_k) &= P(GICC_\lambda(k) - GICC_\lambda(k_n^*) \le 0) \\ &\ge P\left(\{2rem_1(k) + \chi_1^2 + \tilde{T}_k \le 0\} \cap \{\chi_1^2 < a_\epsilon\}\right) \\ &\ge P\left(\{2rem_1(k) + a_\epsilon + \tilde{T}_k \le 0\} \cap \{\chi_1^2 < a_\epsilon\}\right) \\ &\ge P\left(\{2rem_1(k) + a_\epsilon + \tilde{T}_k \le 0\}\right) - P\left(\{\chi_1^2 \ge a_\epsilon\}\right) \\ &\ge P\left(2rem_1(k) \le -(a_\epsilon + \tilde{T}_k)\right) - \epsilon \end{aligned} \tag{3.8}$$

Note that $rem_1(k) \sim N(0, \|(I_n - M_k)\mu\|^2)$. Also note that model $k \in \Gamma_{sub}(1)$ and $\tilde{T}_k < 0$.

- If $a_\epsilon + \tilde{T}_k \leq 0$, then $P\left(2rem_1(k) \leq -(a_\epsilon + \tilde{T}_k)\right) \geq \frac{1}{2}$.
- If $a_\epsilon + \tilde{T}_k > 0$, then $P\left(2rem_1(k) \leq -(a_\epsilon + \tilde{T}_k)\right) = P\left(N(0,1) \leq -\frac{a_\epsilon + \tilde{T}_k}{2\|(I_n - M_k)\mu\|}\right) \geq P\left(N(0,1) \leq -\frac{a_\epsilon}{2\|(I_n - M_k)\mu\|}\right) \to \frac{1}{2}$ since $\|(I_n - M_k)\mu\| \to \infty$ as $n \to \infty$.

For either case, there exists a constant $\epsilon' > 0$ such that $P(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0) \geq \epsilon'$ as $n \to \infty$. In other words, the criterion $GICC$ is not consistent in this case.

Obviously, if there exists a model $k \in \Gamma_{sub}(1)$ with $\tilde{T}_k \leq 0$, then Condition $(G3)$ is violated since $T_k \leq \tilde{T}_k \leq 0$ and $\sum_{k \in \Gamma_{sub}} e^{-cT_k/8} \geq 1$.

Next we give a counterexample of Condition $(G3)$ with a model $k \in \Gamma_w$ and $k \not\subset k_n^*$. Suppose the true model consists of predictors $x_1, x_2, \cdots, x_6$ which are orthogonal to one another. A candidate model k consists of predictors x_1, x_2, x_3, x_4, x_7 with x_7 orthogonal to $x_1, x_2, \cdots, x_6$. Furthermore, suppose $\tilde{T}_k = \|(I_n - M_k)\mu_n\|^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \leq 0$. Clearly, Condition $(G3)$ is violated in this case. We show that $P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \not\to 0$.

Note that in this case,

$$\begin{aligned} GICC_\lambda(k) - GICC_\lambda(k_n^*) &= \|(I_n - M_k)\mu_n\|^2 + 2rem_1(k) + e_n^T(M_{k_n^*} - \tilde{M}_k)e_n \\ &\quad -e_n^T(M_k - \tilde{M}_k)e_n + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\ &= \tilde{T}_k + 2rem_1(k) + \chi_2^2 - \chi_1^2 \end{aligned}$$

Similar to ((3.8)) and the above argument , we have that $P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \not\to 0$.

3.4 Consistency of GICC when σ^2 is unknown

For the case of unknown σ^2, we consider the model selection criterion $GICC'$ with $GICC'_\lambda(k) = n\log(\hat{\sigma}_k^2) + m_k\eta_n + \lambda C_k$, where $\hat{\sigma}_k^2 = RSS_k/n$.

For $w > -1$, let $T_k(w) = \|(I_n - M_k)\mu_n\|^2/\sigma^2 + \frac{1}{1+w}\left[(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})\right] - B_k$.

3.4.1 Results for the case with $p_n = o(n)$

Conditions:

(G4). Assume $p_n = o(n)$ and there exists a partition of $\Gamma = \Gamma_s \cup \Gamma_l$ such that $\limsup_{n\to\infty} \frac{\sup_{k\in\Gamma_s} \|(I_n - M_k)\mu_n\|^2/\sigma^2}{n} = 0$, $\liminf_{n\to\infty} \frac{\inf_{k\in\Gamma_l} \|(I_n - M_k)\mu_n\|^2/\sigma^2}{n} > 0$, and if $\inf_{k\in\Gamma_l}(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) < 0$, then $\liminf_{n\to\infty} \frac{\inf_{k\in\Gamma_l}(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})}{n} = 0$.

(G5). There exist constants $0 < \epsilon < 1, 0 < c < 1$ such that for $k \in \Gamma_w$, $\min\left(T_k(\epsilon),\ T_k(-\epsilon)\right) \geq c\|(I_n - M_k)\mu_n\|^2/\sigma^2$, and $\sum_{k\in(\Gamma_w\cap\Gamma_s)}\left[e^{-cT_k(\epsilon)/8} + e^{-cT_k(-\epsilon)/8}\right] \to 0$.

Theroem 7

For choices of B_k satisfying the constraint ((3.1)), under Conditions (G1)-(G2) and (G4)-(G5), we have

$$P\left(\min_{\{k\in\Gamma, k\neq k_n^*\}} GICC'_\lambda(k) > GICC'_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{3.9}$$

□

Remarks:

1. When $\epsilon \to 0$, Condition $(G5)$ becomes exactly the same as Condition $(G3)$. The difference is due to the uncertainty of $\hat{\sigma}^2$ in estimating σ^2. Condition $(G5)$ has similar interpretations as Condition $(G3)$. See the remarks after Lemma 4 and Lemma 5.

2. The first two equations in Condition $(G4)$ say that for models in Γ_s their approximation errors are uniformly of a smaller order than n and for models in Γ_l their approximation errors are uniformly of a order n or higher.

3. Note that if p_n is fixed, then for any candidate model either the first equation or the second one in Condition (G4) holds. That is, when p_n is fixed, the partition of Γ always exists. This is not necessarily the case for $p_n \to \infty$.

4. As can be seen from the proof, the first equation in Condition $(G4)$ guarantees the uniformity in Taylor expansion for models in Γ_s. It makes sure that the probability that the term $\mu_n^T(I_n - M_k)e_n$ is large for some models in Γ_s goes to zero as the sample size increases.

5. The second equation in Condition $(G4)$ guarantees that the probability $\hat{\sigma}_k^2 < \hat{\sigma}_{k_n^*}^2$ for some models in Γ_l goes to zero as the sample size n goes to infinity. In other words, it guarantees that with probability going to 1, we have $\hat{\sigma}_k^2 > \hat{\sigma}_{k_n^*}^2$ uniformly over all the models in Γ_l, which is needed because otherwise, there will be a certain probability $GICC_\lambda'(k) < GICC_\lambda'(k_n^*)$ for some models in Γ_l, especially for those with $m_k < m_{k_n^*}$ and having a smaller penalty term than the true model.

6. The third equation in Condition $(G4)$ says that the difference of the penalty terms between models in Γ_l and the true model should not be too large in the negative direction. Intuitively, if a candidate model has a penalty term that is much smaller than that of the true model, then its approximation error has to be really large in order to have selection consistency. This point has also been discussed in the remarks after Lemma 4.

7. Note that the true model $k_n^* \in \Gamma_s$ since $\|(I_n - M_{k_n^*})\mu_n\|^2 = 0$. Similarly, $\Gamma_{sup} \subset \Gamma_s$.

Corollary 5

For the case $\eta_n = \log(n)$ and $C_k = m_k \log(p_n)$, take $B_k = 3\log(\nu_k)$, where $\nu_k = m_{k_n^*}^{(m_{k_n^*} - \tilde{m}_k)} \cdot p_n^{(m_k - \tilde{m}_k)}$, if $\lambda > 2$, then we have consistency ((3.15)) under Conditions

$(G4) - (G5)$. □

Remarks:

1. Similar to Corollary 3, the choice of B_k here has similar implication on the requirement of the true model. See the remarks after Corollary 3.

3.4.2 Results for arbitrary p_n with additional information on σ^2

Note that Theorem 7 handles the case $p_n = o(n)$. To handle the case with arbitrarily large p_n, we need a little technical condition.

Conditions:

(G6). Suppose we have a reference estimate of σ^2, $\hat{\sigma}^2_{ref}$, such that $P\left(a < \frac{\hat{\sigma}^2_{k_n^*}}{\hat{\sigma}^2_{ref}} < b\right) \to 1$ as $n \to \infty$ for some $0 < a < b < \infty$, where $\hat{\sigma}^2_{k_n^*} = RSS_{k_n^*}/n$.

(G7). Denote $\hat{\Gamma} = \{k : k \in \Gamma,\ a < \frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{ref}} < b\}$, $c_1 = \frac{\log(b/a)}{b/a-1}$, $c_2 = \frac{\log(b/a)}{1-a/b}$. Note that $0 < c_1 < 1$, $c_2 > 1$. Suppose there exist constants $0 < c < 1$, $0 < w_1 < c_1$, and $w_2 > c_2$ such that for $k \in \Gamma_w$, $\min\left(T_k(w_1 - 1),\ T_k(w_2 - 1)\right) \geq c\|(I_n - M_k)\mu_n\|^2/\sigma^2$, and $\sum_{k\in\Gamma_w}\left[e^{-cT_k(w_1-1)/8} + e^{-cT_k(w_2-1)/8}\right] \to 0$.

(G8). There exist $0 < \alpha < 1$ and $0 \leq \zeta_n \to 0$ such that for all j and n,

$$\sum_{k\in\Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*})\cdot\frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]\cdot\frac{1}{w_2}} \leq e^{\zeta_n\eta_n j}.$$

Theroem 8

For choices of B_k satisfying the constraint ((3.1)), under Conditions (G1) and (G6)-(G8), we have $P(k_n^* \in \hat{\Gamma}) \to 1$ and

$$P\left(\min_{\{k\in\Gamma\cap\hat{\Gamma},\ k\neq k_n^*\}} GICC'_\lambda(k) > GICC'_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{3.10}$$

□

Remarks:

1. Theorem 8 says if we have a reference estimate of σ^2, $\hat{\sigma}^2_{ref}$, such that the ratio $\frac{\hat{\sigma}^2_{k^*_n}}{\hat{\sigma}^2_{ref}}$ is bounded away from zero and infinity with probability going to 1, then we can consider only those models whose ratio $\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{ref}}$ is within a certain range and among these models $GICC$ identifies the true model with probability going to 1.

2. We are ignoring those models whose ratio $\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{ref}}$ tends to either zero or infinity for two reasons. The first reason is for technical convenience. With the ratio bounded away from zero and infinity, we can find lower bounds for the the term $\log(\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{ref}})$. The second reason is from practical concern. We can choose a range (a, b) large enough to include all interesting models.

3. When p_n is larger than n, it is possible that some overfitting models have ranks close to n and thus have $\hat{\sigma}^2_k \approx 0$. Then the term $n \log(\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{k^*_n}})$ tends to $-\infty$ nastily fast.

4. For those models with $\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{ref}}$ tends infinity, the term $n \log(\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{k^*_n}})$ also tends to ∞. But when p_n is huge, it is still possible that the difference of the penalty terms goes to $-\infty$ even faster since $\lim_{x\to\infty} \frac{x}{\log(x)} = \infty$.

5. In a sense, Theorem 7 is a special case of Theorem 8. It can be seen from the proof that under the assumptions of Theorem 7, since $p_n = o(n)$, for models in Γ_s, the ratio $\frac{\hat{\sigma}^2_k}{\hat{\sigma}^2_{k^*_n}}$ is bounded away from zero and infinity, and that for models in Γ_l, the ratio is bounded away from zero and if the ratio goes to infinity for models in Γ_l, then the difference of the penalty terms goes to $-\infty$ at a slower rate than n.

6. Condition (G6) is met if the estimate $\hat{\sigma}^2_{ref}$ is independent of $\hat{\sigma}^2_{k^*_n}$ and the ratio

$\frac{\sigma^2}{\hat{\sigma}^2_{ref}}$ is bounded away from zero and infinity with probability going to 1. In general, $\hat{\sigma}^2_{ref}$ does not have to be independent of $\hat{\sigma}^2_{k_n^*}$.

7. Instead of using $\hat{\sigma}^2_{ref}$, an alternative to Condition (G6) is that we have an upper bound and a lower bound on σ^2 and then choose $\hat{\Gamma}$ to be the set of model indices with $\hat{\sigma}^2_k$ within a certain range.

8. Condition (G7) is essentially the same as Condition (G3) or Condition (G5). We are just multiplying the difference of the penalty terms by a constant $\frac{1}{w_1}$ or $\frac{1}{w_2}$ due to the estimation uncertainty of σ^2. Notice that $0 < w_1 < 1$ and $w_2 > 1$.

9. Similarly, Condition (G8) is almost the same as Condition (G2) except that we are multiplying the difference of the penalty terms by a constant $\frac{1}{w_2}$.

3.5 Consistency of GICC without parametric assumptions

Having provided the conditions for consistency of the selection criteria $GICC$ and $GICC'$ in the previous sections, we now generalize the concept of consistency.

A criticism of theories on model selection consistency is that the true model cannot be finite dimensional and then the concept of selecting the true (finite-dimensional) model makes no sense. This critical view certainly has a point that it is often better to treat all the candidate models as approximations to the infinite-dimensional truth (which itself is just a mathematical simplification of the reality). For a discussion on practical parametricness/nonparametricness, see [51]. Below we slightly extend the concept of consistency to mean the ability to select the model with a best trade-off between approximation and model dimension. Note that consistency of cross validation in selecting the best regression estimator (in terms of a global mean squared

error) has been established in [74].

Suppose k_n^* stands for the model that minimizes $\|(I_n - M_k)\mu_n\|^2 + r_k\eta_n^{'}\sigma^2$, where $\eta_n^{'} \geq 0$ may not be the same as η_n in the selection criterion $GICC$ and $\eta_n^{'}$ could be constant. We define consistency to be that the probability of choosing model k_n^* goes to 1 as the sample size goes to infinity. We show that the conditions in previous sections naturally extend to this new concept with minor revision. We still focus on selection criteria $GICC$ and $GICC^{'}$.

Note that if $\eta_n^{'} = 1$, then k_n^* stands for the model with the best tradeoff between the approximation error and the estimation error since $E\|\hat{\mathbf{Y}}_n - \mu_n\|^2 = \|(I_n - M_k)\mu_n\|^2 + r_k\sigma^2$.

3.5.1 σ^2 known

general conditions

In parallel to Conditions (G2) and (G3), we first provide general conditions on consistency for this new concept.

Note that

$$\begin{aligned} GICC_\lambda(k) - GICC_\lambda(k_n^*) &= \left(\|(I_n - M_k)\mathbf{Y}_n\|^2 - \|(I_n - M_{k_n^*})\mathbf{Y}_n\|^2\right)/\sigma^2 \\ &\quad +(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}). \end{aligned}$$

Denote $T_k^{'} = \mu_n^T(M_{k_n^*} - M_k)\mu_n/\sigma^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) - B_k$.

Conditions:

$(G2^{'})$. There exist $0 \leq \alpha < 1$ and $0 \leq \zeta_n \to 0$ such that

$$\limsup_{n\to\infty} \sup_{k\in\Gamma_{sup}} \frac{2\|(M_k - M_{k_n^*})\mu_n\|^2/\sigma^2 + (r_k - r_{k_n^*})}{\|(M_k - M_{k_n^*})\mu_n\|^2/\sigma^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})} \leq 1 - \frac{\sqrt{4 - (1-\alpha)^2}}{2}$$

and that for all j and n,

$$\sum_{k\in\Gamma_{sup}(j)} e^{-\frac{1}{2}\left[(\lambda(C_k-C_{k_n^*})+\|(M_k-M_{k_n^*})\mu_n\|^2/\sigma^2)\cdot\frac{1+\alpha}{2}+\alpha\eta_n(m_k-m_{k_n^*})\right]} \leq e^{\zeta_n\eta_n j}.$$

$(G3^{'})$. For $k \in \Gamma_w$, $T_k^{'} \geq c\|(M_{k_n^*} - M_k)\mu_n\|^2/\sigma^2$ for some constant $0 < c < 1$, and $\sum_{k\in\Gamma_w} e^{-cT_k^{'}/8} \to 0$.

Theroem 9

For choices of B_k satisfying the constraint ((3.1)), under Conditions $(G1)$, $(G2^{'})$, and $(G3^{'})$, we have

$$P\left(\min_{\{k\in\Gamma,k\neq k_n^*\}} GICC_\lambda(k) > GICC_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n\to\infty. \qquad \square$$

Remarks:

1. The Conditions $(G2^{'})$ and $(G3^{'})$ are almost the same as Conditions $(G2)$ and $(G3)$. The difference is due to the fact that we not necessarily have $\|(I_n - M_{k_n^*})\mu_n\|^2 = 0$ any more. In the case that k_n^* is the true model, it is easy to check that Conditions $(G2)$ and $(G3)$ are exactly the same as Conditions $(G2^{'})$ and $(G3^{'})$.

2. The first inequality in Condition $(G2^{'})$ says that for an overfitting model $k \supset k_n^*$, the reduction in the approximation error, $\|(M_k - M_{k_n^*})\mu_n\|^2/\sigma^2$, should be smaller than a fraction of the increase of the penalty term, $(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})$. This makes intuitive sense. Otherwise, an overfitting model is going to better than the true model in terms of the criterion $GICC$. In the case k_n^* is the true model, this inequality is typically automatically satisfied since $\|(M_k - M_{k_n^*})\mu_n\|^2 = 0$ for any model $k \supset k_n^*$.

3. The second inequality in Condition $(G2^{'})$ is the similar to Condition $(G2)$ which controls the overfitting probability.

4. Condition $(G3^{'})$ controls the probability of choosing a wrong model. It has similar interpretation to Condition $(G3)$, as is discussed in the remarks after Lemma 4 and Lemma 5.

order selection

Similar to Theorem 6, we also provide simplified conditions for this general consistency in the order selection setting. Again suppose we have a ordered list of predictors $x_1, \cdots, x_{p_n}$ and the candidate list Γ contains only those models following the order of the predictors. We point out that the p_n predictors could be the basis functions of a series expansion in a nonparametric situation or just regular predictors. The model k_n^* minimizes $\|(I_n - M_k)\mu_n\|^2 + r_k\eta_n^{'}\sigma^2$.

Theroem 10

In the above order selection setting, choose $\lambda = 0$, if $\eta_n \to \infty$ and there exist $0 < c < 1$ and $\delta_n \to \infty$ such that $(1-c)\|(M_{k_n^*} - M_k)\mu_n\|^2/\sigma^2 - (m_{k_n^*} - m_k)\eta_n \geq \frac{8}{c}(\log(m_{k_n^*}) + \delta_n)$ for $k \in \Gamma_w$ and $\eta_n > \frac{2-c_0}{c_0}\eta_n^{'}$ for some $0 < c_0 < 1 - \frac{\sqrt{3}}{2}$, then

$$P\left(\min_{\{k\in\Gamma, k\neq k_n^*\}} GICC_\lambda(k) > GICC_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \qquad \square$$

Remarks:

1. Still since the candidate list Γ is simple enough, we do not need additional complexity penalty C_k to control the probability of overfitting. As long as $\eta_n \to \infty$ and $\eta_n > \frac{2-c_0}{c_0}\eta_n^{'}$, the Condition $(G2^{'})$ is satisfied.

2. The condition in Theorem 10 says for models in Γ_w, the increase in the approximation error, $\|(M_{k_n^*} - M_k)\mu_n\|^2/\sigma^2$, needs to be larger than the number of

predictors those models missed times η_n plus a multiple of $(\log(m_{k_n^*}) + \delta_n)$.

3. Typically, $\eta_n^{'}$ is bounded, for instance $\eta_n^{'} = 2$, then the condition $\eta_n > \frac{2-c_0}{c_0}\eta_n^{'}$ is well satisfied.

4. The rate at which p_n or δ_n goes to infinity is arbitrary.

We now take a closer look at the conditions in Theorem 10 to gain a better understanding. For simplicity, we consider a series expansion of the regression function: $f(\mathbf{X}) = \sum_{i=1}^{\infty} \beta_i \phi_i(\mathbf{X})$ with $\phi_i(\mathbf{X})$ being orthogonal at the design points. We assume that $\phi_i(\mathbf{X})$ is standardized such that $\|\phi_i(\mathbf{X})\|^2/\sigma^2 = \sum_{j=1}^{n} \phi_i^2(\mathbf{X}_j)/\sigma^2 = n$. We first demonstrate that the conditions can not be satisfied in the gradual decay case with the coefficient $\beta_i \sim i^{-\gamma}$ for some $\gamma > 0$. Note that in this scenario, the approximation error of model k, $\|(M_{k_n^*} - M_k)\mu_n\|^2/\sigma^2$, is $\sum_{i=k+1}^{n} n\beta_i^2 + \eta_n^{'}k = n\sum_{i=k+1}^{n} i^{-2\gamma} + \eta_n^{'}k = \frac{n}{2\gamma-1}(k+1)^{1-2\gamma} + \eta_n^{'}k$ when $\gamma > 1/2$. Thus the best trade-off is: $k_n^* \approx \left[\left(\frac{n}{(2\gamma-1)\eta_n^{'}}\right)^{\frac{1}{2\gamma}} - 1\right]$. But for model $k \in \Gamma_w$ with $k = k_n^* - 1$, we have $(1-c)\|(M_{k_n^*} - M_k)\mu_n\|^2/\sigma^2 - (m_{k_n^*} - m_k)\eta_n = (1-c)n\beta_{k_n^*}^2 - \eta_n = (1-c)n\,(k_n^*)^{-2\gamma} - \eta_n \approx (1-c)(2\gamma-1)\eta_n^{'} - \eta_n$. Since $\eta_n > \eta_n^{'}$, the conditions in Theorem 10 cannot be satisfied in this scenario. The intuitive answer is that in this gradual decay case, there are models which are very close to the best model in terms of approximation error. Thus it is almost impossible to identify the best model consistently. Note that in a parametric scenario, all the coefficients β_i become zero for $i > k_n^*$. In parallel to this situation, in a nonparametric scenario, if the first part of the coefficient sequence is large enough and then become small thereafter, we can still have consistency.

Corollary 6

Consider a series expansion of the regression function: $f(\mathbf{X}) = \sum_{i=1}^{\infty} \beta_i \phi_i(\mathbf{X})$ with $\phi_i(\mathbf{X})$ being orthogonal at the design points and $\phi_i(\mathbf{X})$ standardized such that $\|\phi_i(\mathbf{X})\|^2/\sigma^2 = \sum_{j=1}^{n} \phi_i^2(\mathbf{X}_j)/\sigma^2 = n$. Let Γ be the collection of models in the order selection setting.

Choose $\lambda = 0$ and assume $\eta_n \to \infty$ and $\eta_n > \frac{2-c_0}{c_0}\eta_n^{'}$ for some $0 < c_0 < 1 - \frac{\sqrt{3}}{2}$. If there exist $0 < c < 1$ and $\delta_n \to \infty$ such that $|\beta_i| > \sqrt{\frac{\eta_n}{n(1-c)} + \frac{8(\log(K_n)+\delta_n)}{c(1-c)n}}$ for all $i \leq K_n$ and $|\beta_i| < \sqrt{\frac{\eta_n^{'}}{n}}$ for $i > K_n$, where $K_n \in \mathbb{Z}^+$, then $k_n^* = K_n$ and

$$P\left(\min_{\{k\in\Gamma,\ k\neq k_n^*\}} GICC_\lambda(k) > GICC_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \qquad \square$$

Remarks:

1. The scenario in the Corollary 6 is a very simplified situation in the order selection setting. What we show here is just the essence.

2. Note that the scenario in the Corollary 6 is a natural generalization of the parametric scenario where $\beta_i = 0$ for $i > K_n$. Essentially, this is a "practically parametric" scenario. For more discussion on practical parametricness, see [51].

3. The K_n could diverge to infinity, in which case we require $K_n < n$.

3.5.2 σ^2 unknown

For the case where σ^2 is unknown, we still consider model selection criterion: $GICC_\lambda^{'}(k) = n\log(\hat{\sigma}_k^2) + m_k\eta_n + \lambda C_k$ with $0 < \eta_n \to \infty$, where $\hat{\sigma}_k^2 = \frac{RSS_k}{n}$.

For $w > -1$, denote $T_k^{'}(w) = \mu_n^T(M_{k_n^*} - M_k)\mu_n/\sigma^2 + \frac{1}{1+w}\left[(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})\right] - B_k$.

Conditions:

($G4^{'}$). Assume $p_n = o(n)$ and there exists a partition of $\Gamma = \Gamma_s \cup \Gamma_l$ such that $k_n^* \in \Gamma_s$, $\limsup_{n\to\infty} \frac{\sup_{k\in\Gamma_s} \|(I_n - M_k)\mu_n\|^2/\sigma^2}{n} = 0$, $\liminf_{n\to\infty} \frac{\inf_{k\in\Gamma_l} \|(I_n - M_k)\mu_n\|^2/\sigma^2}{n} > 0$, and if $\inf_{k\in\Gamma_l}(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) < 0$, then $\liminf_{n\to\infty} \frac{\inf_{k\in\Gamma_l}(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})}{n} = 0$.

($G5'$). There exist constants $0 < \epsilon < 1, 0 < c < 1$ such that for $k \in \Gamma_w$, $\min\left(T_k'(\epsilon), \quad T_k'(-\epsilon)\right) \geq c\|(M_{k_n^*} - M_k)\mu_n\|^2/\sigma^2$, and $\sum_{k\in(\Gamma_w\cap\Gamma_s)}\left[e^{-cT_k'(\epsilon)/8} + e^{-cT_k'(-\epsilon)/8}\right] \to 0$.

Theroem 11

For choices of B_k satisfying the constraint ((3.1)), under Conditions ($G1$), ($G2'$), ($G4'$) and ($G5'$) we have

$$P\left(\min_{\{k\in\Gamma, k\neq k_n^*\}} GICC_\lambda'(k) > GICC_\lambda'(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{3.14}$$

□

Remarks:

1. Conditions ($G4'$) and ($G5'$) are almost the same as Conditions ($G4$) and ($G5$). They have the same interpretation, as is presented in the remarks after Theorem 7.

2. The difference between Conditions ($G4'$) and ($G5'$) and Conditions ($G4$) and ($G5$) is again due to the fact that we not necessarily have $\|(I_n - M_{k_n^*})\mu_n\|^2 = 0$ any more. In the case that k_n^* is the true model, they are exactly the same.

Similar to Theorem 8, for an arbitrarily large p_n, in the general concept of consistency, we need a little modification of the Conditions ($G2'$) and ($G3'$).

Conditions:

($G7'$). Denote $\hat{\Gamma} = \{k : k \in \Gamma,\ a < \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{ref}^2} < b\}$, $c_1 = \frac{\log(b/a)}{b/a-1}$, $c_2 = \frac{\log(b/a)}{1-a/b}$. Note that $0 < c_1 < 1$, $c_2 > 1$. Suppose there exist constants $0 < c < 1$, $0 < w_1 < c_1$, and $w_2 > c_2$ such that for $k \in \Gamma_w$, $\min\left(T_k'(w_1 - 1),\ T_k'(w_2 - 1)\right) \geq c\|(M_{k_n^*} - M_k)\mu_n\|^2/\sigma^2$, and $\sum_{k\in\Gamma_w}\left[e^{-cT_k'(w_1-1)/8} + e^{-cT_k'(w_2-1)/8}\right] \to 0$.

$(G8')$. There exist $0 \leq \alpha < 1$ and $0 \leq \zeta_n \to 0$ such that

$$\limsup_{n\to\infty} \sup_{k\in\Gamma_{sup}} \frac{2\|(M_k - M_{k_n^*})\mu_n\|^2/\sigma^2 + (r_k - r_{k_n^*})}{\|(M_k - M_{k_n^*})\mu_n\|^2/\sigma^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})} \leq 1 - \frac{\sqrt{4-(1-\alpha)^2}}{2}$$

and that for all j and n,

$$\sum_{k\in\Gamma_{sup}(j)} e^{-\frac{1}{2}\left[(\lambda(C_k - C_{k_n^*}) + \|(M_k - M_{k_n^*})\mu_n\|^2/\sigma^2)\cdot\frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]\cdot\frac{1}{w_2}} \leq e^{\zeta_n\eta_n j}.$$

Theroem 12

For choices of B_k satisfying the constraint ((3.1)), under Conditions (G1),(G6), $(G7')$, and $(G8')$ we have $P(k_n^* \in \hat{\Gamma}) \to 1$ and

$$P\left(\min_{\{k\in\Gamma\cap\hat{\Gamma},\ k\neq k_n^*\}} GICC'_\lambda(k) > GICC'_\lambda(k_n^*)\right) \to 1 \quad \text{as} \quad n \to \infty. \tag{3.15}$$

□

Remarks:

1. For the interpretations of the conditions, see the remarks after Theorem 8 and Theorem 9.

3.6 A risk bound for GICC on regression estimation

In this section, we provide a risk bound for the regression estimator based on the model selection criterion $GICC$. This bound holds regardless whether the selected model is the true model or not. The risk bound is due to the incorporation of the complexity penalty term C_k. We reference [71] for a detailed explanation and

illustration of complexity penalty. The result here is essentially the same as the one in [71].

Let $f(X) = X^T\beta$ be the regression function and denote $f_n = (f(X_1), \cdots, f(X_n))^T$. Denote

$$R_n(f;k) = \frac{1}{n}\|(I_n - M_k)f_n\|^2 + \frac{(m_k \log(n) - r_k)\sigma^2}{n} + \frac{\lambda\sigma^2 C_k}{n},$$

$$R_n^{(0)}(f;\Gamma) = \min_{k\in\Gamma} R_n(f;k), \quad k_n^{(0)} = arg\min R_n(f;k).$$

Let $ASE(k) = \|f_n - \hat{\mathbf{Y}}_k\|^2/n$ denote the average squared error of the estimator of the regression function from model k. Let $\hat{k}_n$ be the model selected by minimizing $GICC_\lambda(k)$ over $k \in \Gamma$. Denote $C_k' = C_k + (m_k \log(n) - 2r_k)/\lambda$.

Theroem 13

Assume $\sum_{k\in\Gamma} e^{-C_k'} \le 1$. When $\lambda \ge 8, \log(n) > 2$, we have

$$E_n\left(ASE(\hat{k}_n) + \frac{(m_{\hat{k}_n}\log(n) - 2r_{\hat{k}_n})\sigma^2}{n} + \frac{\lambda\sigma^2 C_{\hat{k}_n}}{n}\right) \le \xi R_n^{(0)}(f;\Gamma), \tag{3.16}$$

where ξ is a constant depending only on λ. □

Remarks:

1. Note that $R_n^{(0)}(f;\Gamma)$ characterizes the best trade-off among the approximation error, estimation error, and model complexity over all the models in Γ.

2. The Theorem says that the statistical risk of the regression estimator based on the selected model $\hat{k}_n$ with the criterion $GICC$ is well bounded by that of the best model in the candidate list Γ.

3. There is no restriction on the size of the list Γ. For more discussion on the interpretation, we reference [71].

4. Note that when we require $\sum_{k\in\Gamma} e^{-C_k} \leq 1$, as is explained in the subsection Section 3.2, we automatically have $\sum_{k\in\Gamma} e^{-C_k'} \leq 1$.

3.7 Conclusions

Model/variable selection with a diverging number of predictors has received much attention recently not only because it is an interesting theoretical exercise, but more because it has a lot of applications in many scientific fields. Traditional selection criteria such as BIC have been shown to be problematic in these situations. The recent successful efforts by Chen and Chen [18] and by Wang et al [69] showed that modifications of BIC can lead to consistency in variable selection. They also call for a more general understanding of consistency of information criteria for high dimensional data, which is not only of great interest on its own, but also serves as a benchmark when studying computationally fast model selection rules. In this paper, we have provided sufficient conditions for consistency of the selection criteria $GICC$ and $GICC'$, which is a generalization of the BIC-type criteria to deal with much increased complexity of the list of models for high-dimensional regression. We have showed that our result is more general in that the restriction on the size of the true model is lifted and the requirement on the approximation errors of wrong models (or the magnitude of the coefficients in the true model) is relaxed without restrictive conditions on correlations of the predictors. By providing counterexamples, it is shown that the sufficient conditions are needed in general. We have also generalized the result to a new concept of consistency and derived a risk bound for estimating the regression function.

3.8 Proofs

In this section, we provide the proofs of the Theorems and Lemmas in previous sections. W.L.O.G., we assume $\sigma^2 = 1$ in the following proofs for the cases where σ^2 is known.

The following two facts will be used in our proofs (see [71]).

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

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

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

Proof 3.1 (Proof of Theorem 4)

We first consider the case $k \in \Gamma_{sup}$ and show that

$$\sum_{k \in \Gamma_{sup}} P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \to 0. \tag{3.17}$$

Notice that in this case

$$GICC_\lambda(k) - GICC_\lambda(k_n^*) = -e_n^T(M_k - M_{k_n^*})e_n + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \tag{3.18}$$

Since $k_n^* \subset k$, if $r_k - r_{k_n^*} = 0$, then $M_k = M_{k_n^*}$. According to Condition (G1), we have $P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) = 0$. In the following, we assume $r_k - r_{k_n^*} > 0$.

By fact 2, we have

$$\begin{aligned}
& P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \\
= \;& P\left(\chi^2_{r_k - r_{k_n^*}} \geq (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})\right) \\
\leq \;& e^{-\frac{r_k - r_{k_n^*}}{2}\left[\left(\frac{\lambda(C_k - C_{k_n^*}) + (m_k - m_{k_n^*})\eta_n}{r_k - r_{k_n^*}} - 1\right) - \log\left(\frac{\lambda(C_k - C_{k_n^*}) + (m_k - m_{k_n^*})\eta_n}{r_k - r_{k_n^*}}\right)\right]}.
\end{aligned}$$

Note that for any given $0 \leq \alpha < 1$, there exists a constant $A > 0$ such that $x - \log(1+x) > \frac{1+\alpha}{2}x$ for $x > A$. By Condition $(G1)$, $\frac{\lambda(C_k - C_{k_n^*}) + (m_k - m_{k_n^*})\eta_n}{r_k - r_{k_n^*}} > \log(\eta_n)$ and since $\eta_n \to \infty$, then there exists n_α such that when $n > n_\alpha$ we have $\log(\eta_n) > A + 1$ and

$$\begin{aligned}
& \sum_{k \in \Gamma_{sup}} P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \\
\leq & \sum_{k \in \Gamma_{sup}} e^{-\frac{r_k - r_{k_n^*}}{2}\left[\left(\frac{\lambda(C_k - C_{k_n^*}) + (m_k - m_{k_n^*})\eta_n}{r_k - r_{k_n^*}} - 1\right) - \log\left(\frac{\lambda(C_k - C_{k_n^*}) + (m_k - m_{k_n^*})\eta_n}{r_k - r_{k_n^*}}\right)\right]} \\
\leq & \sum_{k \in \Gamma_{sup}} e^{-\frac{r_k - r_{k_n^*}}{2}\left(\frac{\lambda(C_k - C_{k_n^*}) + (m_k - m_{k_n^*})\eta_n}{r_k - r_{k_n^*}} - 1\right) \cdot \frac{1+\alpha}{2}} \\
\leq & \sum_{j=1}^{p_n - m_{k_n^*}} \sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*}) + (\eta_n - 1)(m_k - m_{k_n^*})\right] \cdot \frac{1+\alpha}{2}} \\
= & \sum_{j=1}^{p_n - m_{k_n^*}} e^{-\frac{j}{2} \cdot (\frac{1-\alpha}{2}\eta_n - \frac{1+\alpha}{2})} \sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*}) \cdot \frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} \\
< & \sum_{j=1}^{p_n - m_{k_n^*}} e^{-\frac{j}{2} \cdot (\frac{1-\alpha}{2}\eta_n - 1)} \cdot e^{\zeta_n \eta_n j} \longrightarrow 0.
\end{aligned}$$

We now consider the case $k \in \Gamma_w$ and show that

$$\sum_{k \in \Gamma_w} P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \to 0. \tag{3.19}$$

Note that

$$\begin{aligned}
GICC_\lambda(k) - GICC_\lambda(k_n^*) = & \|(I_n - M_k)\mu_n\|^2 + 2\mu_n^T(I_n - M_k)e_n + e_n^T(M_{k_n^*} - \tilde{M}_k)e_n \\
& - e_n^T(M_k - \tilde{M}_k)e_n + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}).
\end{aligned}$$

By constraint ((3.1)), $\sum_{k \in \Gamma_w} P\left(e_n^T(M_k - \tilde{M}_k)e_n \geq B_k\right) \to 0$.

By Condition $(G3)$,

$$\begin{aligned}\sum_{k\in\Gamma_w} P\left(2rem_1(k)+T_k\leq 0\right) &\leq \frac{1}{2}\sum_{k\in\Gamma_w} e^{-\frac{T_k^2}{8\|(I_n-M_k)\mu_n\|^2}} \\ &\leq \frac{1}{2}\sum_{k\in\Gamma_w} e^{-cT_k/8}\to 0.\end{aligned}$$

Thus

$$\begin{aligned}&\sum_{k\in\Gamma_w} P\left(GICC_\lambda(k)-GICC_\lambda(k_n^*)\leq 0\right) \\ \leq\ &\sum_{k\in\Gamma_w}\left[P\left(e_n^T(M_k-\tilde{M}_k)e_n\geq B_k\right)+P\left(2rem_1(k)+T_k\leq 0\right)\right]\to 0.\end{aligned}$$

((3.17)) and ((3.19)) together imply ((3.3)). □

Proof 3.2 (Proof of Lemma 1)

Let $\nu_k^{'}$ be the number of models that have the same m_k and $\tilde{m}_k$. Note that $\nu_k^{'} < \begin{pmatrix} m_{k_n^*} \\ \tilde{m}_k \end{pmatrix}\cdot\begin{pmatrix} p_n \\ m_k-\tilde{m}_k \end{pmatrix} < \nu_k$. Also note that $\frac{\log(\nu_k)}{r_k-\tilde{r}_k} > \log(p_n)\to\infty$ as $n\to\infty$.

Since $c > 2$, then $\frac{1}{2} < \frac{2+c}{2c} < 1$. Thus by fact 2,

$$\begin{aligned}
& \sum_{k\in\Gamma_w} P\left(e_n^T(M_k - \tilde{M}_k)e_n \geq B_k\right) \\
\leq & \sum_{k\in\Gamma_{temp}} e^{-\frac{r_k-\tilde{r}_k}{2}\left[\frac{c\log(\nu_k)}{r_k-\tilde{r}_k}-1-\log(\frac{c\log(\nu_k)}{r_k-\tilde{r}_k})\right]} \\
\leq & \sum_{k\in\Gamma_w} e^{-\frac{r_k-\tilde{r}_k}{2}\left[\frac{c\log(\nu_k)}{r_k-\tilde{r}_k}\right]\cdot\frac{2+c}{2c}} \qquad \text{when } n \text{ is large enough} \\
= & \sum_{k\in\Gamma_w} e^{-\frac{c+2}{4}\log(\nu_k)} < \sum_{m_k,\ \tilde{m}_k} \nu_k \cdot e^{-\frac{c+2}{4}\log(\nu_k)} < \sum_{\tilde{m}_k=0}^{m_{k_n^*}-1} \sum_{m_k=\tilde{m}_k}^{p_n} \nu_k^{-\frac{c-2}{4}} \\
= & \sum_{\tilde{m}_k=0}^{m_{k_n^*}-1} \left[m_{k_n^*}^{-\frac{c-2}{4}}\right]^{m_{k_n^*}-\tilde{m}_k} \cdot \sum_{m_k=\tilde{m}_k}^{p_n} \left[p_n^{-\frac{c-2}{4}}\right]^{m_k-\tilde{m}_k} \\
< & \sum_{\tilde{m}_k=0}^{m_{k_n^*}-1} \left[m_{k_n^*}^{-\frac{c-2}{4}}\right]^{m_{k_n^*}-\tilde{m}_k} \cdot \frac{1}{1-p_n^{-\frac{c-2}{4}}} < \frac{(m_{k_n^*})^{-\frac{c-2}{4}}}{1-(m_{k_n^*})^{-\frac{c-2}{4}}} \cdot \frac{1}{1-p_n^{-\frac{c-2}{4}}}
\end{aligned}$$

Thus, if $m_{k_n^*} \to \infty$, then the proof is complete. If $m_{k_n^*}$ is bounded, then we can replace $m_{k_n^*}$ in the above by $m_{k_n^*}\log(n)$ and the same argument follows. □

Proof 3.3 (Proof of Lemma 2)

In this case, Condition $(G1)$ is obviously satisfied since $r_k - r_{k_n^*} \leq m_k - m_{k_n^*}$ for $k_n^* \subset k$. We just need to verify Condition $(G2)$ for $\lambda > 2$.

Note that the function $u(x) = \frac{4}{1+x}$ is continuous and decreasing in $x \in (0,1)$ and that $\lim_{x\uparrow 1} u(x) = 2$. Since $\lambda > 2$, then there exists $\alpha \in (0,1)$ such that $u(\alpha) < \lambda$. In

other words, $\lambda \cdot \frac{1+\alpha}{4} > 1$. Thus

$$\begin{aligned}
& \sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*}) \cdot \frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} \\
< \quad & \binom{p_n}{j} \cdot e^{-\frac{1}{2}\left[\lambda j \log(p_n) \cdot \frac{1+\alpha}{2} + \alpha j \eta_n\right]} \\
< \quad & e^{-(\frac{1+\alpha}{4}\lambda - 1) j \log(p_n)} \cdot e^{-\frac{\alpha}{2} j \eta_n} \\
\leq \quad & e^{-\frac{\alpha}{2} j \eta_n}.
\end{aligned}$$

□

The existence of ζ_n is obvious. For instance, we can take $\zeta_n = 0$.

Proof 3.4 (Proof of Lemma 3)

Condition $(G1)$ is obviously satisfied in this case.

If $0 < \kappa < 1/2$, then we show that Condition $(G2)$ is satisfied for $0 < \alpha = 2\kappa < 1$. Note that

$$\begin{aligned}
& \sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*}) \cdot \frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} \\
= \quad & \sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda j \log(p_n) \cdot \frac{1+\alpha}{2} + \alpha j \log(n)\right]} \\
< \quad & \binom{p_n}{j} \cdot e^{-\frac{1}{2}\left[\lambda j \log(p_n) \cdot \frac{1+\alpha}{2} + \alpha j \log(n)\right]} \\
< \quad & p_n^j \cdot e^{-\frac{1}{2}\left[\lambda j \log(p_n) \cdot \frac{1+\alpha}{2} + \alpha j \log(n)\right]} \\
= \quad & e^{-\frac{1}{2}\left[\lambda j \log(p_n) \cdot \frac{1+\alpha}{2} + \alpha j \log(n) - 2j \log(p_n)\right]} \\
= \quad & e^{-\frac{1}{2}\left[\lambda j \kappa \log(n) \cdot \frac{1+\alpha}{2} + \alpha j \log(n) - 2\kappa j \log(n)\right]} \\
= \quad & e^{-\frac{1}{2}\left[\lambda j \kappa \log(n) \cdot \frac{1+\alpha}{2}\right]} \to 0 \quad \text{as} \quad n \to \infty \quad \text{since} \quad \lambda > 0.
\end{aligned}$$

If $\kappa \geq 1/2$, then we show the existence of $0 < \alpha < 1$ such that Condition $(G2)$ is satisfied. Note that when $\kappa \geq 1/2$, the function $g(x) = 2(1 - \frac{x}{2\kappa}) \cdot \frac{2}{1+x}$ is non-negative

and decreasing in $x \in (0,1)$. Also note that $\lim_{x\uparrow 1} g(x) = 2(1-\frac{1}{2\kappa})$. Since $2(1-\frac{1}{2\kappa}) < \lambda$ and $g(x)$ is continuous, then there exists $\alpha \in (0,1)$ such that $g(\alpha) = 2(1-\frac{\alpha}{2\kappa})\cdot\frac{2}{1+\alpha} < \lambda$. That is, $(\lambda \cdot \frac{1+\alpha}{2} - 2)\kappa + \alpha > 0$. Then

$$\begin{aligned}
& \sum_{k\in\Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*})\cdot\frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} \\
< \quad & \binom{p_n}{j} \cdot e^{-\frac{1}{2}\left[\lambda j\log(p_n)\cdot\frac{1+\alpha}{2} + \alpha j\log(n)\right]} \\
= \quad & e^{-\frac{1}{2}\left[\lambda j\log(p_n)\cdot\frac{1+\alpha}{2} + \alpha j\log(n) - 2j\log(p_n)\right]} \\
= \quad & e^{-\frac{1}{2}\left[\lambda\cdot\kappa\cdot\frac{1+\alpha}{2} + \alpha - 2\kappa\right]j\log(n)} \\
= \quad & e^{-\frac{1}{2}\left[(\lambda\cdot\frac{1+\alpha}{2} - 2)\kappa + \alpha\right]j\log(n)} \\
\longrightarrow \quad & 0 \quad \text{as} \quad n\to\infty \quad \text{since} \quad (\lambda\cdot\frac{1+\alpha}{2} - 2)\kappa + \alpha > 0. \qquad \square
\end{aligned}$$

Proof 3.5 (Proof of Lemma 4)

Note that

$$\sum_{k\in\Gamma_w} e^{-cT_k/8} \quad = \quad \sum_{a=1}^{p_n - m_{k_n^*}} \sum_{j=1}^{m_{k_n^*}-1} \sum_{\substack{m_{k_n^*} - \tilde{m}_k = j \\ m_k - \tilde{m}_k = a \\ k\in\Gamma_w}} e^{-\frac{c}{8}T_k}$$

and for any positive sequence $\delta_n \to \infty$, we have

$$\sum_{a=1}^{p_n - m_{k_n^*}} \sum_{j=1}^{m_{k_n^*}-1} \sum_{\substack{m_{k_n^*} - \tilde{m}_k = j \\ m_k - \tilde{m}_k = a \\ k\in\Gamma_w}} e^{-\log\left[\binom{m_{k_n^*}}{j}\cdot\binom{p_n - m_{k_n^*}}{a}\right] - \log(m_{k_n^*}) - \log(p_n) - \delta_n} \quad \le \quad e^{-\delta_n} \to 0. \qquad \square$$

Thus as long as $\frac{c}{8}T_k \ge \log\left[\binom{m_{k_n^*}}{m_{k_n^*} - \tilde{m}_k}\cdot\binom{p_n - m_{k_n^*}}{m_k - \tilde{m}_k}\right] + \log(m_{k_n^*}) + \log(p_n) + \delta_n$ for all $k \in \Gamma_w$, then we have $\sum_{k\in\Gamma_w} e^{-cT_k/8} \to 0$.

Note that $T_k = \|(I_n - M_k)\mu_n\|^2 + \eta_n(m_k - m_{k_n^*}) + \lambda(C_k - C_{k_n^*}) - B_k$. The inequality

in Lemma 4 guarantees the above inequality holds and that $T_k \geq c\|(I_n - M_k)\mu_n\|^2$. Therefore, Condition $(G3)$ is met.

Proof 3.6 (Proof of Corollary 3)

The corollary follows immediately from Lemma 2 and Theorem 4. □

Proof 3.7 (Proof of Theorem 5)

According to Lemma 2, Conditions (G1) and (G2) are satisfied under the assumptions of Theorem 5. We now verify that Condition (G3) is also satisfied.

Similar to the proof of Lemma 4, note that

$$\sum_{k\in\Gamma_w} e^{-cT_k/8} \;=\; \sum_{j=1}^{p_n} \sum_{\substack{m_k=j \\ k\in\Gamma_w}} e^{-\frac{c}{8}T_k}$$

and for any positive sequence $\delta_n \to \infty$, we have

$$\sum_{j=1}^{p_n} \sum_{\substack{m_k=j \\ k\in\Gamma_w}} e^{-\log\left(\binom{p_n}{j}\right)-\log(p_n)-\delta_n} \;\leq\; e^{-\delta_n} \to 0.$$

Thus as long as $\frac{c}{8}T_k \geq \log\left(\binom{p_n}{m_k}\right) + \log(p_n) + \delta_n$ for all $k \in \Gamma_w$, then we have $\sum_{k\in\Gamma_w} e^{-cT_k/8} \to 0$.

Note that $T_k = \|(I_n - M_k)\mu_n\|^2 + \eta_n(m_k - m_{k_n^*}) + \lambda(C_k - C_{k_n^*}) - B_k$. Since $\lambda > 2 + \frac{8}{c}$, then there exists $a > 2$ such that $\lambda > a + \frac{8}{c}$. Take $B_k = am_k \log(p_n)$. Then it suffices to show that there exists $\delta_n \to \infty$ such that $(1-c)\|(I_n - M_k)\mu_n\|^2 + \eta_n(m_k - m_{k_n^*}) + \lambda(C_k - C_{k_n^*}) - B_k > \frac{8}{c}\left[\log\left(\binom{p_n}{m_k}\right) + \log(p_n) + \delta_n\right]$. Or equivalently, $(1-c)\|(I_n - M_k)\mu_n\|^2 + \eta_n m_k + (\lambda - a - \frac{8}{c})m_k \log(p_n) > m_{k_n^*}(\eta_n + \lambda\log(p_n))\frac{8}{c}\left[\log(p_n) + \delta_n\right]$, which is implied by the assumption in Theorem 5 since $\lambda > a + \frac{8}{c}$, $m_k \geq 1$, and $\eta_n \to \infty$.

Therefore, Condition $(G3)$ is met. This completes the proof. □

Proof 3.8 (Proof of Theorem 6)

It is obvious that Condition (G1) is met since $\lambda = 0$ and $0 < \eta_n \to \infty$.

We check that Condition (G2) is met. Note that there is only one model in $\Gamma_{sup}(j)$ for every $j \in \mathbb{Z}^+$. Then for any $\alpha > 0$, we have

$$\sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*}) \cdot \frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} = e^{-\alpha\eta_n j} \to 0.$$

Thus Condition (G2) is well satisfied.

We now check Condition (G3). Again there is only one model of each dimension. Note that

$$\sum_{k \in \Gamma_w} e^{-cT_k/8} \;=\; \sum_{m_k=1}^{m_{k_n^*}-1} e^{-\frac{c}{8}T_k}$$

and $\sum_{j=1}^{m_{k_n^*}-1} e^{-\log(m_{k_n^*}) - \delta_n} \to 0$ for any $\delta_n \to \infty$.

Take $B_k = 0$. Since $\frac{c}{8}\left[(1-c)\|(I_n - M_k)\mu_n\|^2 - (m_{k_n^*} - m_k)\eta_n\right] \geq \log(m_{k_n^*}) + \delta_n$ for $k \in \Gamma_w$, then $\sum_{k \in \Gamma_w} e^{-cT_k/8} \to 0$ and $T_k \geq c\|(I_n - M_k)\mu_n\|^2$. Thus Condition (G3) is satisfied. □

Proof 3.9 (Proof of Corollary 4)

The proof is the same as the one of Theorem 6 except that we take $B_k = 2(r_k - \tilde{r}_k) + \frac{2}{1-\log(2)}(\log(p_n) + \delta_n')$ for any $\delta_n' \to \infty$. □

Proof 3.10 (Proof of Theorem 7)

Note that

$$\begin{aligned} GICC_\lambda'(k) - GICC_\lambda'(k_n^*) &= n\log(1 + \frac{\hat{\sigma}_k^2 - \hat{\sigma}_{k_n^*}^2}{\hat{\sigma}_{k_n^*}^2}) + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\ &= n\log(1 + \frac{RSS_k - RSS_{k_n^*}}{n\hat{\sigma}_{k_n^*}^2}) + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}). \end{aligned}$$

We first consider models in Γ_s.

By Taylor expansion, for $x \to 0$, we have $\log(1+x) = x + o(x)$. We claim that $\sup_{k\in\Gamma_s} \frac{|RSS_k - RSS_{k_n^*}|}{n\hat{\sigma}^2_{k_n^*}} \xrightarrow{p} 0$. Under this claim, we have

$$\begin{aligned} GICC'_\lambda(k) - GICC'_\lambda(k_n^*) &= \frac{RSS_k - RSS_{k_n^*}}{\hat{\sigma}^2_{k_n^*}} + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) + o_p(1) \\ &= \frac{RSS_k - RSS_{k_n^*}}{\sigma^2}(1+o_p(1)) + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) + o_p(1) \end{aligned}$$

The rest of the proof follows similarly to the one of Theorem 4. Now we prove the claim.

Note that $RSS_k - RSS_{k_n^*} = \|(I_n - M_k)\mu_n\|^2 + 2\mu_n^T(I_n - M_k)e_n + e_n^T(M_{k_n^*} - M_k)e_n$. By fact 2, $\sum_{k\in\Gamma} P\left(e_n^T M_k e_n \geq 3\log(\binom{p_n}{m_k}) + 3\log(p_n) + \log(n)\right) \to 0$. Since $m_k \leq p_n = o(n)$, then with probability going to 1, $\sup_{k\in\Gamma} e_n^T M_k e_n = o(n)$.

By fact 1,

$$\begin{aligned} \sum_{k\in\Gamma_s} P\left(|\mu_n^T(I_n - M_k)e_n| \geq \sqrt{n}\|(I_n - M_k)\mu_n\|\right) &\leq \sum_{k\in\Gamma_s} 2e^{-\frac{n}{2}} \\ &\leq 2^{p_n} \cdot 2e^{-\frac{n}{2}} \\ &= 2 \cdot e^{p_n \log(2)} \cdot e^{-\frac{n}{2}} \to 0. \end{aligned}$$

Thus with probability going to 1, $\sup_{k\in\Gamma_s} |\mu_n^T(I_n - M_k)e_n| < \sup_{k\in\Gamma_s} \sqrt{n}\|(I_n - M_k)\mu_n\| = o(n)$.

According to Condition $(G4)$, we have $\sup_{k\in\Gamma_s} \frac{|RSS_k - RSS_{k_n^*}|}{n} \xrightarrow{p} 0$. Notice that $\hat{\sigma}^2_{k_n^*} \xrightarrow{p} \sigma^2 > 0$. Thus $\sup_{k\in\Gamma_s} \frac{|RSS_k - RSS_{k_n^*}|}{n\hat{\sigma}^2_{k_n^*}} \xrightarrow{p} 0$.

Now we consider models in Γ_l.

By fact 1, for $0 < c < 1/2$

$$\begin{aligned}\sum_{k\in\Gamma_l} P\left(|\mu_n^T(I_n - M_k)e_n| \geq c\|(I_n - M_k)\mu_n\|^2\right) &\leq \sum_{k\in\Gamma_l} 2e^{-\frac{c^2}{2}\|(I_n-M_k)\mu_n\|^2} \\ &\leq 2^{p_n}\cdot 2e^{-\frac{c^2}{2}\inf_{k\in\Gamma_l}\|(I_n-M_k)\mu_n\|^2} \\ &= 2\cdot e^{p_n\log(2)}\cdot e^{-\frac{c^2}{2}\inf_{k\in\Gamma_l}\|(I_n-M_k)\mu_n\|^2}.\end{aligned}$$

According to Condition $(G4)$, we know with probability going to 1, $|\mu_n^T(I_n - M_k)e_n| < c\|(I_n - M_k)\mu_n\|^2$ for all $k \in \Gamma_l$.

Note that with probability going to 1, $\sup_{k\in\Gamma} e_n^T M_k e_n = o(n)$. Thus with probability going to 1, $RSS_k - RSS_{k_n^*} > (1-2c)\|(I_n - M_k)\mu_n\|^2 - o(n)$ for all $k \in \Gamma_l$, which implies that $\liminf_{n\to\infty}\inf_{k\in\Gamma_l}\log(1 + \frac{RSS_k - RSS_{k_n^*}}{n\hat{\sigma}^2_{k_n^*}}) > 0$ with probability going to 1. Since $\inf_{k\in\Gamma_l}(m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) = o(n)$, then with probability going to 1, we have $GICC'_\lambda(k) - GICC'_\lambda(k_n^*) > 0$ for all $k \in \Gamma_l$ when n is large enough.

This completes the proof. □

Proof 3.11 (Proof of Corollary 5)

This Corollary follows immediately from Lemma 2 and Theorem 7. □

Proof 3.12 (Proof of Theorem 8)

First, it is trivial that $P(k_n^* \in \hat{\Gamma}) \to 1$. We now prove the second part.

Note that

$$GICC'_\lambda(k) - GICC'_\lambda(k_n^*) = n\log(\frac{\hat{\sigma}_k^2}{\hat{\sigma}^2_{k_n^*}}) + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}).$$

Since $k \in \hat{\Gamma}$, then $a < \frac{\hat{\sigma}_k^2}{\hat{\sigma}^2_{ref}} < b$ and according to Condition (G6), $\frac{a}{b} < \frac{\hat{\sigma}_k^2}{\hat{\sigma}^2_{k_n^*}} < \frac{b}{a}$.

Notice that for $1 \leq x < \frac{b}{a}$, we have $\log(x) \geq c_1(x-1)$, where $0 < c_1 = \frac{\log(b/a)}{b/a-1} < 1$, and for $\frac{a}{b} < x \leq 1$, we have $\log(x) \geq c_2(x-1)$, where $c_2 = \frac{\log(b/a)}{1-a/b} > 1$.

Thus for $k \in \hat{\Gamma}$, we have

$$\begin{aligned} GICC'_\lambda(k) - GICC'_\lambda(k_n^*) &\geq c_1 n(\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k_n^*}^2} - 1) + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\ &= c_1 n \frac{\hat{\sigma}_k^2 - \hat{\sigma}_{k_n^*}^2}{\hat{\sigma}_{k_n^*}^2} + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\ &= c_1 \frac{RSS_k - RSS_{k_n^*}}{\hat{\sigma}_{k_n^*}^2} + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\ &= c_1 \frac{RSS_k - RSS_{k_n^*}}{\sigma^2}(1 + o_p(1)) + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \end{aligned}$$

or

$$GICC'_\lambda(k) - GICC'_\lambda(k_n^*) \geq c_2 \frac{RSS_k - RSS_{k_n^*}}{\sigma^2}(1 + o_p(1)) + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \quad \square$$

The rest of the proof follows similarly as the one of Theorem 4.

Proof 3.13 (Proof of Theorem 9)

The proof is similar to the one of Theorem 4.

We first consider the case $k \in \Gamma_{sup}$ and show that

$$\sum_{k \in \Gamma_{sup}} P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \to 0.$$

In this case,

$$GICC_\lambda(k) - GICC_\lambda(k_n^*) = -\|(M_k - M_{k_n^*})\mathbf{Y}_n\|^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})$$

and $\|(M_k - M_{k_n^*})\mathbf{Y}_n\|^2 \sim \chi^2_{(r_k - r_{k_n^*}),\ \|(M_k - M_{k_n^*})\mu_n\|^2}$.

According to the Lemma 8.1 in [**?**], for a non-central χ^2 variable with D degrees of freedom and non-centrality parameter $B \geq 0$, we have $P\left(\chi^2_{D,\ B} \geq A\right) \leq$

$e^{-\frac{1}{2}\left[A+B-\sqrt{(2A-D)(D+2B)}\right]}$ for any $A \geq D+B$.

Then

$$\begin{aligned} & P\left(GICC_\lambda(k) - GICC_\lambda(k_n^* \leq 0\right) \\ = \; & P\left(\chi^2_{(m_k - m_{k_n^*}),\ \|(M_k - M_{k_n^*})\mu_n\|^2} \geq (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})\right) \\ \leq \; & e^{-\frac{1}{2}\left[t-\sqrt{(2t-x)x}\right]}, \end{aligned}$$

where $t = \|(M_k - M_{k_n^*})\mu_n\|^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})$ and $x = 2\|(M_k - M_{k_n^*})\mu_n\|^2 + (r_k - r_{k_n^*})$.

Note that when $0 \leq u < 1 - \frac{\sqrt{4-(1-\alpha)^2}}{2}$ for some $0 \leq \alpha < 1$, then $\sqrt{(2-u)u} < \frac{1-\alpha}{2}$. According to Condition $(G2')$, we have $\frac{x}{t} < 1 - \frac{\sqrt{4-(1-\alpha)^2}}{2}$ when n is large enough and $t - \sqrt{(2t-x)x} > \frac{1+\alpha}{2}t$ since $t > 0$.

Thus under Condition $(G2')$, we have

$$\begin{aligned} & \sum_{k \in \Gamma_{sup}} P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \\ \leq \; & \sum_{k \in \Gamma_{sup}} e^{-\frac{1}{2}\left[\|(M_k - M_{k_n^*})\mu_n\|^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})\right] \cdot \frac{1+\alpha}{2}} \\ = \; & \sum_{j=1}^{p_n - m_{k_n^*}} \sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[\lambda(C_k - C_{k_n^*}) + \|(M_k - M_{k_n^*})\mu_n\|^2 + (m_k - m_{k_n^*})\eta_n\right] \cdot \frac{1+\alpha}{2}} \\ = \; & \sum_{j=1}^{p_n - m_{k_n^*}} e^{-\frac{j}{2} \cdot \frac{1-\alpha}{2}\eta_n} \sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[(\lambda(C_k - C_{k_n^*}) + \|(M_k - M_{k_n^*})\mu_n\|^2) \cdot \frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} \\ < \; & \sum_{j=1}^{p_n - m_{k_n^*}} e^{-\frac{j}{2} \cdot \frac{1-\alpha}{2}\eta_n} \cdot e^{\zeta_n \eta_n j} \longrightarrow 0. \end{aligned}$$

We now consider the case $k \in \Gamma_w$ and show that

$$\sum_{k \in \Gamma_w} P\left(GICC_\lambda(k) - GICC_\lambda(k_n^*) \leq 0\right) \to 0.$$

In this case,

$$\begin{aligned}
& GICC_\lambda(k) - GICC_\lambda(k_n^*) \\
= \; & \|(I_n - M_k)\mathbf{Y}_n\|^2 - \|(I_n - M_{k_n^*})\mathbf{Y}_n\|^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\
= \; & \|(I_n - M_k)\mu_n\|^2 + 2\mu_n^T(I_n - M_k)e_n + e_n^T(I_n - M_k)e_n \\
& - \left(\|(I_n - M_{k_n^*})\mu_n\|^2 + 2\mu_n^T(I_n - M_{k_n^*})e_n + e_n^T(I_n - M_{k_n^*})e_n\right) \\
& + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\
= \; & \mu_n^T(M_{k_n^*} - M_k)\mu_n + 2\mu_n^T(M_{k_n^*} - M_k)e_n + e_n^T(M_{k_n^*} - M_k)e_n \\
& + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}) \\
= \; & \mu_n^T(M_{k_n^*} - M_k)\mu_n + 2\mu_n^T(M_{k_n^*} - M_k)e_n + e_n^T(M_{k_n^*} - \tilde{M}_k)e_n \\
& - e_n^T(M_k - \tilde{M}_k)e_n + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*}).
\end{aligned}$$

The rest of the proof is the same as the one in Theorem 4. □

Proof 3.14 (Proof of Theorem 10)

Similar to the proof of Theorem 6, it is easy to check that Condition $(G3')$ is met. We now check Condition $(G2')$.

Since there is only one model in $\Gamma_{sup}(j)$, then for any $\alpha > 0$, we have

$$\sum_{k \in \Gamma_{sup}(j)} e^{-\frac{1}{2}\left[(\lambda(C_k - C_{k_n^*}) + \|(M_k - M_{k_n^*})\mu_n\|^2/\sigma^2)\cdot\frac{1+\alpha}{2} + \alpha\eta_n(m_k - m_{k_n^*})\right]} \leq e^{-\frac{\alpha}{2}j\eta_n} \to 0.$$

Thus the existence of α and ζ_n is obvious.

Note that for every $k \in \Gamma_{sup}$,

$$\|(I_n - M_k)\mu_n\|^2 + \eta_n' m_k \geq \|(I_n - M_{k_n^*})\mu_n\|^2 + \eta_n' m_{k_n^*}.$$

Since $\frac{c_0}{2-c_0}\eta_n > \eta_n^{'}$ and $m_k \geq m_{k_n^*}$ for $k \in \Gamma_{sup}$, we have for $k \in \Gamma_{sup}$,

$$\|(I_n - M_k)\mu_n\|^2 + \frac{c_0}{2-c_0}\eta_n m_k \geq \|(I_n - M_{k_n^*})\mu_n\|^2 + \frac{c_0}{2-c_0}\eta_n m_{k_n^*}$$

. That is, for $k \in \Gamma_{sup}$, $\|(M_k - M_{k_n^*})\mu_n\|^2 \leq \frac{c_0}{2-c_0}\eta_n(m_k - m_{k_n^*})$.

Thus

$$\limsup_{n\to\infty} \sup_{k\in\Gamma_{sup}} \frac{2\|(M_k - M_{k_n^*})\mu_n\|^2 + (r_k - r_{k_n^*})}{\|(M_k - M_{k_n^*})\mu_n\|^2 + (m_k - m_{k_n^*})\eta_n + \lambda(C_k - C_{k_n^*})} \leq c_0.$$

Since $c_0 < 1 - \frac{\sqrt{3}}{2}$, there exists $0 < \alpha < 1$ such that $c_0 < 1 - \frac{\sqrt{4-(1-\alpha)^2}}{2}$. Then Condition $(G2^{'})$ is satisfied. Since $\lambda = 0$ and $\eta_n \to \infty$, Condition (G1) is also satisfied. This completes the proof. □

Proof 3.15 (Proof of Corollary 6)

The Corollary 6 follows immediately from Theorem 10. □

Proof 3.16 (Proof of Theorem 11)

The proof is similar to the one for Theorem 7 and shall be omitted here. □

Proof 3.17 (Proof of Theorem 12)

The proof is similar to the one for Theorem 8 and shall be omitted here. □

Proof 3.18 (Proof of Theorem 13)

W.L.O.G., we assume $\sigma^2 = 1$.

Recall that $rem_1(k) = e_n^T(f_n - M_k f_n)$ and denote $rem_2(k) = m_k - e_n^T M_k e_n$. For simplicity, we denote $R_n(f;k), R_n^{(0)}(f;\Gamma)$ by $R_n(k), R_n^{(0)}$, respectively.

Note that

$$\begin{aligned} GICC_\lambda(k) &= \|(I_n - M_k)\mathbf{Y}_n\|^2 + m_k \log(n) + \lambda C_k \\ &= \|(I_n - M_k)f_n\|^2 + 2e_n^T(I_n - M_k)f_n + \|(I_n - M_k)e_n\|^2 + m_k \log(n) + \lambda C_k \\ &= \|(I_n - M_k)f_n\|^2 + m_k \log(n) + \lambda C_k + 2rem_1(k) + e_n^T(I_n - M_k)e_n + r_k - r_k \\ &= nR_n(k) + 2rem_1(k) + rem_2(k) + e_n^T e_n \end{aligned} \tag{3.20}$$

and

$$\begin{aligned} & \|f_n - \hat{\mathbf{Y}}_k\|^2 + (m_k \log(n) - 2r_k) + \lambda C_k \\ = \; & \|(I_n - M_k)f_n\|^2 + e_n^T M_k e_n + (m_k \log(n) - 2r_k) + \lambda C_k \\ = \; & nR_n(k) - rem_2(k). \end{aligned} \tag{3.21}$$

Then

$$\begin{aligned} nR_n(\hat{k}_n) &= GICC_\lambda(\hat{k}_n) - 2rem_1(\hat{k}_n) - rem_2(\hat{k}_n) - e_n^T e_n \\ &\leq GICC_\lambda(k_n^{(0)}) - 2rem_1(\hat{k}_n) - rem_2(\hat{k}_n) - e_n^T e_n \\ &= nR_n^{(0)} + 2rem_1(k_n^{(0)}) + rem_2(k_n^{(0)}) - 2rem_1(\hat{k}_n) - rem_2(\hat{k}_n). \end{aligned}$$

Thus

$$\begin{aligned} \frac{R_n(\hat{k}_n)}{R_n^{(0)}} &\leq \frac{nR_n^{(0)} + 2rem_1(k_n^{(0)}) + rem_2(k_n^{(0)}) - 2rem_1(\hat{k}_n) - rem_2(\hat{k}_n)}{nR_n^{(0)}} \\ &= 1 + \frac{2rem_1(k_n^{(0)}) + rem_2(k_n^{(0)}) - 2rem_1(\hat{k}_n) - rem_2(\hat{k}_n)}{nR_n^{(0)}}. \end{aligned}$$

We claim that for $\lambda \geq 8$, there exist constants $\tau_1 > 0, \tau_2 > 0, 2\tau_1 + \tau_2 < 1$ such that for any $0 < \delta < 1$, for each sample size n, with probability no less than $1 - 3\delta$, we

have

$$|rem_1(k)| \leq \tau_1(nR_n(k) + g(\delta)), \tag{3.22}$$

$$|rem_2(k)| \leq \tau_2(nR_n(k) + g(\delta)) \tag{3.23}$$

for all $k \in \Gamma$, where $g(\delta) = \lambda \log_2(1/\delta)$.

Under the above claim, we have with probability no less than $1 - 3\delta$,

$$\frac{R_n(\hat{k}_n)}{R_n^{(0)}} \leq 1 + (2\tau_1 + \tau_2)\frac{R_n(\hat{k}_n)}{R_n^{(0)}} + (2\tau_1 + \tau_2)\left(1 + \frac{2g(\delta)}{nR_n^{(0)}}\right). \tag{3.24}$$

Since $nR_n^{(0)} \geq m_k(\log(n) - 1) \geq m_k \geq 1$, we know that with probability no less than $1 - 3\delta$,

$$\frac{R_n(\hat{k}_n)}{R_n^{(0)}} \leq \frac{1 + (2\tau_1 + \tau_2)(1 + 2g(\delta))}{1 - 2\tau_1 - \tau_2}, \tag{3.25}$$

and according to ((3.21)),

$$\begin{aligned} & \frac{\|f_n - \hat{\mathbf{Y}}_{\hat{k}_n}\|^2 + (m_{\hat{k}_n}\log(n) - 2r_{\hat{k}_n}) + \lambda C_{\hat{k}_n}}{nR_n^{(0)}} \qquad (3.26) \\ \leq & \frac{nR_n(\hat{k}_n) + |rem_2(\hat{k}_n)|}{nR_n^{(0)}} \\ \leq & (1 + \tau_2)\frac{R_n(\hat{k}_n)}{R_n^{(0)}} + \tau_2 g(\delta) \\ \leq & (1 + \tau_2)\frac{1 + (2\tau_1 + \tau_2)(1 + 2g(\delta))}{1 - 2\tau_1 - \tau_2} + \tau_2 g(\delta). \qquad (3.27) \end{aligned}$$

Let $Z_n = \frac{\|f_n - \hat{\mathbf{Y}}_{\hat{k}_n}\|^2 + r_{\hat{k}_n}(\log(n)-2) + \lambda C_{\hat{k}_n}}{nR_n^{(0)}}$, $\xi_1 = \frac{(1+\tau_2)(1+2\tau_1+\tau_2)}{1-2\tau_1-\tau_2}$, and $\xi_2 = \lambda\left[\frac{2(1+\tau_2)(2\tau_1+\tau_2)}{1-2\tau_1-\tau_2} + \tau_2\right]$. Then conditioned on $X_i, 1 \leq i \leq n$, $P(Z_n \geq \xi_1 + \xi_2 \log_2(1/\delta)) \leq 3\delta$. Thus

$$E_n\left(\frac{Z_n - \xi_1}{\xi_2}\right)^+ = \int_0^\infty P\left(\frac{Z_n - \xi_1}{\xi_2} \geq t\right) dt \leq 3\int_0^\infty 2^{-t} dt = \frac{3}{\log(2)}.$$

Let $\xi = \frac{3\xi_2}{\log(2)} + \xi_1$, then we have $E_n\left(ASE(\hat{k}_n) + \frac{r_{\hat{k}_n}(\log(n)-2)}{n} + \frac{\lambda C_{\hat{k}_n}}{n}\right) \leq \xi R_n^{(0)}$.

To complete the proof, we only need to prove the above claim. First, notice that $rem_1(k) \sim N(0, \|(I_n - M_k)f_n\|^2)$. If $\|(I_n - M_k)f_n\|^2 = 0$, then ((3.22)) obviously holds. So we only consider the case where $\|(I_n - M_k)f_n\|^2 \neq 0$.

By fact 1, $P(|rem_1(k)|/\|(I_n - M_k)f_n\| \geq t_k) \leq e^{-t_k^2/2}$. Taking $t_k^2 = 2(C_k' - \log(\delta))$, we have

$$\begin{aligned} P\left(\sup_{k\in\Gamma} \frac{|rem_1(k)|}{\|(I_n - M_k)f_n\| t_k} \geq 1\right) &\leq \sum_{k\in\Gamma} P\left(\frac{|rem_1(k)|}{\|(I_n - M_k)f_n\|} \geq t_k\right) \\ &\leq \sum_{k\in\Gamma} e^{-(C_k' - \log(\delta))} \\ &\leq \delta. \end{aligned}$$

Thus with probability no less than $1 - \delta$, we have for all $k \in \Gamma$,

$$|rem_1(k)| \leq \|(I_n - M_k)f_n\| \sqrt{2(C_k' - \log(\delta))}.$$

For the term $rem_2(k)$, taking $\rho_{1,k} > 0$ and $0 < \rho_{2,k} < 1$ such that

$$\begin{aligned} \frac{r_k}{2}(\rho_{1,k} - \log(1 + \rho_{1,k})) &= C_k' - \log(\delta), \quad (3.28) \\ \frac{r_k}{2}(-\rho_{2,k} - \log(1 - \rho_{2,k})) &= C_k' - \log(\delta), \quad (3.29) \end{aligned}$$

we obtain that

$$\begin{aligned} P\left(rem_2(k) \leq -\rho_{1,k} r_k \text{ for some } k \in \Gamma\right) &\leq \sum_{k\in\Gamma} e^{-(C_k' - \log(\delta))} \leq \delta, \\ P\left(rem_2(k) \geq \rho_{2,k} r_k \text{ for some } k \in \Gamma\right) &\leq \sum_{k\in\Gamma} e^{-(C_k' - \log(\delta))} \leq \delta. \end{aligned}$$

Next if we can show that when λ is large enough (not depending on δ),

$$\begin{aligned} \|(I_n - M_k)f_n\|\sqrt{2(C_k' - \log(\delta))} &\leq \tau_1(nR_n(k) + g(\delta)) \\ \rho_{1,k}r_k &\leq \tau_2(nR_n(k) + g(\delta)) \\ \rho_{2,k}r_k &\leq \tau_2(nR_n(k) + g(\delta)), \end{aligned}$$

then ((3.22)) and ((3.23)) hold, so does the claim. Equivalently, we need

$$\begin{aligned} \|(I_n - M_k)f_n\|\sqrt{2(C_k' - \log(\delta))} &\leq \tau_1\left[\|(I_n - M_k)f_n\|^2 + r_k + \lambda(C_k' - \log(\delta))\right], \\ \rho_{1,k}r_k &\leq \tau_2\left[\|(I_n - M_k)f_n\|^2 + r_k + \lambda(C_k' - \log(\delta))\right], \\ \rho_{2,k}r_k &\leq \tau_2\left[\|(I_n - M_k)f_n\|^2 + r_k + \lambda(C_k' - \log(\delta))\right]. \end{aligned}$$

Let $s = \frac{\|(I_n - M_k)f_n\|}{\sqrt{C_k' - \log(\delta)}}$. It suffices to require that for all $s > 0, \rho_{1,k} > 0$,

$$\begin{aligned} \frac{\sqrt{2}}{\tau_1}s &\leq s^2 + \lambda, \qquad (3.30) \\ (\rho_{1,k}/\tau_2 - 1)r_k &\leq \lambda(C_k' - \log(\delta)). \qquad (3.31) \end{aligned}$$

Using ((3.28)), then ((3.31)) reduces to $(\rho_{1,k}/\tau_2 - 1)\frac{2}{\rho_{1,k} - \log(1+\rho_{1,k})} \leq \lambda$. Thus it suffices to require that

$$\lambda \geq h(\tau_1, \tau_2) = \max\left(\sup_{s>0}\frac{\sqrt{2}}{\tau_1}s - s^2, \quad \sup_{\rho>0}\frac{2(\rho/\tau_2 - 1)}{\rho - \log(1+\rho)}\right).$$

It is easily seen that $h(\tau_1, \tau_2)$ is less than infinity for any $\tau_1 > 0, \tau_2 > 0$. In fact,

$$\begin{aligned} \sup_{s>0}\frac{\sqrt{2}}{\tau_1}s - s^2 &= \frac{1}{2\tau_1^2}. \\ \sup_{\rho>0}\frac{2(\rho/\tau_2 - 1)}{\rho - \log(1+\rho)} &= 2(1 + \frac{1}{\rho}) \quad \text{for some } \rho > \tau_2. \end{aligned}$$

□

Let $\lambda_0 = \min_{0<\tau_2<1} \left(h(\frac{1-\tau_2}{2}, \tau_2)\right)$. Then if $\lambda > \lambda_0$, by continuity, there exist $\tau_1 > 0, \tau_2 > 0, 2\tau_1 + \tau_2 < 1$ such that conditions ((3.30)) and ((3.31)) are satisfied. For example, taking $\tau_1 = 7/24, \tau_2 = 1/3$, then $h(\tau_1, \tau_2) < 8$.

Chapter 4

Summary and future research

Model selection criteria such as AIC and BIC are widely used in practice. They are derived from distinct perspectives: AIC aims at minimizing the Kullback-Leibler divergence between the true distribution and the estimate from a candidate model and BIC tries to select a model that maximizes the posterior model probability. It is well known that AIC is minimax-rate optimal for estimating regression functions in nonparametric scenarios and BIC is consistent in parametric scenarios; however, the general forms of AIC and BIC make it very clear that they and similar criteria (such as GIC in [57]) cannot simultaneously enjoy the properties of consistency in a parametric scenario and asymptotic optimality in a nonparametric scenario [73]. The conflict between AIC and BIC has generated a large debate in the literature. Some say the goal of statistics modeling should be targeting at prediction accuracy, while others claim that it is important to discover the underlying principles of the data. Some researchers have no problem with assuming the existence of the true model, while others think there is no such thing as a true model at all. Despite the debate, from both theory and application perspectives, it is important to be able to statistically distinguish a parametric scenario from a nonparametric one. The measure we developed, PI, has the desired property that with probability going to 1, PI separates typical parametric and nonparametric scenarios. It also advises on

whether identifying the true or best candidate model is feasible at the given sample size or not and on whether AIC is likely better than BIC or not for the data at hand. We also provided our perspectives on the existence of a true model by advocating a practical view.

Traditionally, the consistency property of BIC type of criteria for model selection is derived with a fixed number of predictors. Nowadays, research and applications frequently deal with problems with the number of predictors increasing with the sample size. A natural question arises on whether or not the consistency property still holds in this kind of high dimensional setting. The answer is in the positive direction [18, 69]. We provided further understanding on the conditions of consistency for similar types of criteria and showed that our results are more general and our conditions are more relaxed. We also generalized the concept of consistency and provided similar results to this new concept.

There are a few directions for our future research.

1. In distinguishing between parametric scenarios and nonparametric ones, we only handled the linear regression problems with Gaussian errors. We would like to investigate similar measures for generalized linear models and different distributions of errors.

2. In the assessment of the measure PI, more understanding is needed on the choices of parameters such as λ_n and d and on the choice of the best cutoff value c.

3. We would also like to see the performance of PI with other consistent model selection criteria.

4. We would like to investigate the conditions of consistency for BIC type of criteria with computationally efficient methods such as the SCAD and LASSO.

References

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

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

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

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

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

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

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

[8] Berger, J. O. and Pericchi, L. (2001). Objective Bayesian methods for model selection: introduction and comparison, in *Model Selection*, ed. P. Lahiri, Institute

of Mathematical Statistics Lecture Notes – Monograph Series, **38**, Beachwood Ohio, 135–207.

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

[10] Birgé, L. (2001). An Alternative Point of View On Lepski's Method. *State of the art in probability and statistics (Leiden, 1999)*, 113 - 133, IMS Lecture Notes Monogr. Ser., 36, Inst. Math. Statist., Beachwood, OH.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[28] Fan, J. and Li, R. (2006) Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *In Proc. Int. Congr. Mathematicians*, vol. III (eds M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera), pp. 595-622. Zurich: European Mathematical Society.

[29] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high-dimensional feature space. *J. Roy. Statist. Soc. Ser. B*, **70**, 849-911.

[30] Fan, J. and Peng, H. (2004) Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *Ann. Statist.*, **32**, 928-961.

[31] Faraway, J.J. (1992). On the Cost of Data Analysis. *J. Computational and Graphical Statist.*, **1**, 213-229.

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

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

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

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

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

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

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

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

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

[41] Huang, J., Horowitz, J. and Ma, S. (2008) Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.*, **36**, 587-613.

[42] Huang, J., Ma, S. and Zhang, C. H. (2008) Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica*, **18**, 1603-1618.

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

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

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

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

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

[48] Lam, C. and Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters.*Ann. Statist.*, **36**, 2232-2260.

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

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

[51] Liu, W. and Yang, Y. (2010) Parametric or Nonparametric? A Parametricness Index for Model Selection. *Revised for Ann. Statist.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[69] Wang, H., Li, B., and Leng, C. (2009). Shirinkage Tuning Parameter Selection with a Diverging Number of Parameters. *J. Roy. Statist. Soc. Ser. B*, **71**, 671-683.

[70] Wasserman, L. and Roeder, K. (2009) High-Dimensional Variable Selection. *Ann. Statist.*, **37**, 2178-2201.

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

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

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

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

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

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

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

[78] Zhang, C. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894-942.

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

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

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

[82] Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.*, **105**, 312-323.

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

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

[85] Zou, H. and Zhang, H. H. (2009). On The Adaptive Elastic-Net With A Diverging Number of Parameters. *Ann. Statist.*, **37**, 1733-1751.