KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Electrical Engineering

The Quest for Robust Model Selection Methods in Linear Regression

PRAKASH BORPATRA GOHAIN

Stockholm, Sweden 2022

The Quest for Robust Model Selection Methods in Linear Regression

PRAKASH BORPATRA GOHAIN

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Monday the 31st October 2022, at 10:00 a.m. in room F3, Lindstedtsvägen 26, KTH, Stockholm.

Doctoral Thesis in Electrical Engineering
KTH Royal Institute of Technology
Stockholm, Sweden 2022

ISBN 978-91-8040-368-9
TRITA-EECS-AVL-2022:61

Printed by: Universitetsservice US-AB, Sweden 2022

Dedicated to my parents and sister

Abstract

A fundamental requirement in data analysis is fitting the data to a model that can be used for the purpose of prediction and knowledge discovery. A typical and favored approach is using a linear model that explains the relationship between the response and the independent variables. Linear models are simple, mathematically tractable, and have sound explainable attributes that make them widely ubiquitous in many different fields of applications. Nonetheless, finding the best model (or true model if it exists) is a challenging task that requires meticulous attention.

In this PhD thesis, we consider the problem of model selection (MS) in linear regression with a greater focus on the high-dimensional setting when the parameter dimension is quite large compared to the number of available observations. Most of the existing methods of MS struggle in two major areas, viz., consistency and scale-invariance. Consistency refers to the property of the MS method to be able to pick the true model as the sample size grows large or/and when the signal-to-noise-ratio (SNR) increases. Scale-invariance indicates that the performance of the MS method is invariant and stable to any kind of data scaling. These two properties are very crucial for any MS method. In the field of MS employing information criteria, the Bayesian Information Criterion (BIC) is undoubtedly the most popular and widely used method. However, the new BIC forms including the extended versions designed for the high-SNR scenarios are not invariant to data-scaling and our results indicate that their performance is quite unstable under different scaling scenarios. To eradicate this problem we proposed improved versions of the BIC criterion viz., BIC_{R} and EBIC_{R} where the subscript 'R' stands for robust. BIC_{R} is based on the classical setting of order selection, whereas EBIC_{R} is the extended version of BIC_{R} to handle MS in the high-dimensional setting where it is quite possible that the parameter dimension p also grows with the sample size N. We analyze their performance as N grows large as well as when the noise variance diminishes towards zero, and provide detailed analytical proofs to guarantee their consistency in both cases. Simulation results indicate that the performance of the proposed MS criteria is robust to any data scaling and offers significant improvement in correctly picking the true model. Additionally, we generalize EBIC_{R} to handle the problem of MS in block-sparse high-dimensional general linear regression. Block-sparsity is a phenomenon that is seen in many applications. Nevertheless, the existing MS methods based on information criteria are not designed to handle the block structure of the linear model. The proposed generalization handles the block structure effortlessly and can be employed for MS in any type of linear regression framework.

Sammanfattning

Ett grundläggande behov i dataanalys är att anpassa data till en modell som kan användas för prediktion eller ge ny kunskap. Ett typiskt och föredraget tillvägagångsätt är att använda en linjär modell som förklarar sambandet mellan svaret och de oberoende variablerna. Linjära modeller är enkla, matematiskt hanterbara och har goda förklarande egenskaper som gör dem allmänt förekommande inom många olika användningsområden. Det är fortfarande en utmanande uppgift att hitta den bästa modellen (eller sanna modellen om den finns) och det kräver en noggrann behandling. I denna doktorsavhandling behandlar vi problemet med modellval (MV) i linjär regression med ett speciellt fokus på det högdimensionella fallet när parameterdimensionen är relativt stor jämfört med antalet tillgängliga observationer. De flesta av de befintliga metoderna för MV har problem med antingen konsistens eller skalningsinvarians. Konsistens hänvisar till egenskapen hos MV-metoden att kunna välja den sanna modellen när sampelstorleken blir stor eller/och när signal-brus-förhållandet (SNR) ökar. Skalningsinvarians indikerar att prestandan för MV-metoden är invariant och stabil för alla typer av dataskalning. Dessa två egenskaper är mycket avgörande för alla MV-metoder. Inom området för MV som använder informationskriterier är Bayesian Information Criterion (BIC) utan tvekan den mest populära och mest använda metoden. De nya BIC-formuleringarna inklusive de utökade versionerna designade för scenarierna med högt SNR är dock inte oberoende av dataskalning och våra resultat indikerar att deras prestanda är ganska instabila under olika skalningsscenarier. För att eliminera detta problem föreslår vi förbättrade versioner av BIC-kriteriet, nämligen $\mathrm{BIC_R}$ och $\mathrm{EBIC_R}$ där tillägget 'R' står för robust. $\mathrm{BIC_R}$ är baserad på det klassiska problemet för val av modellordning, medan $\mathrm{EBIC_R}$ är den utökade versionen av $\mathrm{BIC_R}$ för att hantera MV i högdimensionella problem där det är mycket möjligt att parameterdimensionen p också växer med antal sampel N. Vi analyserar deras prestanda när N växer sig stor såväl som när brusvariansen går mot noll och ger detaljerade analytiska bevis för att garantera överensstämmelse i bägge fallen. Simuleringsresultat indikerar att prestandan för de föreslagna MV-kriterierna är robusta för alla dataskalningar och erbjuder betydande förbättringar när det gäller att korrekt välja den sanna modellen. Dessutom generaliserar vi $\mathrm{EBIC_R}$ för att hantera problemet med MV i blockgles högdimensionell allmän linjär regression. Blockglesa modeller förekommer i många tillämpningar. Ändå är de befintliga MV-metoderna baserade på informationskriterier inte utformade för att hantera den linjära modellens blockstruktur. Den föreslagna generaliseringen hanterar blockstrukturen utan ansträngning och kan användas för MV i vilken typ av linjär regression som helst.

List of Papers

This thesis is based on material from the following papers:

I ***Relative Cost based Model Selection for Sparse High-Dimensional Linear Regression Models***
Prakash B. Gohain, Magnus Jansson
IEEE International Conference on Acoustics, Speech and Signal Processing (2020)

II ***Scale-Invariant and Consistent Bayesian Information Criterion for Order Selection in Linear Regression Models***
Prakash B. Gohain, Magnus Jansson
Elsevier Signal Processing (2022)

III ***New Improved Criterion for Model Selection in Sparse High-Dimensional Linear Regression models***
Prakash B. Gohain, Magnus Jansson
IEEE International Conference on Acoustics, Speech and Signal Processing (2022)

IV ***Robust Information Criterion for Model Selection in Sparse High-Dimensional Linear Regression Models***
Prakash B. Gohain, Magnus Jansson
Submitted to IEEE Transaction on Signal Processing (2022) (under revision)

V ***Model Selection in Block-Sparse High-Dimensional Linear Regression***
Prakash B. Gohain, Magnus Jansson
Submitted to IEEE Signal Processing Letters (2022)

Other contributions/collaborations by the author not included in the thesis.

I ***Neural Greedy Pursuit for Feature Selection***
Sandipan Das, Alireza M. Javid, **Prakash B. Gohain**, Yonina C. Eldar, Saikat Chatterjee
IEEE World Congress on Computational Intelligence (2022)

II ***Statistical Model-Based Evaluation of Neural Networks***
Sandipan Das, **Prakash B. Gohain**, Alireza M. Javid, Yonina C. Eldar, Saikat Chatterjee
arXiv preprint arXiv:2011.09015

Acknowledgement

The time spent at KTH has been a wonderful and great learning experience. This thesis would not have been possible without the guidance and support of many people. Here, I wish to extend my gratitude to all those people who helped me in successfully completing this phase of my life.

First and foremost I would like to offer my deepest gratitude to my supervisor Prof. Magnus Jansson for offering me this opportunity to be part of the doctoral program at KTH. I am eternally grateful for your constant support, guidance, and patience throughout my PhD journey. I have always valued our conversations and meetings which have been a great source of knowledge and wisdom. Thank you!

I am also thankful to the Swedish Research Council and the European Research Council for being generous and supporting all our work financially.

I would also like to say special thanks to Assoc. Prof. Saikat, Sandipan, Dr. Alireza, and Prof. Yonina Eldar with whom I had the privilege to collaborate on two remarkable projects. It was a great learning experience and I really enjoyed working with the team.

I would also like to extend my sincere thanks to Prof. Tobias Oechtering for being my advanced reviewer, providing constructive feedback on the thesis, and also for being such a wonderful office neighbour. I am also thankful to Assoc. Prof. Markus Flierl for agreeing to be the chair of the public defense. Taking this opportunity I would like to offer my sincere thanks to all our professors at ISE: Prof. Mikael Skoglund, Prof. Joakim Jaldén, Prof. Mats Bengtsson, Assoc. Prof. Ragnar Thobaben, Prof. James Gross, and Assoc. Prof. Ming Xiao for their support in general.

I would also like to express my sincere gratitude to Prof. K.V.S Hari from the Indian Institute of Science, Bengaluru for acting as the opponent and to the grading committee members: Prof. Rebecka Jörnsten, from Chalmers University, Prof. Mattias Villani from the Stockholm University, Assoc. Prof. Gustaf Hendeby from the Linköping University and Prof. Cristian Rojas from KTH.

Further, I would like to acknowledge my past and present PhD colleagues Ramana, Lissy, Alexander, Dr. Hao, Sahar, Vishnu, Antoine, Anubhab, Amaury, Borja, Wendi, Dr. Håkan, Sara, Movitz, Baptiste, Samie, Linghui, Michail, Javier, Martin, Amirreza, Hamid, Xuechun, Dr. Sina, Dr. Hasan, Dr. Vidit, Manuel, and many others for the pleasant time that we have shared during the course of this

journey. I will cherish our lunch times, Friday fikas, kick-offs, retreats, and other social activities that we did together.

Heartfelt thanks go out to our HR administrators Anneli Ström and Anna Mård for kindly helping me with everything from the very beginning.

Last, but certainly not least, I would like to dedicate this thesis to my parents and my sister. I could not have done this without them. They have always encouraged me through tough times and have been incredibly loving and supportive. Thank you very much!

Prakash Borpatra Gohain
Stockholm, October 2022

Contents

Acronyms

List of commonly used acronyms:

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BMMV	Block Multiple Measurement Vector
BSMV	Block Single Measurement Vector
BOMP	Block Orthogonal Matching Pursuit
EBIC	Extended Bayesian Information Criterion
EFIC	Extended Fisher Information Criterion
FIC	Fisher Information Criterion
FIM	Fisher Information Matrix
GRRT	Generalized Residual Ratio Thresholding
LASSO	Least Absolute Shrinkage and Selection Operator
LARS	Least Angle Regression
MAP	Maximum a-Posteriori
MDL	Minimum Description Length
MBT	Multi Beta Test
MMV	Multiple Measurement Vector
OMP	Orthogonal Matching Pursuit
PAL	Penalizing Adaptively the Likelihood
PCMS	Probability of Correct Model Selection
PCOS	Probability of Correct Order Selection
RRT	Residual Ratio Thresholding
SMV	Single Measurement Vector
SNR	Signal-to-Noise-Ratio
SOMP	Simultaneous Orthogonal Matching Pursuit
cdf	Cumulative Distribution Function
pdf	Probability Density Function
i.i.d.	Independent and Identically Distributed
w.r.t	With Respect To

Mathematical Notation

$\mathbb{R}$	The set of real numbers
$\mathbb{C}$	The set of complex numbers
$\mathbb{N}$	The set of non-negative natural numbers
$\mathcal{I}$	A subset of $\{\mathbb{N}\}$
$\mathbf{A}$	Matrices are presented in upper-case bold fonts
$\mathbf{a}$	Vectors are presented in lower-case bold fonts
$\|\mathbf{a}\|_p$	The p-norm of $\mathbf{a} \in \mathbb{R}^N$ defined as $\|\mathbf{a}\|_p = \left(\sum_{i=1}^N \|a_i\|^p\right)^{1/p}$
$\mathbf{a}_i$	The i-th column of the matrix $\mathbf{A}$
$\mathbf{a}_\mathcal{I}$	The subvector formed from the collection of the elements of the vector $\mathbf{a}$ with support $\mathcal{I}$
$a_{i,j}$	The $(i,j)-$th element of the matrix $\mathbf{A}$
$\mathbf{A}^T$	The transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	The inverse of the non-singular square matrix $\mathbf{A}$
$\|\mathbf{A}\|$	Determinant of the matrix $\mathbf{A}$
$\|\mathbf{A}\|_F$	Frobenius norm of the matrix $\mathbf{A} \in \mathbb{R}^{N\times p}$ defined as follows $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^N \sum_{j=1}^p \|a_{i,j}\|^2}$
$\mathbf{A}_\mathcal{I}$	The submatrix formed from the collection of columns of the matrix $\mathbf{A}$ with support $\mathcal{I}$
$\mathbf{A}_k$	The submatrix formed from the collection of first k columns $\mathbf{A}$
$\mathbf{\Pi}_\mathcal{I}$	The orthogonal projection matrix onto the span of the matrix $\mathbf{A}_\mathcal{I}$
$\mathbf{\Pi}_\mathcal{I}^\perp$	The orthogonal projection matrix onto the null space of the matrix $\mathbf{A}_\mathcal{I}^T$
$\mathbf{I}_N$	The $N \times N$ identity matrix
$\mathbf{A} \otimes \mathbf{B}$	The Kronecker product of $\mathbf{A}$ and $\mathbf{B}$
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$	The Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$
$\chi_k^2(\lambda)$	Non-central chi-squared distribution with k degrees of freedom and non-centrality parameter λ
χ_k^2	Central chi-squared distribution with k degrees of freedom

$\mathcal{B}(\alpha, \beta)$	Beta distribution with parameters α and β
$\mathcal{B}^{-1}(\alpha, \beta)$	Inverse Beta cdf with parameters α and β
$\mathbb{E}(X)$	Expectation of the random variable X
$\text{Var}(X)$	Variance of the random variable X
$X \to Y$	The random variable X converges in distribution to Y.

List of Figures

List of Tables

Chapter 1

Introduction

"Data is the new oil."
-Clive Humby

1.1 Motivation

The 21st century is the era of data that has revolutionized the manner in which decisions are made. In general, large volumes of data can be examined to gain additional insights and extract important information to further enhance the analysis and the decision making process. In this regard, a typical approach used by statisticians, analysts, and data scientists is to employ different statistical methods or machine-learning approaches to fit a model using the available data for making predictions. However, since the true model is unknown, there can be several possible candidate models that can be used to describe the data. The goal herein is then to find the best model among the available candidate models. This is a pivotal step in data analysis because wrong or improper choice of model can produce incorrect predictions resulting in misleading conclusions. Thus, model selection is the task of selecting a model from a set of candidate models, given a set of data [1, 2].

The scope of model selection is quite extensive. It plays a central role in statistical inference in many areas of science, engineering, finance, economics, biology, ecology, etc. In a broader sense, model selection may involve various tasks such as finding the best subset of a linear regression model, estimating the order of a polynomial regression or autoregressive process, estimating the required components in a mixture model, evaluating the number of change points in time series models, and estimating the true variables of a non-linear system. In the deep learning domain model selection may be concerned with finding the optimal numbers of neurons in a layer or choosing the number of hidden layers for a deep neural network, etc. However, we keep the scope of this thesis to statistical models. Below, we provide a brief overview of statistical models and their role in data analysis.

1.2 Statistical Models

Generally, the meaning of the term "model" might vary depending on the situation or the context at hand. In scientific studies, experts belonging to a particular field may use the term model to refer to a particular thing in their domain, which may completely differ from another domain/field. In this thesis, our focus is on statistical models.

Formally, a statistical model is defined as a family of probability distributions, $\mathcal{P}$, on a sample space $\mathbb{S}$, constructed to enable inferences to be drawn or decisions made from data [3, 4]. The reasoning behind the definition is as follows. Typically it is assumed that there exists a "true" probability distribution that generates the data. In this regard, $\mathcal{P}$ represents a collection of distributions that contains a distribution that accurately approximates the true distribution. However, in practice, $\mathcal{P}$ may not always include the true distribution. Hence, the goal here is to approximate the true structure as accurately as possible using the available data. The parameter θ is typically used to specify the distribution in $\mathcal{P}$, where θ belongs to the parameter space Θ. Therefore, the set $\mathcal{P}$ is parameterized, i.e., $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where P_θ is a probability distribution on the sample space $\mathbb{S}$ with respect to parameter $\theta \in \Theta$. The parameterization is said to be identifiable if distinct parameter values give rise to distinct distributions, i.e., $P_{\theta_1} = P_{\theta_2}$ implies $\theta_1 = \theta_2$ [4]. Below we describe the types of statistical models available in the literature.

1. **Parametric:** This is the class of statistical models that has a finite number of parameters no matter the amount of data available. These parameters are of a fixed size, which means that the model already knows the number of parameters it requires. The model's complete information is represented within its parameters. The only information needed to predict future or unknown values from the current value is the parameters [1, 5].

   Examples: A simple example of a parametric model is the family of Gaussian distributions parametrized by $\theta = [\mu, \sigma]$ where $\mu \in \mathbb{R}$ denotes the mean value also called the location parameter, and $\sigma > 0$ is the standard deviation also known as the scale parameter. Any Gaussian distribution can be completely described by just these two parameters. So the knowledge of μ and σ are sufficient to know everything about the statistical model

   $$\mathcal{P} = \left\{ P_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu)^2}{2\sigma^2} \right) \middle| \mu \in \mathbb{R}, \sigma > 0 \right\}. \tag{1.1}$$

   Other popular noteworthy examples of a parametric model include linear regression model [6], logistic regression [7], linear support vector machine [8].

2. **Non-parametric:** In non-parametric models, the data distribution cannot be defined in terms of a finite set of parameters. Instead, the parameters are

often defined by assuming an infinite dimensional[1] θ. Typically, θ is thought of as a function. The amount of information that θ can capture about the data can grow as the amount of data grows. This makes them more flexible. Here, the structure of the model is not fixed, but very often grows in size to accommodate the complexity of the data [9]. Well-known methods for non-parametric models are Decision Trees [10], K-Nearest Neighbor [11], and Support Vector Machines with Gaussian Kernels [12].

3. **Semi-parametric:** A semi-parametric model contains both finite and infinite dimensional parameters, i.e., it has parametric and non-parametric components [13]. However, the estimation of the finite-dimensional parametric component is of more interest and the non-parametric component is treated as a nuisance parameter.

1.3 Why Statistical Model

In the previous section, we described what is a statistical model and the types of models that are being used for modeling the data. In this section, we provide a discussion emphasizing the need for such models.

Models play a central role in the field of data analysis. Several inferences, including prediction, control, information extraction, knowledge discovery, validation, risk assessment, and decision making, may be made after a model has been established. Therefore, constructing and developing appropriate models is essential for solving difficult real-world problems. Below we highlight some of the key requirements of statistical models as mentioned in [1].

1.3.1 Representation of Stochastic Structures

The primary need of a statistical model is to approximate the unknown true distribution of probabilistic events. As shown in Fig. 1.1, the observed data is utilized to estimate a statistical model, F, that closely mimics the true distribution G. However, model mismatches are bound to occur in the modeling process. Hence, F is never exactly identical to G. This is rightly stated by Burnham & Anderson, "A model is a simplification or approximation of reality and hence will not reflect all of reality" [14].

1.3.2 Predictions Purposes

A major requirement in any field employing advanced data analytics is making predictions about future outcomes using historical and current data. In the present era, organizations use predictive analytics to find patterns, behaviour, and trends

[1] Infinite-dimensional linear space means a space that cannot be spanned by any finite set of elements in the set. An example of an infinite-dimensional linear space is the space of continuous functions defined on the real line.

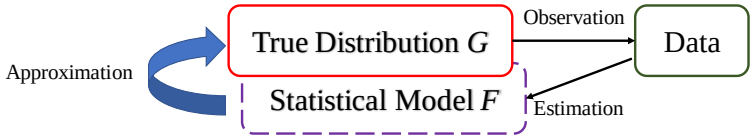


Figure 1.1: Estimating true distributions using statistical model (Fig. 1.1 of [1]).

in the data to identify risks and opportunities. In this regard, statistical models can be used to predict data as accurately as possible. Making predictions is one of the most significant roles of statistical models. Fig. 1.2 highlights the predictive mode of statistical models.

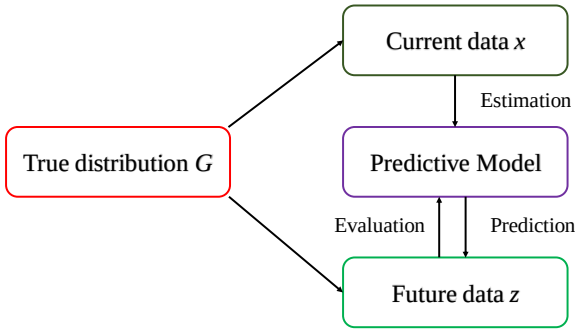


Figure 1.2: Statistical models for predictive analysis (Fig. 1.2 of [1]).

1.4 Model Selection

In the previous section, we presented the definition of a statistical model and discussed the need for such models. However, while modeling stochastic structures there may not be any unique form that can be determined deterministically. In fact, since the true model is unknown, the stochastic structures can be modeled using a variety of forms that differ from each other. This gives rise to the problem of model selection, i.e., how do we select or estimate a model from a set of candidate models given a set of data? Or in other words, how do we analyze the goodness of a model given the observed data? However, since the scope of this thesis encompasses linear regression models, hence we explicitly explore the methods of statistical model selection particularly used in linear regression analysis. This is specifically because linear regression models are perhaps the most popular and widely used models for inference and predictions. Furthermore, their simplicity,

ease of use, and mathematical tractability make them the most desired parametric models in various fields of science, engineering, business, environmental studies, and many other domains [15].

1.5 Thesis Outline and Contributions

Chapter 2: Model Selection - A Brief Overview

In this chapter, we start with a brief explanation of the linear regression architecture and discuss its low and high-dimensional (HD) scenarios with some examples. We highlight the challenges in the HD setting and present two popular predictor/subset selection algorithms that are widely used for selecting significant variables in a linear regression. The model selection problem in linear regression is formally established. It is followed by a detailed literature review of the existing different statistical model selection methods. Furthermore, we also discuss the drawbacks of classical model selection methods when dealing with HD data, where the number of available measurements is quite small compared to the parameter dimension.

Chapter 3: Multi-Beta-Test

In this chapter, we introduce a novel method for model selection method called multi-beta-test (MBT) based on the hypothesis testing framework. MBT is specifically designed to perform model selection in high-dimensional linear regression that employs a greedy predictor selection algorithm for picking the set of the most probable candidate models. The candidate models chosen in this fashion should possess a nested structure such that any smaller model is a subset of a bigger model. This nested design is required for MBT to operate. In this regard, the orthogonal matching pursuit is utilized for selecting the initial set of candidate models with models starting with dimension one to a maximum dimension of K. Eventually, MBT is used to estimate the true model. Simulation results have shown that MBT does quite well in estimating the true model. However, it is sensitive to a tuning parameter, which needs to be selected with care.

This chapter is based on the following published paper.

- Prakash B. Gohain, and Magnus Jansson. "Relative cost based model selection for sparse high-dimensional linear regression models." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

Chapter 4: Bayesian Information Criterion - Robust

In this chapter, we re-investigate one of the most popular model selection criteria, the Bayesian information criterion (BIC), and its high signal-to-noise-ratio (SNR) forms proposed in [16]. These high-SNR forms of the BIC suffer from a data-scaling problem. This data-scaling problem is a byproduct of the data-dependent penalty design, which generates irregular penalties when the data is scaled and often leads to

greater underfitting and overfitting losses in some scenarios when the noise variance is too small or large respectively. We discuss this problem in detail. Furthermore, to alleviate this problem, we present a new form of the BIC, called BIC-Robust or BIC_R in short. BIC_R is invariant to data-scaling and we provide analytical proofs to show that BIC_R is a consistent criterion, i.e., it selects the true model as the sample size grows large and/or when the SNR increases.

This chapter is based on the following published paper.

- Prakash B. Gohain, and Magnus Jansson. "Scale-invariant and consistent Bayesian information criterion for order selection in linear regression models." Signal Processing 196 (2022): 108499.

Chapter 5: Extended Bayesian Information Criterion - Robust

In the high-dimensional setting, in which the number of available measurements is quite small compared to the parameter dimension, the classical methods of model selection including BIC_R underperform and fail to achieve consistency, especially in cases when the parameter dimension grows with the sample size. In this regard, extended BIC (EBIC) [17], which is an extended version of the original BIC, and extended Fisher information criterion (EFIC) [18], which is a combination of EBIC and Fisher information criterion, were proposed. Both EBIC and EFIC are consistent estimators of the true model as the number of measurements grows very large. However, EBIC is not consistent in high-SNR scenarios where the sample size is fixed and EFIC is not invariant to data-scaling resulting in unstable behaviour. In this chapter, we propose a new form of the EBIC criterion called EBIC-Robust or EBIC_R in short, which is invariant to data-scaling and consistent in both large sample size and high-SNR scenarios. Analytical proofs are presented to guarantee its consistency.

This chapter is based on the following publications.

- Prakash B. Gohain, and Magnus Jansson. "New Improved Criterion for Model Selection in Sparse High-Dimensional Linear Regression Models." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

- Prakash B. Gohain, and Magnus Jansson. "Robust Information Criterion for Model Selection in Sparse High-Dimensional Linear Regression Models." *Submitted to IEEE Transactions on Signal Processing.* Preprint available at arXiv preprint arXiv:2206.08731 (2022).

Chapter 6: Model Selection in Block-Sparse Linear Regression Models

In this chapter, we discuss model selection in block-sparse general linear regression models. This is also termed the block multiple measurement vector (BMMV) regression model. The BMMV model is the most general form of a linear regression model. The different variations of BMMV include (i) single measurement vector

(SMV) (ii) multiple measurement vector (MMV) (iii) block single measurement vector (BSMV). We investigate model selection only in BMMV model since the method can be easily adapted to all other forms of linear regression models. EBIC_{R} proposed in Chapter 5 is modified or generalized to perform model selection in sparse BMMV regression models. In this regard, we provide the necessary steps to show how this can be achieved. Simulation results are provided to analyze the behaviour of the methods in this setting and compare the performance with state-of-the-art methods for model selection in the BMMV scenario.

This chapter is based on the following publications.

- Prakash B. Gohain, and Magnus Jansson. "Model Selection in High-Dimensional Block-Sparse General Linear Regression Models." *Submitted to IEEE Signal Processing Letters.* Preprint available at arXiv preprint arXiv:2209.01460

Chapter 7: Discussion and Conclusion

In the final chapter, we provide a comprehensive summary of the thesis, discussing the proposed methods their advantages, and some disadvantages. Furthermore, we highlight some of the potential future research problems in this context of model selection.

Copyright Notice

Chapter 2

Background

"A model should be as simple as it can be but no simpler."
—*Albert Einstein (1879–1955)*

In the previous chapter we motivated the idea of model selection in data analysis and described briefly statistical models and their purpose. We further mention that this thesis focuses on model selection in linear regression due to its popularity and ubiquitousness. In this chapter, we provide the necessary background and a survey of different model selection methods.

2.1 Linear Regression

Generally speaking, regression analysis is a statistical technique for investigating and modeling the relationship between a dependent variable (or response variable) and a set of independent variables (or predictor variables) [19]. In the case of a linear regression model, this relationship between the dependent and independent variables is modeled using a linear approach whose unknown model parameters are estimated from the data. The literature on linear regression is quite extensive [6, 19, 20]. Consider the set of observations $\{y_i\}_{i=1}^{N}$ where N is the sample size. For the ith observation y_i, let the associated predictor variables be $[a_{i1}, \ldots, a_{ip}]$. The linear model that maps the relationship between the response y_i and the predictor variables $\{a_{ij}\}_{j=1}^{p}$ is given as [6]

$$y_i = x_1 a_{i1} + \ldots + x_p a_{ip} + e_i, \tag{2.1}$$

for $i = 1, \ldots, N$. Here, e_i is the error term that models the misfit. Traditionally it is often assumed that $e_i \sim \mathcal{N}(0, \sigma^2)$ where σ^2 is the true noise variance. In the matrix form, we can write

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{2.2}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the observation (also called measurement or response) vector and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p] \in \mathbb{R}^{N \times p}$ is termed as the design matrix and $\mathbf{a}_j = [a_{1j}, \ldots, a_{Nj}]^T \in$

$\mathbb{R}^{N\times 1}$ and $j = 1, \ldots, p$. We now consider $N \gg p$, which is the low-dimensional scenario. The $p \gg N$, which is the high-dimensional setting will be discussed later. $\mathbf{x} \in \mathbb{R}^p$ is the unknown parameter (or regression coefficient) vector. $\mathbf{e} \in \mathbb{R}^N$ is the associated noise vector whose elements are assumed to be i.i.d. following a Gaussian distribution, i.e., $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. The fundamental problem in linear regression is estimating $\mathbf{x}$ given $\mathbf{y}$ and $\mathbf{A}$. The classical solution of $\mathbf{x}$ is obtained using the method of least-squares that minimizes the error between the observed and the estimated response

$$\hat{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2. \tag{2.3}$$

The least squares estimate is equivalent to the maximum likelihood (ML) estimate and it is the optimal unique solution of $\mathbf{x}$ assuming $\mathbf{A}$ is full ranked. The closed form solution is [20]

$$\hat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}. \tag{2.4}$$

Note that a mandatory requirement for evaluating the solution in (2.4) is that $\mathbf{A}^T\mathbf{A}$ should be full rank such that its inverse exists. Henceforth, the fitted response vector is

$$\hat{\mathbf{y}} = \mathbf{A}\hat{\mathbf{x}} = \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y} = \mathbf{\Pi}\mathbf{y} \tag{2.5}$$

where $\mathbf{\Pi}$ is the orthogonal projection matrix. It projects $\mathbf{y}$ onto the $span(\mathbf{A})$. In many real-world scenarios, not all the predictor variables are significant. Perhaps out of the p variables say only k_0 of them actually contribute substantially in generating the data. In this case, model or variable selection entails finding the most significant set of predictor variables out of the available p variables. In a more classical setting, if the models are nested such that a model with k parameters is a subset of the model with $k + 1$ parameters, then the model selection problem is also termed as the order selection problem.

2.1.1 High-dimensional Scenario

The high-dimensional case arises when $p \gg N$. Many real applications exhibit this scenario such as in genome-wide association studies [21], high-resolution magnetic resonance imaging [22], and high-resolution radar systems [23]. For example, in genome-wide studies, one measures micro-array datasets built from a large amount of profile genes expression. Here, y_i denotes the expression level of one gene on the ith sample and $[a_{i1}, \ldots, a_{ip}]$ the biological signals (DNA micro-arrays). Typically the number of available samples N is in the order of hundreds while the number of genes (predictors) p is in the order of thousands. From a statistical model selection point of view, it is desired to estimate the most important variables out of the p variables. Thus, the primary objective is to select the significant components among the available genes in order to establish a meaningful relationship between DNA and the gene expression level.

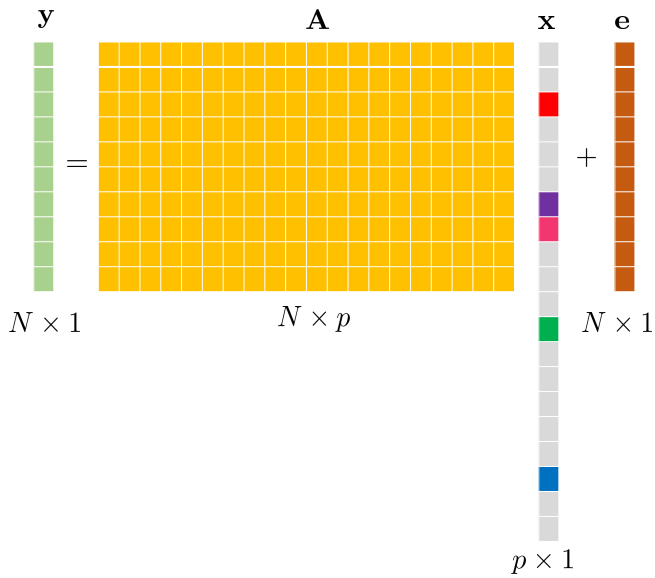


Figure 2.1: Sparse high-dimensional linear regression illustration.

The high-dimensional case is an underdetermined problem, where it is quite possible that $\mathbf{A}$ contains linearly dependent columns given that p is so large. In this case, the Gram matrix $\mathbf{A}^T\mathbf{A}$ is ill-conditioned such that most of the eigenvalues are zero and hence the Gram matrix is not invertible. As such, solving for $\mathbf{x}$ using the least-squares solution (2.4) is not feasible. To circumvent this problem a widely used assumption is considering the parameter vector $\mathbf{x}$ to be sparse. By sparse $\mathbf{x}$ it means that only a few of the elements of $\mathbf{x}$ are non-zero and the majority are zero elements. This translates to that only a very few of the predictors out of p are relevant to the data. This concept of sparsity was introduced in the compressed sensing literature, where solving inverse problems in underdetermined scenarios are frequently encountered. The sparsity concept is motivated by the fact most of the high-dimensional data that occur naturally are genuinely sparse in nature, i.e., a small subset of the signals can be used to effectively represent the data while ignoring the rest. It is quite common nowadays to frequently encounter high-dimensional datasets where the number of available measurements N is quite small compared to the number of features (or parameters). However, in many practical scenarios, just the sheer number of features does not imply that all of them are significant. In fact, it has been observed in many cases that just a few of the parameters are important to describe the data and perform good predictions. Hence, the sparsity assumption does play an important role in simplifying the subset-selection problem in linear regression without being invalid. Fig. 2.1 presents

an illustration of the sparse high-dimensional linear model structure. The non-zero elements of the parameter vector $\mathbf{x}$ are marked as colored squares and the gray squares indicate the zero elements. This provides an idea of the sparse nature of the parameter vector with very few non-zero elements that actually contribute to the observed data.

Model Selection Problem: For the linear regression model in (2.2) we can formally define the model selection problem in general. Whether it be a low-dimensional setting or a high-dimensional setting the goal of the model selection problem remains the same in both cases. Let us denote $\mathcal{S}$ as the true support of $\mathbf{x}$, i.e., $\mathcal{S} = \{i : x_i \neq 0\}$ having cardinality $card(\mathcal{S}) = k_0 \leq p$ and $\mathbf{A}_{\mathcal{S}}$ as the set of columns of $\mathbf{A}$ corresponding to the support $\mathcal{S}$. In the sparse high-dimensional setting we have $p \gg N$ and $k_0 \ll p$. The goal of model selection (also called best subset selection) is estimating the unknown true support $\mathcal{S}$ given $\mathbf{y}$ and $\mathbf{A}$.

2.2 Predictor Selection Algorithms

To perform model selection in linear regression, we need to first have at our disposal a set of candidate models, whence we can choose the best model using a suitable method or criterion. Now observe that, for the linear model in 2.2 with p number of parameters, if we take the combinatorial approach the total number of possible candidate models is $2^p - 1$. A naive approach is evaluating each model using a model selection criterion to find the true model among all $2^p - 1$ available candidates. However, it is quite obvious that as p grows large, the candidate model space grows exponentially. In such a situation, testing model by model is infeasible and in fact impractical. To address this problem a viable approach is to perform a pre-screening by employing a certain algorithm that can provide us with a set of most important candidate models. We refer to such algorithms that provide us with an initial set of candidate models as predictor selection algorithms. Below we discuss two types of algorithms widely used to find sparse solutions in high-dimensional settings.

2.2.1 Greedy Methods

A greedy method operates by favoring a sequence of locally optimal variables thus dumping the need for an exhaustive search over the entire parameter space. A typical greedy method is an iterative process that starts from an initially empty predictor index set and at each step, expands that set by one additional column index. The column selected at each stage maximally lowers the residual l_2 error in predicting the data $\mathbf{y}$ from the currently active columns in the index set. The algorithm stops when the stopping criterion is fulfilled which is necessary in order to avoid selecting all the predictors. Popular methods in this genre include matching pursuit [24], basis pursuit [25], orthogonal matching pursuit (OMP) [26]. Here, we focus on OMP for our application which is discussed in detail below.

OMP is a widely used greedy algorithm for recovering sparse signals in an underdetermined system of linear equations [27]. Here, we specifically consider the

Algorithm 2.1 OMP

1: **Inputs:** Design matrix $\mathbf{A}$, observation vector $\mathbf{y}$.
2: **Initialization:** $\mathbf{r}^0 = \mathbf{y}, \mathcal{S}^0_{\text{OMP}} = \emptyset$, $i = 1$
3: **repeat**
4: Find next column index $d^i = \arg\max_j \left|\mathbf{a}_j^T \mathbf{r}^{i-1}\right|$
5: Add current index: $\mathcal{S}^i_{\text{OMP}} = \mathcal{S}^{i-1}_{\text{OMP}} \cup \{d^i\}$
6: Update residual: $\mathbf{r}^i = \left(\mathbf{I}_n - \mathbf{\Pi}_{\mathcal{S}^i_{\text{OMP}}}\right)\mathbf{y}$
7: Increment counter: $i = i + 1$
8: **until** the stopping condition is achieved
9: **Outputs:** True support estimate $\hat{\mathcal{S}}_0 = \mathcal{S}^i_{\text{OMP}}$.

recovery of the k_0-sparse vector $\mathbf{x}$ in the linear model 2.2. Given the design matrix $\mathbf{A} \in \mathbb{R}^{N \times p}$, the OMP (Algorithm 1) iteratively selects the optimal columns one by one until the stopping criterion is fulfilled. A required step in this regard is to first normalize all the columns of $\mathbf{A}$ to be unit norm, i.e., $\|\mathbf{a}_i\|_2 = 1,\ \forall\ i = 1, 2, \ldots, p$. OMP starts by initializing the residual vector $\mathbf{r}^0 = \mathbf{y}$, setting the counter $i = 1$ and the initial empty index set $\mathcal{S}^0_{\text{OMP}} = \emptyset$. A column's index is selected if that column has the maximum absolute correlation value with the residual vector $\mathbf{r}^{i-1}$, i.e., $d^i = \arg\max_j |\mathbf{a}_j^T \mathbf{r}^{i-1}|$ where d^i denotes the column index at i^{th} iteration and $\mathbf{a}_j$ represents the j^{th} column of $\mathbf{A}$. $\mathbf{A}_{\mathcal{S}^i_{\text{OMP}}}$ denotes the sub-matrix of $\mathbf{A}$ formed using the columns indexed by $\mathcal{S}^i_{\text{OMP}}$. Next, based on the current support $\mathcal{S}^i_{\text{OMP}}$, the least-squares estimate of $\mathbf{y}$, i.e., $\mathbf{\Pi}_{\mathcal{S}^i_{\text{OMP}}}\mathbf{y}$ is evaluated where $\mathbf{\Pi}_{\mathcal{S}^i_{\text{OMP}}} = \mathbf{A}_{\mathcal{S}^i_{\text{OMP}}}\mathbf{A}^{\dagger}_{\mathcal{S}^i_{\text{OMP}}}$ denotes the projection matrix onto the $span(\mathbf{A}_{\mathcal{S}^i_{\text{OMP}}})$ and $\mathbf{A}^{\dagger}_{\mathcal{S}^i_{\text{OMP}}} = \left(\mathbf{A}^T_{\mathcal{S}^i_{\text{OMP}}}\mathbf{A}_{\mathcal{S}^i_{\text{OMP}}}\right)^{-1}\mathbf{A}^T_{\mathcal{S}^i_{\text{OMP}}}$ is the Moore-Penrose pseudo inverse of $\mathbf{A}_{\mathcal{S}^i_{\text{OMP}}}$. The estimate $\mathbf{\Pi}_{\mathcal{S}^i_{\text{OMP}}}\mathbf{y}$ is used to update the residual $\mathbf{r}^i = \mathbf{y} - \mathbf{\Pi}_{\mathcal{S}^i_{\text{OMP}}}\mathbf{y} = \left(\mathbf{I}_N - \mathbf{\Pi}_{\mathcal{S}^i_{\text{OMP}}}\right)\mathbf{y}$. The selection process is halted when the stopping criterion is achieved or the number of regressors to be chosen is fixed *a priori*.

2.2.2 Shrinkage Methods

The least-squares solution is too liberal and does not impose any constraint on the value of the parameter coefficients. As such the solution of least-squares produces an all non-zero $\mathbf{x}$ vector. In the large-p, small-N case, the least-squares solution is infeasible since this is an underdetermined problem and the gram matrix $\mathbf{A}^T\mathbf{A}$ is most likely singular and hence not invertible. A popular way to solve for $\mathbf{x}$ in the underdetermined situation is using penalized regression methods that combines the least-squares loss with a constraint or bound on the sum of the absolute values of the coefficients. They are often termed as shrinkage techniques since they shrink

the estimates of the parameters towards zero. The typical Lagrangian form is given as [8]

$$\hat{\mathbf{x}}(\lambda) = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \frac{1}{2N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_q^q \right\}, \tag{2.6}$$

where $\lambda \geq 0$ is a regularization (or tuning) parameter that controls the level of the penalty or in other words it is a way of balancing the goodness of fit and shrinking the coefficients. If we choose $\lambda = 0$, we end up with the usual least-squares solution, while for $\lambda = \infty$ gives an all zero $\mathbf{x}$ vector. For other values of λ within the range $0 < \lambda < \infty$ it produces different estimates of $\mathbf{x}$. Generally, as we move from $\lambda = \infty$ towards $\lambda = 0$, the number of non-zero coefficients in $\hat{\mathbf{x}}$ increases. Depending on the choice of $0 \leq q \leq 2$ we arrive at different versions of the penalized regression. The penalized regression with $p = 2$ is called ridge regression. If we are looking for a sparser solution ridge is not a good option since even though it tries to shrink the coefficients to zero, they are never exactly zero unless of course for $\lambda = \infty$ when all components are exactly zero. Thus, obtaining a much sparser solution of $\mathbf{x}$ is not feasible using ridge regression. The alternative to this problem is setting $q = 1$, which leads to the well-known LASSO estimator [8, 28, 29]. LASSO stands for Least Absolute Shrinkage and Selection Operator. Unlike ridge regression, the beauty of LASSO is that it produces sparse solutions with exactly zero components. Luckily the LASSO problem is convex w.r.t. $\mathbf{x}$ and hence can be solved. Efficient algorithms to solve the LASSO for large-p do exist. A popular way of finding the LASSO solution is using the algorithm called modified least angle regression (LARS) [30] since it also provides the required sequence of regularization parameters for which the support changes. Thus, one can obtain a possible set of candidate linear regression models using the LARS algorithms where each candidate model corresponds to a unique value of the regularization parameter λ. The goal then is to find the best model using a suitable model selection method.

2.2.3 Support Recovery conditions for OMP

In order for OMP to correctly recover the true signal support, certain conditions need to be fulfilled. These are referred to as support recovery conditions. Most of these conditions are based on specific features of the design matrix $\mathbf{A}$. In the compressed sensing literature, different aspects of the design matrix have been proposed to analyze the support recovery performance such as Mutual Coherence [31], and Restricted Isometry Property (RIP) [32]. We first formally define both these terms and then describe how they are used in the case of OMP to express the support recovery guarantees.

Mutual Coherence: Assuming that the columns of the design matrix are normalized, i.e., $\mathbf{a}_i^T \mathbf{a}_i = 1$ the mutual coherence is then defined by [33]

$$M = \max_{1\leq i\neq j\leq p} \left| \mathbf{a}_i^T \mathbf{a}_j \right|. \tag{2.7}$$

Thus, mutual coherence is a measure of the maximum cross correlation between the columns of a matrix.

Restricted Isometry Property (RIP): A matrix $\mathbf{A}$ satisfies the RIP of order k_0 if there exists a constant $\delta_{k_0} \in [0, 1)$ such that

$$(1 - \delta_{k_0})\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_{k_0})\|\mathbf{x}\|_2^2 \tag{2.8}$$

holds for all $\mathbf{x} \in \mathbb{R}^p$ with $card(supp(\mathbf{x})) \leq k_0$. In particular, the smallest constant δ_{k_0} satisfying (2.8) is called the restricted isometry constant (RIC). In general, a small δ_k indicates that any k collection of columns of the matrix $\mathbf{A}$ is approximately orthonormal.

In the noise-free case (i.e., $\mathbf{e} = \mathbf{0}$) if $\delta_{k_0} < \frac{1}{\sqrt{k_0+1}}$ then OMP is guaranteed to exactly recover k_0-sparse signal $\mathbf{x}$ in exactly k_0 iterations [34,35]. In the noisy case, the latest result on the sufficient recovery condition is given in Theorem 1 of [36] which we present as a theorem below:

Theorem 2.1 *Under $\|\mathbf{e}\|_2^2 \leq \epsilon$, suppose that $\mathbf{A}$ satisfies the RIP of order $k_0 + 1$ with $\delta_{k_0+1} < \frac{1}{\sqrt{k_0+1}}$. Then OMP with the stopping criterion $\|\mathbf{r}^k\|_2^2 \leq \epsilon$ can exactly recover $supp(\mathbf{x})$ in k_0 iterations provided that*

$$\min_{i \in supp(\mathbf{x})} |x_i| > \frac{\epsilon}{\sqrt{1 - \delta_{k_0+1}}} + \frac{\sqrt{1 + \delta_{k_0+1}}\epsilon}{1 - \sqrt{k_0 + 1}\delta_{k_0+1}}. \tag{2.9}$$

Next, we present some support recovery guarantees of OMP based on the mutual coherence property. This is a slightly easier alternative to gauge the suitability of $\mathbf{A}$ as compared to the RIP condition since it does not require an exhaustive search over a collection of subsets. In the noiseless case, $M < \frac{1}{2k_0-1}$ is a sufficient condition for exactly recovering a k_0-sparse vector. In the bounded Gaussian noise case, Theorem 7 of [37] provides an insight to the recovery performance of OMP. This is stated below.

Theorem 2.2 *Suppose $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$, $M < \frac{1}{2k_0-1}$ and that the non-zero coefficients of $\mathbf{x}$ fulfill*

$$|x_i| \geq \frac{2\sigma\sqrt{N + 2\sqrt{N \log N}}}{1 - (2k_0 - 1)M}. \tag{2.10}$$

Then OMP with the stopping rule $\|\mathbf{r}^i\|_2 \leq \sigma\sqrt{N + 2\sqrt{N \log N}}$ selects the true support set $\mathcal{S}$ with probability at least $1 - 1/N$.

In a practical scenario, typically both the noise variance σ^2 and true sparsity k_0 are unknown quantities. As such, implementing OMP with $\|\mathbf{r}^k\|_2^2 \leq f(\sigma^2)$ where $f(\sigma^2)$ is some function of σ^2 or with k_0 iterations as the stopping criterion is not feasible for correct subset selection in the sparse high-dimensional setting. This motivates the need for sophisticated and practical methods for model selection in the absence of knowledge about k_0 and σ^2.

2.2.4 Support Recovery Guarantees In LASSO

Similar to OMP, there are certain conditions that need to hold in order for LASSO to correctly recover the true sparse solution in the high-dimensional setting. In the literature, following recovery guarantees of LASSO, several conditions have been proposed such as the Irrepresentable condition [38], restricted eigenvalue condition, Restricted Isometry Property, and Sparse Riesz condition [39]. Here, we briefly discuss only the Irrepresentable condition for brevity.

Irrepresentable Condition: The authors in [38] show that for LASSO to select the true support set both in the classical fixed p setting and in the large p setting as the sample size N grows large, the Irrepresentable Condition is necessary, and sufficient. Let $C_{11} = N^{-1}\mathbf{A}_{\mathcal{S}}^T\mathbf{A}_{\mathcal{S}}$ and $C_{21} = N^{-1}\mathbf{A}_{\mathcal{N}}^T\mathbf{A}_{\mathcal{S}}$ where $\mathcal{S}$ is the true support set with $card(supp(\mathcal{S})) = k_0$ and $\mathcal{N} = \{1, \ldots, p\} \setminus \mathcal{S}$ with $card(supp(\mathcal{N})) = p - k_0$. The matrix $\mathbf{A}$ satisfies the Irrepresentable Condition if

$$\left|C_{21}(C_{11})^{-1} sign(\mathbf{x}_{\mathcal{S}})\right| < \mathbf{1} \tag{2.11}$$

where $\mathbf{1}$ is a $(p - k_0) \times 1$ vector of ones and the inequality holds element-wise. The $sign(\beta)$ is 1 for $\beta > 0$, -1 for $\beta < 0$ and 0 if $\beta = 0$. Note that the irrepresentable condition requires that C_{11} is invertible.

2.3 Model Selection Methods

Here we present a brief survey of the existing model selection methods. We discuss three popular approaches to model selection widely used in the literature, i.e., methods based on hypothesis testing, information criteria, and cross-validation.

2.3.1 Hypothesis Testing

Model selection in linear regression can be performed using hypothesis testing if there is a nested order in the candidate models. By nested we mean that any smaller model is a subset of a bigger model. In this regard, we can reformulate the linear model in (2.2) for the hypothesis testing problem as follows

$$\mathcal{H}_k : \mathbf{y} = \mathbf{A}_k\mathbf{x}_k + \mathbf{e}_k. \tag{2.12}$$

Here, $\mathcal{H}_k$ denotes the hypothesis that the data $\mathbf{y} \in \mathbb{R}^{N\times 1}$ is generated according to the candidate model with k parameters. $\mathbf{A}_k \in \mathbb{R}^{N\times k}$ is a sub-design matrix formed using the first k columns of the full design matrix $\mathbf{A} \in \mathbb{R}^{N\times p}$ where $k = 1, \ldots, p$. $\mathbf{x}_k \in \mathbb{R}^{k\times 1}$ is the corresponding unknown regression coefficient vector. $\mathbf{e}_k \in \mathbb{R}^{N\times 1}$ is the associated noise vector, such that $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2\mathbf{I}_N)$, where σ_k^2 is the unknown noise variance corresponding to hypothesis $\mathcal{H}_k$.

A sequence of hypotheses, $\mathcal{H}_k$ holds true versus $\mathcal{H}_{k+1}$ holds true $k = 1, 2, ...,$ can be tested sequentially. In this case, $\mathcal{H}_k$ is the null hypothesis and $\mathcal{H}_{k+1}$ is the

alternate hypothesis. In hypothesis testing based model selection, a test statistic is involved that is evaluated using the data available. Typically, based on the assumed statistical properties of the random quantity, the test statistic would follow a certain distribution under the null hypothesis. To perform the test itself and make a decision, a threshold is required, which is obtained based on the significance level chosen by the user. To decide what hypothesis is true, the test statistic is compared with the threshold. As long as the alternate hypothesis $\mathcal{H}_{k+1}$ is true, the test continues to progress ahead starting from $k = 1, 2, \ldots$, and so on. Once the null hypothesis $\mathcal{H}_k$ is accepted, the test procedure stops and the model $\mathcal{H}_k$ is selected.

A classical and widely used approach for accessing the goodness of fit of competing parametric statistical models is the generalized likelihood ratio test (GLRT). Consider the following hypotheses given observed data $\mathbf{y}$

$$\begin{aligned} \mathcal{H}_m &: \mathcal{L}\left(\theta_m|\mathbf{y}\right) \\ \mathcal{H}_q &: \mathcal{L}\left(\theta_q|\mathbf{y}\right), \end{aligned} \tag{2.13}$$

where $\mathcal{L}\left(\theta_m|\mathbf{y}\right)$ and $\mathcal{L}\left(\theta_q|\mathbf{y}\right)$ are the likelihood functions under hypothesis $\mathcal{H}_m$ and $\mathcal{H}_q$, respectively, and θ_m, θ_q being their corresponding unknown parameters. Note that the subscript denotes the dimension or order of the model with $m < q$. Then the GLRT computes the following test statistic

$$\Lambda = \frac{\mathcal{L}\left(\hat{\theta}_m|\mathbf{y}\right)}{\mathcal{L}\left(\hat{\theta}_q|\mathbf{y}\right)} \tag{2.14}$$

where $\hat{\theta}_m$ and $\hat{\theta}_q$ are maximum likelihood estimates of θ_m and θ_q, respectively evaluated as follows

$$\hat{\theta}_i = \arg\max_{\theta_i} \left\{\mathcal{L}(\theta_i|\mathbf{y})\right\}. \tag{2.15}$$

To decide between the two models with orders m and q, is performed as follows

$$\Lambda \underset{\mathcal{H}_q}{\overset{\mathcal{H}_m}{\lessgtr}} \eta \tag{2.16}$$

where η is a pre-specified threshold chosen for a particular probability of false positive. GLRT can be used to perform model (order) selection in the linear regression problem (2.12). Under the assumption of the i.i.d Gaussian noise elements $e_i \sim \mathcal{N}(0, \sigma^2)$, the likelihood ratio given by 2.14 boils down to the following test statistic

$$F = \frac{\left(\frac{\mathrm{RSS}_m - \mathrm{RSS}_q}{q-m}\right)}{\left(\frac{\mathrm{RSS}_q}{N-q}\right)} = \frac{\frac{X}{d1}}{\frac{Y}{d2}} (\text{say}), \tag{2.17}$$

where RSS stands for residual sum of squares and is computed as follows

$$\mathrm{RSS} = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \tag{2.18}$$

$\hat{\mathbf{y}}$ is the predicted measurement. RSS_m and RSS_q denotes the residual sum of squares of model $\mathcal{H}_m$ and $\mathcal{H}_q$, respectively. Under the null hypothesis that model $\mathcal{H}_q$ is unable to provide a remarkably better fit than model $\mathcal{H}_m$, and the i.i.d. Gaussian assumption of e_i, the X and Y terms in 2.17 are independent Chi-squared distributed random variables with d_1 and d_2 degrees of freedom, respectively, where $d_1 = q - m$ and $d_2 = N - q$. Now from probability theory, the ratio $(X/d_1)/(Y/d_2)$ follows a F-distribution with d_1 and d_2 degrees of freedom. Technically, since $q > m$, as such $\mathcal{H}_q$ will always give a better fit to the data and lower fitting error as compared to model $\mathcal{H}_m$ (or any other model with fewer parameters). The purpose of the F-test is to determine if the added parameters have a significant contribution to the overall fit. The null hypothesis (in this case $\mathcal{H}_m$) is rejected if the F value calculated is greater than the critical value of the F-distribution for some chosen false-rejection probability (Typical values are 0.05 or 0.1).

Another common way to represent the GLRT for the linear regression is by reformulating the test statistic using the log function in the following manner

$$\Lambda_{\text{LLR}} = -2 \ln \left[\frac{\mathcal{L}\left(\hat{\theta}_m | \mathbf{y}\right)}{\mathcal{L}\left(\hat{\theta}_q | \mathbf{y}\right)} \right] = 2 \ln \mathcal{L}\left(\hat{\theta}_q | \mathbf{y}\right) - 2 \ln \mathcal{L}\left(\hat{\theta}_m | \mathbf{y}\right), \tag{2.19}$$

where the subscript LLR stands for log-likelihood ratio. Under the null hypothesis where $q > m \geq k_0$ and bearing the i.i.d Gaussian noise assumption, the test statistic Λ_{LLR} asymptotically follows a central chi-squared distribution with $q - m$ degrees of freedom, i.e., $\Lambda_{\text{LLR}} \sim \chi^2_{(q-m)}$ [40]. Thus, we can decide between two models with orders m and q, respectively using a similar approach as in 2.16 where the threshold is set based on the χ^2_{q-m} distribution for some chosen probability of false selection.

Popular methods for implementing model selection in linear regression using sequential hypothesis testing include forward selection, backward elimination, and stepwise regression.

Forward Selection is often used to provide an initial screening of the candidate variables when a large group of variables exists. For example, suppose we have a hundred or more variables to choose from, this is way beyond the domain of all possible regression procedures. A reasonable strategy would be to use the forward selection procedure to obtain the best ten to fifteen variables in decreasing order of significance and then apply the existing model selection algorithm to the variables in this subset. The method begins with no candidate variables in the model. A variable is selected that has the highest statistically significant improvement of the fit. In this context, quantities like R-Squared or p-values (e.g., based on the F statistic) could be used to measure the statistical significance of the variables. A variable with a higher R-squared value or equivalently smaller p-value is statistically more significant. This process is repeated and at each step, a variable is added that increases the statistical significance the most. The process is stopped when none of the remaining variables improves the model to a statistically significant extent. An

important aspect of this method is that once a variable enters the model, it cannot be deleted [41].

Backward Selection is technically the reverse of the forward selection. It starts with all candidate variables already in the model. At each step, the variable that is the least significant statistically is removed. This process continues until no non-significant variables remain. The user sets the significance level at which variables can be removed from the model. A variable once removed cannot be added again to the model [41].

Stepwise regression is a mix of forward and backward selection approaches. It is a variant of the forward selection in that all candidate variables in the model are examined to see whether their significance has decreased below the predetermined tolerance threshold after each stage in which a variable was added. If a non-significant variable is discovered, it is eliminated from the model. Stepwise regression needs two significance levels: one for adding variables and one for deleting variables. In order to prevent stepwise regression from entering an infinite loop, the cutoff probability for adding variables should be lower than the cutoff probability for deleting variables [42].

Classical statistical hypothesis testing for model selection in linear regression works well when the models have a nested structure. However, in many scenarios, the nested assumption may not hold. In this case, if we want to employ a hypothesis testing based method for model selection, the first step is obtaining a nested sequence of models such that we prefer model $\mathcal{H}_i$ over model $\mathcal{H}_{i+1}$. This can be accomplished by using predictor selection algorithms that generate a monotonic sequence of predictor indices such as OMP. However, using greedy methods (e.g. OMP) to obtain this sequence of models causes the test statistic under the null hypothesis to deviate from the known distribution. As such, we cannot use the test as it is and some finer modifications are necessary to deal with the statistical changes due to the data-dependent greedy selection of predictor variables. This point will be further discussed in Chapter 3 where we propose the Multi-Beta-Test, a model selection method for linear regression based on hypothesis testing and employed along with OMP.

2.3.2 Information Theoretic Criteria

In the previous section, we discussed how hypothesis testing can be used for model selection in linear regression. We also mentioned some issues in using hypothesis testing based methods, primarily the requirement of a nested structure of models to be tested and the deviation of the distribution under the null hypothesis from its true distribution when using greedy algorithms for variable selection in order to obtain a nested sequence of models. These issues can be easily avoided if we employ information theoretic approaches for model selection. In this case, we may reformulate the linear model in (2.2) as follows

$$\mathcal{H}_{\mathcal{I}} : \mathbf{y} = \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}} + \mathbf{e}_{\mathcal{I}}, \tag{2.20}$$

where $\mathcal{I} \subset \{1, \ldots, p\}$, $\mathbf{A}_{\mathcal{I}} \equiv (\mathbf{a}_j, j \in \mathcal{I})$. $\mathbf{x}_{\mathcal{I}}$ and $\mathbf{e}_{\mathcal{I}} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathcal{I}}\mathbf{I}_N)$ are the associated parameter vector and noise vector, respectively. The ML estimate of the noise variance under $\mathcal{H}_{\mathcal{I}}$ is $\hat{\sigma}^2_{\mathcal{I}} = \mathbf{y}\mathbf{\Pi}^{\perp}_{\mathcal{I}}\mathbf{y}/N$.

The literature on information theoretic criteria is quite extensive [2, 14, 43–45]. A typical information criterion based model selection rule picks the best model that minimizes some statistical metric as

$$\hat{\mathcal{S}} = \arg\min_{\mathcal{I}\in\mathcal{J}} \{f(\mathcal{H}_{\mathcal{I}}) + \mathcal{P}(\mathcal{I})\}, \tag{2.21}$$

where $\hat{\mathcal{S}}$ is the true model estimate, $\mathcal{J}$ is the set of candidate models under consideration and $\mathcal{H}_{\mathcal{I}}$ denotes the model with support $\mathcal{I}$. The statistical metric consists of two parts: (1) $f(\mathcal{H}_{\mathcal{I}})$ representing the goodness of fit of model $\mathcal{H}_{\mathcal{I}}$ and (2) $\mathcal{P}(\mathcal{I})$ is the penalty term that compensates for overparameterization. The literature on model selection is quite extensive. Some of the popular classical model selection rules include Akaike information criterion (AIC) [46], Bayesian information criterion (BIC) [47], minimum description length (MDL) [48], gMDL [49], nMDL [50], and penalizing adaptively the likelihood (PAL) [51]. Below a brief summary of some of the popular criteria is provided.

AIC : The first and one of the most well-known information criterion is the Akaike information criterion developed by Hirotugu Akaike in 1974 [46]. The basic tenet of AIC is to estimate the goodness of a statistical model by gauging how closely the prediction distribution specified by the model resembles the actual true distribution. Akaike adopted the Kullback–Leibler information (divergence) to measure this closeness, which led to the derivation of AIC. For the linear regression in (2.20) the AIC value of a model with support $\mathcal{I}$ is defined as

$$\mathrm{AIC}(\mathcal{I}) = N \ln(\hat{\sigma}^2_{\mathcal{I}}) + 2k, \tag{2.22}$$

where $k = card(\mathcal{I})$ is the number of model parameters. Given a set of candidate models for the data, the model with the minimum AIC value is preferred. Notice that AIC assigns goodness of fit to a model based on the likelihood function, but it also contains a penalty that grows in proportion to the number of parameters in the model. The penalty inhibits overfitting, which is desirable because increasing the number of parameters in the model almost always enhances the goodness of fit, but we may end up with a model with too many unwanted parameters. In practice, AIC does a good job of minimizing the underfitting probability, however, it suffers from an overfitting problem. If we assume that the true model generating the data is indeed present in the collection of candidate models, then for AIC it can be shown that the probability of underfitting $\to 0$ and the probability of overfitting $\to$ constant > 0 as the sample size $N \to \infty$. Hence, AIC is not a consistent model selection criterion [52]. This inconsistency is arising because the penalty term of AIC ($2k$) is too small and fails to compensate appropriately as the model size increases, hence the overfitting issue.

BIC : In 1978, Schwarz presented a new information criterion now popularly known as the Bayesian information criterion (BIC) [47]. BIC is formulated from the maximum a-posteriori or MAP estimator, which has its roots in the Bayesian framework. Under this regime, the model that maximizes the posterior probability is selected. The BIC value of a model with support $\mathcal{I}$ is defined as

$$\mathrm{BIC}(\mathcal{I}) = N \ln(\hat{\sigma}_{\mathcal{I}}^2) + k \log N, \tag{2.23}$$

where k is the model dimension and N is the sample size. The BIC picks the model that minimizes (2.23). As compared to AIC, BIC is a consistent criterion, i.e., the probability of correctly detecting the true model $\to 1$ as $N \to \infty$. This is precisely because the penalty value of BIC for an arbitrary model with dimension k ($k \log N$) grows large as the number of measurements $N \to \infty$ and is not fixed as in the case of AIC, thus providing a much higher penalty for overfitting. BIC will be discussed in more detail in Chapter 4, particularly in the context of model selection in linear regression. The different forms of the BIC are also presented and the newly developed BIC-Robust ($\mathrm{BIC_R}$ in short) to handle the data scaling problem that exists in the high-SNR forms of the BIC. Fig. 2.2 provides a comparison of AIC and BIC scores of a linear regression model as a function of the model dimension k. The models are assumed to be nested. The true support is chosen as $\mathcal{S} = \{1, 2, 3, 4\}$, hence $k_0 = 4$. Also for the simulation purpose we consider $N = 60$ and $p = 20$. The model scores are computed up to $k = 12$. If we look at the BIC plot, the model score starts from a high value at $k = 1$ (i.e., the model with only predictor $\mathbf{a}_1$) and it gradually diminishes as k increases, or in other words as we add more predictors to the model. It reaches a minimum value at $k = 4$, which is indeed the true model order. For $k > 4$, the BIC score starts growing as k increases.

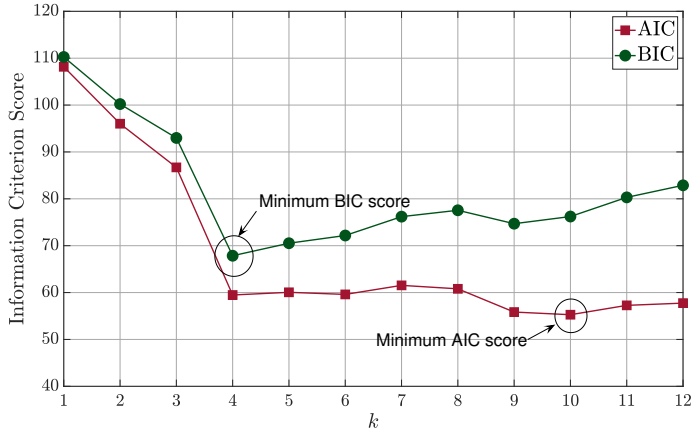


Figure 2.2: Comparing the AIC and BIC score as a function of model dimension k with $N = 60$, $p = 20$, SNR = 3 dB and $\mathcal{S} = \{1, 2, 3, 4\}$.

Therefore, BIC performs a correct model order selection. On the contrary, the AIC score reaches a low value at $k = 4$, however, we do observe that there are values of k for which the AIC score is much lower. Hence, in this case, AIC fails to select the true model order. This small example in a way illustrates the overfitting issue of AIC and underlines the drawback of its penalty to handle larger dimensions.

Another model selection criterion that has a similar structure to BIC but derived from a completely different approach based on coding arguments and the minimum description length principle (MDL) is the MDL criterion [48, 53]. In MDL, the parameters are derived using the MDL principle under the assumption that there is no prior knowledge about the model parameters. There are two popular upgrades of the MDL criterion namely the gMDL [49] and the nMDL [50]. Both these criteria were designed to solve the consistency issue of MDL in the high-SNR regime.

Information criteria are excellent methods of model selection. They have proved to be very useful and reliable in selecting a good model based on the available data. Hence, they are ubiquitous in many fields. However, when dealing with high-dimensional scenarios where $p \gg N$, these classical methods tend to heavily overfit and fail to guarantee consistency especially when p grows with N. This issue has been thoroughly discussed in [17] and [18], where the authors provide extended versions of BIC to handle the large-p small-N problem and to guarantee consistency in both large-N and high-SNR scenarios. In [17] the authors add a binomial coefficient penalty to the BIC's objective function that leads to the extended family of the BIC termed as extended BIC (EBIC). This extra penalty negates the idea of assigning uniform prior probability to the models used in the original BIC and allocates a dynamic prior that depends on the model dimension. As the model dimension increases, the prior probability assigned to the model diminishes. This is in tune with the law of parsimony where we prefer smaller models over larger ones. It is shown that under a suitable asymptotic identifiability condition, EBIC can consistently select the true model as the sample size N grows to infinity. The EBIC score for a linear regression model with support $\mathcal{I}$ is evaluated as

$$\text{EBIC}(\mathcal{I}) = N \ln(\hat{\sigma}_{\mathcal{I}}^2) + k \ln(N) + 2\gamma k \ln(p), \tag{2.24}$$

where $\gamma \in (0, 1)$ is a tuning parameter. Fig. 2.3 compares the model score versus the dimension k for BIC and EBIC. The considered parameters are $N = 100$, $p = 500$ SNR $= 6$ dB. OMP is used to pick a set of initial significant predictors up to maximum cardinality $k = 11$. In this case, OMP indeed recovers the true support $\mathcal{S} = \{1, 2, 3, 4\}$ in the fist $k_0 = 4$ iterations. It is quite clear from the figure that BIC fails in handling the high-dimensional scenario. Thus, the penalty of the classical BIC is insufficient to counteract the large-p case and the greedy selection procedure. On the contrary, EBIC successfully handles the large-p small-N scenario. The minimum EBIC score occurs at $k = 4$, which is the true sparsity. Hence, for the above problem, employing EBIC for model selection will result in the selection of the true model. However, if instead BIC is used we end up with an overfitted model with a total of 11 parameters where 4 are true and 7 are false.

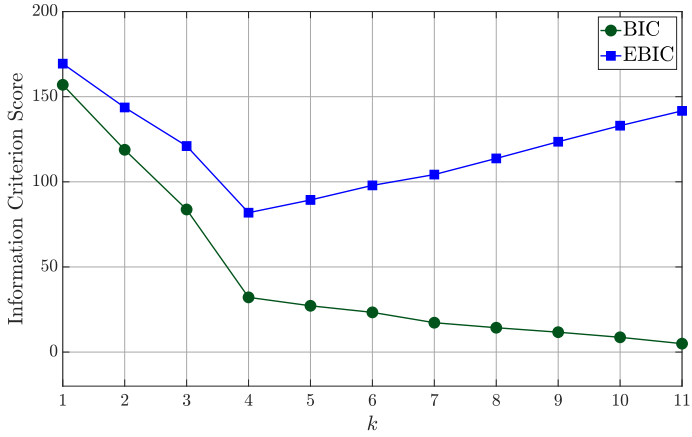


Figure 2.3: Information criterion score versus model dimension k with $N = 100$, $p = 500$, SNR $= 6$ dB and $\mathcal{S} = \{1, 2, 3, 4\}$

The performance of EBIC is quite appreciable in large sample scenarios. However, the empirical performance of EBIC can sometimes be unsatisfactory for practical sizes of N. Moreover, in scenarios when N is fixed but the noise variance, σ^2, tends to zero, results show that EBIC is inconsistent [18]. To handle the consistency issue for decreasing noise variance scenario, the authors in [18] proposed an improved criterion for model selection in the high-dimensional setting known as the extended Fisher information criterion (EFIC). We can view EFIC as a combination of EBIC and FIC [54] and it alleviates the inconsistency problem of EBIC in high-SNR. The EFIC score for a model with support $\mathcal{I}$ is evaluated as

$$\text{EFIC}(\mathcal{I}) = (N - k - 2)\ln\|\mathbf{y}\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 + k\ln(N) + \ln\left|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right| + 2ck\ln(p), \quad (2.25)$$

where $c > 0$ is a tuning parameter. Fig. 2.4 highlights this point. It presents the probability of correct model selection (PCMS) versus SNR in dB. We consider the scenario where N is small and fixed, while $p \gg N$. As SNR increases, EFIC is seen to achieve empirical consistency, i.e., PCMS $\to 1$ as SNR$\to \infty$. This is however not true for EBIC. A significant performance gap is observed between the two methods. Also, the PCMS of EBIC does not tend to one even when SNR reaches a very high value. The authors in [18], show that under a certain asymptotic identifiability condition, EFIC is consistent, i.e., PCMS $\to 1$ as $N \to \infty$ and/or SNR $\to \infty$. However, EFIC suffers from a data scaling problem because its penalty is data-dependent. This point will be further discussed in Chapter 5 where we present the modified criterion called EBIC$_{\text{R}}$ to resolve the data scaling issue and guarantee consistency in both large-N and high-SNR scenarios.

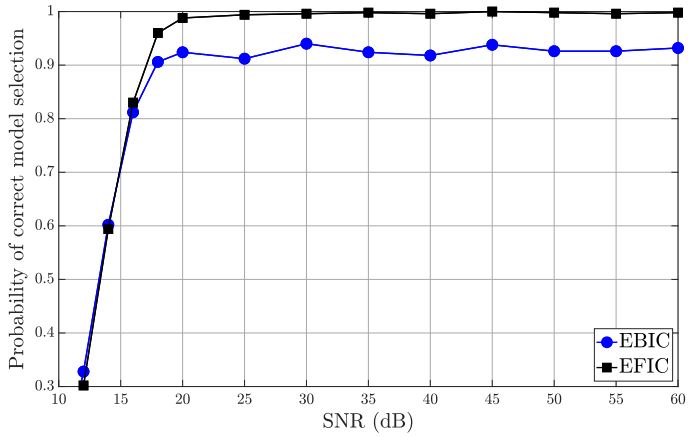


Figure 2.4: PCMS versus SNR (dB) with $N = 55$, $p = 1000$, and $\mathcal{S} = \{1, 2, 3, 4, 5\}$.

2.3.3 Cross-Validation

Cross-validation (CV) is a well-known model selection method that evaluates the best model based on the predictive ability of the models under consideration. CV does not require the knowledge or the need to compute the likelihood of the generating function or any other quantity that depends on the statistical assumption of the underlying model, which makes it a very general approach for model selection. CV is based on a data splitting procedure. Given a dataset, a segment of the data is used for fitting each competing candidate model and the remaining data is used to measure the predictive performances of the models by the validation errors. The model with the best overall predictive performance is selected [55, 56].

Depending on how the data is split for fitting and validation, CV can be classified into different types. If we have N data samples, we hold one sample and use the remaining $N - 1$ for fitting a model, and the withheld one data sample for validation, this form of the CV is called leave-one-out (LOO) or delete-1 CV. Here, the predictive performance is evaluated for all the data points, and the model with the best average performance is selected [57]. It has been observed that LOO is asymptotically equivalent to AIC [58] and is inconsistent as $N \to \infty$. On the other hand, instead of leaving one sample out for validation if we leave N_v samples out, then this is called delete-N_v CV. Under the assumption $N_v/N \to 1$, delete-N_v CV is consistent as $N \to \infty$.

The other approach to CV is the K-fold CV. Here, the entire dataset is randomly divided into K partitions of equal size such that each partition has N/K data samples. Out of the K partitions, data from $K - 1$ partitions are used to fit the model and the remaining data from the single partition is used to validate the predictive performance. The CV process is then repeated K times, with data from

each of the K partitions used exactly once as the validation set. Finally, the K predictive results of each model are averaged to obtain a generalized performance across the K partitions. These averaged results can be scrutinized to select the best model.

CV works well when the size of the candidate model space is small and the sample size N is large. However, in the large-p small-N scenario, employing CV may not be viable as it tends to have high variance [59]. Also, CV-based procedures can be computationally intensive and their performance in high-dimensional problems is not satisfactory [60, 61].

Chapter 3

Multi Beta Test

"All models are wrong but some are useful."
—*George E. P. Box (1919–2013)*

In Chapter 2, we discussed how hypothesis testing can be used for model selection in linear regression. We also further highlighted that, when a greedy approach is employed for predictor selection, the statistical distribution of the test statistic under the null hypothesis changes. This happens because the greedy procedure has intentionally chosen the strongest predictor among all of the available choices, hence the models no longer have a pre-defined nested structure but a sequence that is data dependent [29]. This chapter proposes a novel model selection method named Multi-Beta-Test (MBT) for the sparse high-dimensional linear regression, that employs a greedy algorithm for predictor selection. The estimation of the correct subset in the linear regression problem is formulated as a series of hypothesis tests where the test statistic is based on the relative least-squares cost of successive parameter models. Extensive simulation results are performed to analyze the behaviour of MBT under different data parameter settings. Furthermore, the performance of MBT is compared to existing model selection methods for high-dimensional parameter space such as extended Bayesian information criterion (EBIC), extended Fisher Information criterion (EFIC), residual ratio thresholding (RRT), and orthogonal matching pursuit (OMP) with *a priori* knowledge of the sparsity.

3.1 Proposed Method

Consider the standard linear regression model as described in 2.2,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}. \tag{3.1}$$

Here $\mathbf{y}$ is the N dimensional vector of real measurement responses, $\mathbf{A} \in \mathbb{R}^{N \times p}$ is the known design matrix and comprising of the column vectors $\{\mathbf{a}_1, \ldots, \mathbf{a}_p\}$ also known

as regressors. The vector $\mathbf{e} \in \mathbb{R}^N$ is the error or noise vector whose elements are assumed to be independent and identically Gaussian distributed, $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$. The parameter $\sigma^2 \geq 0$ is the unknown noise variance. Here, we consider the high-dimensional setting where $p > N$ or $p \gg N$, i.e., $\mathbf{A}$ has more columns than rows. This is an underdetermined system with many solutions and classical methods such as ordinary least-squares are no longer applicable. A practical and widely used valid assumption to handle the $p \gg N$ situation is to assume that the underlying unknown parameter vector $\mathbf{x} \in \mathbb{R}^p$ is sparse. By sparse it means that only a few of the elements of $\mathbf{x}$ are non-zero, i.e., the support of $\mathbf{x}$ given by $\mathcal{S}_0 = \{i : x_i \neq 0\}$ has cardinality $card(\mathcal{S}_0) = k_0 \ll \min(N, p)$. In this case, $\mathbf{x}$ is termed as a k_0-sparse vector [62]. In this section, we present in detail a novel method called MBT for estimating the true sparsity k_0. OMP with $K \ll N$ iterations is used for predictor selection as it provides a sequence of nested models up to maximum cardinality K. Here, we assume that $k_0 < K$, which is a necessary condition for MBT to be able to select the true sparsity.

Let $\mathbf{A}_\mathcal{S} \in \mathbb{R}^{N \times s}$ be a matrix constructed using some columns from the design matrix $\mathbf{A}$ with support $\mathcal{S} \subset \{1, 2, \ldots, p\}$ and cardinality $card(\mathcal{S}) = s \ll \min(N, p)$. Let $\mathbf{x}_\mathcal{S} \in \mathbb{R}^s$ be the unknown regressor coefficient vector corresponding to $\mathbf{A}_\mathcal{S}$. The linear regression model then can be rewritten as $\mathbf{y} = \mathbf{A}_\mathcal{S}\mathbf{x}_\mathcal{S} + \mathbf{e}$. The method of least-squares provides estimates of $\mathbf{x}_\mathcal{S}$ by minimizing the least-squares cost function $V_\mathcal{S}(\mathbf{x}_\mathcal{S}) = \|\mathbf{y} - \mathbf{A}_\mathcal{S}\mathbf{x}_\mathcal{S}\|_2^2$ where $\|\cdot\|_2$ denotes the Euclidean vector norm. The minimizer is $\hat{\mathbf{x}}_\mathcal{S} = (\mathbf{A}_\mathcal{S}^T\mathbf{A}_\mathcal{S})^{-1}\mathbf{A}_\mathcal{S}^T\mathbf{y}$ and the least-squares estimate of the output response $\mathbf{y}$ is $\mathbf{\Pi}_\mathcal{S}\mathbf{y}$ [63]. Then, for the design matrix $\mathbf{A}_\mathcal{S}$ with support $\mathcal{S}$, the least-squares cost is

$$V_\mathcal{S} = \|(\mathbf{I} - \mathbf{\Pi}_\mathcal{S})\mathbf{y}\|_2^2 = \left\|\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{y}\right\|_2^2 = \mathbf{y}^T\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{y}, \tag{3.2}$$

where $\mathbf{\Pi}_\mathcal{S}^\perp$ is the orthogonal projection matrix onto the null space of $\mathbf{A}_\mathcal{S}^T$. Now consider a new matrix $\mathbf{A}_\mathcal{I} \in \mathbb{R}^{N \times k}$ with support $\mathcal{I} \subset \{1, \ldots, p\}$ having cardinality $card(\mathcal{I}) = k$ and consisting of columns from the original design matrix $\mathbf{A}$ but not present in $\mathbf{A}_\mathcal{S}$. Concatenating $\mathbf{A}_\mathcal{S}$ and $\mathbf{A}_\mathcal{I}$ forms the new matrix $[\mathbf{A}_\mathcal{S}\ \mathbf{A}_\mathcal{I}]$. The new least-squares cost after incorporating $\mathbf{A}_\mathcal{I}$ is evaluated as

$$\begin{aligned} V_{[\mathcal{S},\mathcal{I}]} = \left\|\mathbf{\Pi}_{[\mathbf{A}_\mathcal{S}\ \mathbf{A}_\mathcal{I}]}^\perp \mathbf{y}\right\|_2^2 &= \left\|\mathbf{y} - \mathbf{\Pi}_\mathcal{S}\mathbf{y} - \mathbf{\Pi}_{\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{A}_\mathcal{I}}\mathbf{y}\right\|_2^2 \\ &= \left\|\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{y} - \mathbf{\Pi}_{\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{A}_\mathcal{I}}\mathbf{y}\right\|^2 \\ &= \left(\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{y} - \mathbf{\Pi}_{\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{A}_\mathcal{I}}\mathbf{y}\right)^T \left(\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{y} - \mathbf{\Pi}_{\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{A}_\mathcal{I}}\mathbf{y}\right) \\ &= V_\mathcal{S} - \mathbf{y}^T\mathbf{\Pi}_{\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{A}_\mathcal{I}}\mathbf{y}, \end{aligned} \tag{3.3}$$

where $\mathbf{\Pi}_{\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{A}_\mathcal{I}}$ is the projection onto the space spanned by $\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{A}_\mathcal{I}$ defined as

$$\mathbf{\Pi}_{\mathbf{\Pi}_\mathcal{S}^\perp \mathbf{A}_\mathcal{I}} = \mathbf{\Pi}_\mathcal{S}^\perp\mathbf{A}_\mathcal{I}\left(\mathbf{A}_\mathcal{I}^T\mathbf{\Pi}_\mathcal{S}^\perp\mathbf{A}_\mathcal{I}\right)^{-1}\mathbf{A}_\mathcal{I}^T\mathbf{\Pi}_\mathcal{S}^\perp. \tag{3.4}$$

Using (3.2) and (3.3), the relative cost is evaluated as

$$w_{\mathcal{I}} = \frac{V_{\mathcal{S}} - V_{[\mathcal{S},\mathcal{I}]}}{V_{\mathcal{S}}} = \frac{\mathbf{y}^T \mathbf{\Pi}_{\mathbf{\Pi}_{\mathcal{S}}^{\perp} \mathbf{A}_{\mathcal{I}}} \mathbf{y}}{\mathbf{y}^T \mathbf{\Pi}_{\mathcal{S}}^{\perp} \mathbf{y}}. \tag{3.5}$$

Substituting (3.4) in (3.5)we get

$$w_{\mathcal{I}} = \frac{\mathbf{y}^T \mathbf{\Pi}_{\mathcal{S}}^{\perp} \mathbf{A}_{\mathcal{I}} (\mathbf{A}_{\mathcal{I}}^T \mathbf{\Pi}_{\mathcal{S}}^{\perp} \mathbf{A}_{\mathcal{I}})^{-1} \mathbf{A}_{\mathcal{I}}^T \mathbf{\Pi}_{\mathcal{S}}^{\perp} \mathbf{y}}{\mathbf{y}^T \mathbf{\Pi}_{\mathcal{S}}^{\perp} \mathbf{y}}. \tag{3.6}$$

Thus, $w_{\mathcal{I}}$ is a measure of the decrease in the least-squares cost relative to $V_{\mathcal{S}}$ after addition of k extra columns of $\mathbf{A}$. Next, we derive the statistical properties of $w_{\mathcal{I}}$. For this let us assume that $\mathcal{S} = \mathcal{S}_0$, the true support of $\mathbf{x}$. Then the true data model is $\mathbf{y} = \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}} + \mathbf{e}$. Furthermore, the projection matrix $\mathbf{\Pi}_{\mathcal{S}}^{\perp}$ can be decomposed as $\mathbf{\Pi}_{\mathcal{S}}^{\perp} = \mathbf{U}\mathbf{U}^T$ where $\mathbf{U} \in \mathbb{R}^{N \times (N-s)}$ is a semi-orthogonal matrix whose columns span the null space of $\mathbf{A}_{\mathcal{S}}$. We also denote $\tilde{\mathbf{e}} = \mathbf{U}^T \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{N-s})$ where $s = card(\mathcal{S})$ and let $\tilde{\mathbf{A}}_{\mathcal{I}} = \mathbf{U}^T \mathbf{A}_{\mathcal{I}} \in \mathbb{R}^{(N-s) \times k}$ where $k = card(\mathcal{I})$ and $N - s > k$. Therefore, when $\mathcal{S}$ is the true subspace we can rewrite (3.6) using the above substitutions and the fact that $\mathbf{\Pi}_{\mathcal{S}}^{\perp} \mathbf{A}_{\mathcal{S}} = 0$ as

$$\begin{aligned}
w_{\mathcal{I}} &= \frac{\mathbf{e}^T \mathbf{U}\mathbf{U}^T \mathbf{A}_{\mathcal{I}} (\mathbf{A}_{\mathcal{I}}^T \mathbf{U}\mathbf{U}^T \mathbf{A}_{\mathcal{I}})^{-1} \mathbf{A}_{\mathcal{I}}^T \mathbf{U}\mathbf{U}^T \mathbf{e}}{\mathbf{e}^T \mathbf{U}\mathbf{U}^T \mathbf{e}} \\
&= \frac{\tilde{\mathbf{e}}^T \tilde{\mathbf{A}}_{\mathcal{I}} (\tilde{\mathbf{A}}_{\mathcal{I}}^T \tilde{\mathbf{A}}_{\mathcal{I}})^{-1} \tilde{\mathbf{A}}_{\mathcal{I}}^T \tilde{\mathbf{e}}}{\tilde{\mathbf{e}}^T \tilde{\mathbf{e}}} = \frac{\tilde{\mathbf{e}}^T \tilde{\mathbf{\Pi}}_{\mathcal{I}} \tilde{\mathbf{e}}}{\tilde{\mathbf{e}}^T \tilde{\mathbf{e}}} \\
&= \frac{\tilde{\mathbf{e}}^T \tilde{\mathbf{\Pi}}_{\mathcal{I}} \tilde{\mathbf{e}}}{\tilde{\mathbf{e}}^T (\mathbf{I}_N - \tilde{\mathbf{\Pi}}_{\mathcal{I}} + \mathbf{\Pi}_{\mathcal{I}}) \tilde{\mathbf{e}}} \\
&= \frac{\tilde{\mathbf{e}}^T \tilde{\mathbf{\Pi}}_{\mathcal{I}} \tilde{\mathbf{e}}}{\tilde{\mathbf{e}}^T \tilde{\mathbf{\Pi}}_{\mathcal{I}} \tilde{\mathbf{e}} + \tilde{\mathbf{e}}^T \tilde{\mathbf{\Pi}}_{\mathcal{I}}^{\perp} \tilde{\mathbf{e}}} = \frac{X_1}{X_1 + X_2},
\end{aligned}$$

where $\tilde{\mathbf{\Pi}}_{\mathcal{I}}$ is the projection matrix onto the span of $\tilde{\mathbf{A}}_{\mathcal{I}}$. Next, observe that X_1 and X_2 are independent random variables distributed as $X_1 \sim \chi^2(k)$ and $X_2 \sim \chi^2(N - s - k)$. It is well known from theory that if X_1 and X_2 are independent chi-squared distributed random variables then $w_{\mathcal{I}}$ follows a Beta distribution with parameters $k/2$ and $(N - s - k)/2$ [64], i.e.,

$$w_{\mathcal{I}} \sim \mathcal{B}\left(\frac{k}{2}, \frac{N - s - k}{2}\right). \tag{3.7}$$

Now, from (3.5) we see that minimizing the least-squares cost $V_{[\mathcal{S},\mathcal{I}]}$ over $\mathcal{I}$ is equivalent to maximizing the relative cost $w_{\mathcal{I}}$. Let us denote

$$w_{\mathcal{I}^*} = \max_{\mathcal{I}_i} w_{\mathcal{I}_i} \quad ; \; \mathcal{I}_i \subset \{1, 2, \ldots, p\} \setminus \mathcal{S} \quad ; \; card(\mathcal{I}_i) = k,$$

where $\{1, 2, \ldots, p\} \setminus \mathcal{S}$ denotes the set difference between the two sets and $i = 1, 2, \ldots, \binom{p-s}{k}$. Now, the probability that the maximum relative cost $w_{\mathcal{I}^*}$ is less than some threshold γ can be expressed as

$$\begin{aligned}
\Pr\left(w_{\mathcal{I}^*} < \gamma\right) &= \Pr\left(w_{\mathcal{I}_1} < \gamma \;\&\; \ldots \;\&\; w_{\mathcal{I}_{\binom{p-s}{k}}} < \gamma\right) \\
&= 1 - \Pr\left(w_{\mathcal{I}_1} > \gamma \text{ or } \ldots \; w_{\mathcal{I}_{\binom{p-s}{k}}} > \gamma\right) \\
\therefore \mathrm{P}(w_{\mathcal{I}^*} < \gamma) &\geq 1 - \binom{p-s}{k} \Pr\left(w_{\mathcal{I}} > \gamma\right), \qquad (3.8)
\end{aligned}$$

where the last inequality follows from the union bound. Hence, (3.8) defines a lower bound on the probability $\Pr(w_{\mathcal{I}^*} < \gamma)$. Setting this lower bound probability to some value $\beta \in (0, 1)$ we get

$$\begin{aligned}
\binom{p-s}{k} \Pr\left(w_{\mathcal{I}} > \gamma\right) &= 1 - \beta \\
\Pr\left(w_{\mathcal{I}} < \gamma\right) &= 1 - \frac{1-\beta}{\binom{p-s}{k}} = \rho \text{ (say)}. \qquad (3.9)
\end{aligned}$$

Since $w_{\mathcal{I}} \sim \mathcal{B}\left(k/2, (N-s-k)/2\right)$, the threshold γ can be evaluated as

$$\gamma = \mathcal{B}^{-1}\left(\rho \,;\, \frac{k}{2}, \frac{N-s-k}{2}\right), \qquad (3.10)$$

where $\mathcal{B}^{-1}(\cdot)$ is the inverse beta cumulative distribution function.

The MBT is summarized in Algorithm 2. First, specify the design matrix $\mathbf{A}$ and the measurement vector $\mathbf{y}$. Then run K iterations of OMP to identify the most appropriate K column vectors of $\mathbf{A}$ in order of decreasing significance. This gives us the OMP generated index set $\mathcal{S}^K_{\mathrm{OMP}}$. Next, at each iteration $s = 1, \ldots, (K-1)$, compute the least-square cost $V_{\mathcal{S}^s_{\mathrm{OMP}}}$ and generate a sequence of relative cost $\{w_{\mathcal{I}}(k)\}$ and the corresponding sequence of threshold $\{\gamma(k)\}$ for $k = 1, \ldots, (K-s)$. The true sparsity k_0 is estimated as the value of s at which $w_{\mathcal{I}}(k) < \gamma(k)$, $\forall k = 1, \ldots, (K-s)$.

3.2 Simulation Results

In this section, simulation results are presented to evaluate the performance of MBT. The performance is measured in terms of the probability of correct model selection (PCMS) versus two varying parameters (1) number of measurements, N and (2) dimension of parameter space, p. Since a high-dimensional scenario is considered, $p > N$ in all cases. The design matrix $\mathbf{A} \in \mathbb{R}^{N \times p}$ is generated with independent entries following normal distribution $\mathcal{N}(0, 1)$. The columns of $\mathbf{A}$ are normalized to have unit Euclidean (l_2) norm. Since the parameter vector $\mathbf{x}$ is

Algorithm 3.1 MBT as used with OMP

Input: Design matrix $\mathbf{A}$, measurement vector $\mathbf{y}$
Run K iterations of OMP to get index set $\mathcal{S}^K_{\text{OMP}}$
for $s = 1$ to $K - 1$ **do**
    Compute $V_{\mathcal{S}^s_{\text{OMP}}} = \mathbf{y}^T \mathbf{\Pi}^{\perp}_{\mathcal{S}^s_{\text{OMP}}} \mathbf{y}$
    for $k = 1$ to $K - s$ **do**
        $w_{\mathcal{I}}(k) = \frac{V_{\mathcal{S}^s_{\text{OMP}}} - V_{\mathcal{S}^{s+k}_{\text{OMP}}}}{V_{\mathcal{S}^s_{\text{OMP}}}}$
        Compute threshold $\gamma(k)$
    end for
    if $w_{\mathcal{I}}(k) < \gamma(k), \forall k$ **then**
        break
    else
        Continue
    end if
end for
Estimated true sparsity $\hat{k}_0 = s$
Estimated true support $\hat{\mathcal{S}}_0 = \mathcal{S}^s_{\text{OMP}}$
Estimated parameter vector $\hat{\mathbf{x}}_{\hat{\mathcal{S}}_0} = (\mathbf{A}^T_{\hat{\mathcal{S}}_0} \mathbf{A}_{\hat{\mathcal{S}}_0})^{-1} \mathbf{A}^T_{\hat{\mathcal{S}}_0} \mathbf{y}$.

assumed to be sparse, the true support cardinality is fixed at $k_0 = 5$. The non-zero entries of $\mathbf{x}$ are taken as $x_i = 1$ where the indices i are taken uniformly at random from $\{1, \ldots, p\}$. The SNR in dB is SNR (dB) $= 10 \log_{10}(\sigma_s^2/\sigma^2)$, where σ_s^2 and σ^2 denote signal and noise power, respectively. The signal power is computed as $\sigma_s^2 = \|\mathbf{Ax}\|_2^2/N$. Based on σ_s^2 and the chosen SNR (dB), the noise power is set as $\sigma^2 = \sigma_s^2/10^{\text{SNR (dB)}/10}$. The PCMS is estimated over 1000 Monte Carlo trials. To maintain randomness in the data, a new design matrix $\mathbf{A}$ is generated at each Monte Carlo trial. For OMP we set $K = 20$. After K iterations we have the OMP generated index set $\mathcal{S}^K_{\text{OMP}}$ where $card\left(\mathcal{S}^K_{\text{OMP}}\right) = K$. For example, say we get the OMP generated index set as $\mathcal{S}^K_{\text{OMP}} = [10, 2, 4, 8, 31, 96, 5]$, where $K = 7$ in this case. Then the candidate models are $\{10\}$, $\{10, 2\}$, $\{10, 2, 4\}$, $\{10, 2, 4, 8\}$, $\{10, 2, 4, 8, 31\}$, $\{10, 2, 4, 8, 31, 96\}$, $\{10, 2, 4, 8, 31, 96, 5\}$. The order of each successive model grows by one. Hence, OMP generates a monotonic nested set of candidate models where any candidate model is a subset of its successor, i.e., $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_K$. Here, we compare the performance of MBT, with *viz.*, EBIC, EFIC, and RRT. EBIC and EFIC are defined in (2.24) and (2.25). In RRT, the following residual ratio statistic is evaluated as [65]

$$RR(k) = \|\mathbf{r}^k\|_2 / \|\mathbf{r}^{k-1}\|_2, \tag{3.11}$$

where $\|\mathbf{r}\|_2^k = \|\mathbf{y}\mathbf{\Pi}^{\perp}_{\mathcal{I}}\mathbf{y}\|_2^2$ and $k = 1, \ldots, K$. The RRT algorithm is shown in Algorithm 3.2 where it is implemented along with OMP with K iterations and the

Algorithm 3.2 RRT with OMP

1: **Inputs:** Design matrix $\mathbf{A}$, observation vector $\mathbf{y}$.
2: **Step 1** Run K iterations of OMP
3: **Step 2** Compute $RR(k)$ for $k = 1, \ldots, K$
4: **Step 3** Compute $k_{RRT} = \max\{k : RR(k) \leq \Gamma^{\alpha}_{RRT}(k)\}$
5: **Outputs:** True support estimate $\hat{\mathcal{S}}_0 = \mathcal{S}^{k_{RRT}}_{\text{OMP}}$.

quantity $\Gamma^{\alpha}_{RRT(k)}$ is evaluated as

$$\Gamma^{\alpha}_{RRT(k)} = \sqrt{\mathcal{B}^{-1}\left(\rho; \frac{N-k}{2}, \frac{1}{2}\right)}, \tag{3.12}$$

where $\alpha \in (0, 1]$ is a tuning parameter and $\rho = \frac{\alpha}{K(p-k-1)}$. In all of the methods described above, the tuning parameter plays a very crucial role in their performance in selecting the true model. Different values of the tuning parameter for the same method may produce diverse performance curves for the same data set, which makes it hard to compare different methods. For our convenience, we chose the following values of the tuning parameter $\gamma = 1$ (EBIC), $c = 1$ (EFIC), $\alpha = 0.1$ (RRT). These values are motivated by their respective original papers.

Fig. 3.1 presents the plot for the probability of correct model selection versus the number of measurements, N. For this case the parameters considered are SNR = 2 dB, $p = 500$, $k_0 = 5$ and two different β values are taken into account, $\beta = [0.95, 0.99]$, to highlight the effect of β in the performance of MBT. It can be seen from the figure that for the given setup and $\beta = 0.95$, the proposed method

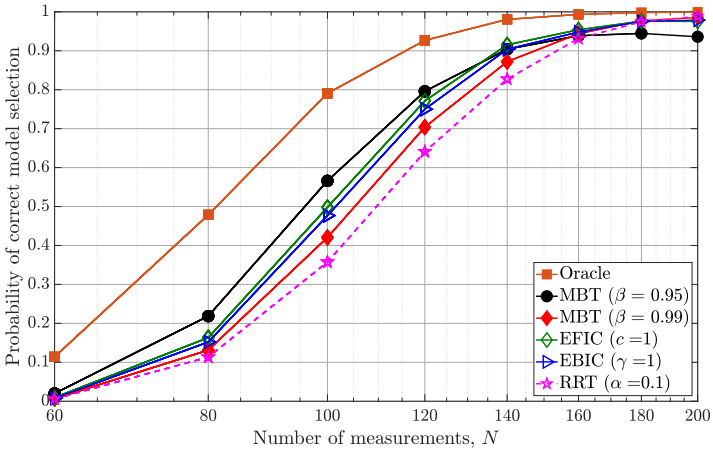


Figure 3.1: PCMS versus N when SNR = 2 dB, $p = 500$, and $k_0 = 5$.

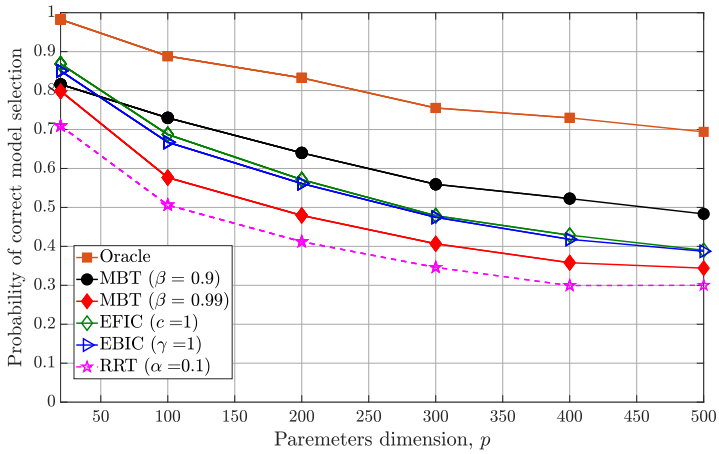


Figure 3.2: PCMS versus p when SNR = 3 dB, $N = 80$, and $k_0 = 5$.

MBT gives a slightly higher probability of correct model selection compared to EFIC, EBIC, and RRT for lower values of $N(< 120)$. On increasing N further, the maximum PCMS for MBT ($\beta = 0.95$) settles at close to 0.95. On the contrary, MBT with $\beta = 0.99$ achieves a maximum probability close to 0.99 with the increase in N, but at lower measurements ($N < 160$) its performance degrades compared to EFIC and EBIC. The performance is also compared to OMP-oracle, which is OMP with known *a priori* knowledge of sparsity k_0 and is the optimal performance that OMP can achieve. RRT with $\alpha = 0.1$ has a lower performance for $N < 160$, however, as N grows the PCMS goes close to one.

Fig. 3.2 illustrates the probability of correct model selection versus parameter dimension, p for SNR = 3 dB, $N = 80$, $k_0 = 5$ and $\beta = [0.90, 0.99]$. The performance of all the methods decreases with the increase in the parameter dimension p. However, it is seen that MBT (with $\beta = 0.90$) provides a higher probability of correct selection as compared to EFIC, EBIC, and RRT for higher model dimensions under low N and SNR values.

3.2.1 Effect of β On The Performance of MBT

The choice of the tuning parameter β is crucial to the performance of MBT. If the value of β is too close to one, it will result in high underfitting losses in the small-N and low-SNR regions. This is because a high β value corresponds to a high threshold value used in the hypothesis test. As such, regression coefficients having small magnitudes have a high probability of failing the test and hence getting dropped from the final support leading to an underfitted model. A higher value of β is desired when the sample size is large and/or the SNR is high such that the noise level is relatively lower than the signal components. This will ensure

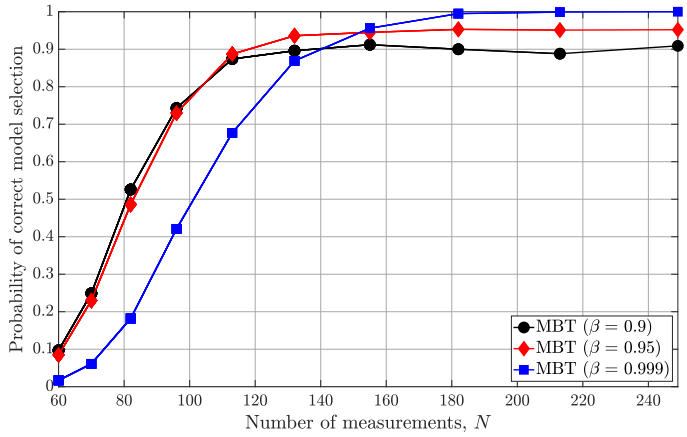


Figure 3.3: Performance of MBT versus N for different values of β. Here SNR = 3 dB, $p = 500$ and $k_0 = 5$.

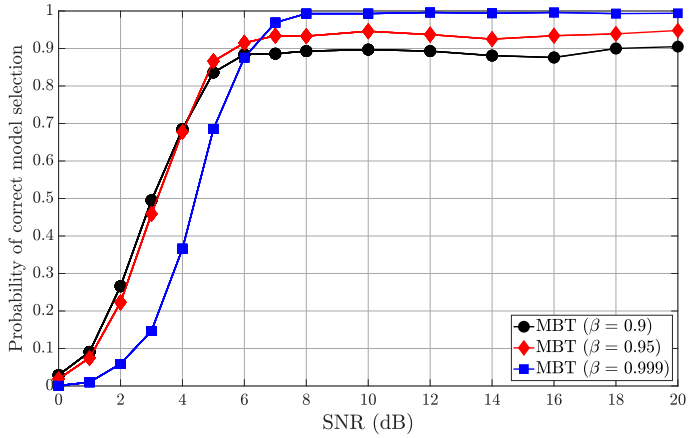


Figure 3.4: Performance of MBT versus SNR for different values of β. Here $N = 80$, $p = 500$ and $k_0 = 5$.

that the true predictors are not dropped prematurely and at the same time avoid picking noisy components. On the other hand, a small β value is equivalent to a lower threshold. This will lead to higher overfitting loss even in the large-N and high-SNR scenarios since the threshold might be too close to the noise level, thus picking up false components.

Fig. 3.3 and 3.4 presents the performance of MBT versus N and SNR, respectively for three different values of β, viz., 0.90, 0.95, and 0.999. It is clearly observed from both the figures that MBT with $\beta = 0.9,\ 0.95$, achieves higher prob-

ability of correct model selection than MBT with $\beta = 0.999$ for $N < 140$ and SNR < 6 dB for the considered data setting. However, as N and SNR increases, MBT with $\beta = 0.999$ reaches correct detection probability very close to one. But the probability of correct detection for MBT with $\beta = 0.90$ and 0.95 gets saturated around 0.90 and 0.95, respectively, and does not rise further even as N and SNR increases. These results clearly highlight that the choice of the tuning parameter can significantly affect MBT's accuracy in correctly selecting the true model.

3.3 Summary

In this chapter, a novel model selection method called Multi-Beta-Test or MBT is proposed that efficiently estimates the true support in a sparse high-dimensional linear regression model. The working principle of MBT is based on a hypothesis testing framework that uses a test statistic computed from the residual. Numerical simulations with synthetic data and performance comparison with the state-of-the-art methods have shown that MBT can provide performance similar to or slightly better than the existing methods in some scenarios with a specific value of the tuning parameter. Results also show that MBT is sensitive to the tuning parameter β, which needs to be adjusted a-priori to achieve optimal results. However, in practical scenarios, this can be a bit challenging to decide on the right value of β. This is part of future work to formulate a way for dynamically setting β in an automatic data-driven fashion.

Chapter 4

Bayesian Information Criterion - Robust

"Science is a way of thinking much more than it is a body of knowledge."
—*Carl Sagan (1934–1996)*

The Bayesian Information Criterion (BIC) is one of the most well-known criteria used for model order estimation in linear regression models. However, in its popular form, BIC is inconsistent as the noise variance tends to zero given that the sample size is small and fixed. Several modifications of the original BIC have been proposed that takes into account the high-SNR consistency, but it has been recently observed that the performance of the high-SNR forms of the BIC highly depends on the scaling of the data. This data-scaling problem is a byproduct of the data-dependent penalty design, which generates irregular penalties when the data is scaled and often leads to greater underfitting or overfitting losses in some scenarios. In this chapter, we present a new form of the BIC for order selection in linear regression models where the parameter vector dimension is small compared to the sample size. The proposed criterion eliminates the data-scaling problem and at the same time is consistent for both large sample sizes and high-SNR scenarios.

4.1 Introduction and Problem Formulation

In Chapter 3, we presented a scheme for estimating the "true" subset in a sparse high-dimensional linear regression model. In this chapter, we take a step back and re-examine the classical problem of model order selection in the linear regression setting. Model order selection is a very fundamental problem in many fields of science, engineering, and statistics wherever data fitting and prediction procedures are employed. The goal of model order selection is to estimate the best (or true) dimension of the parametric model using the observed data available. The outcome of a model order selection is an integer-value describing the dimension of the model

[66]. Consider the following linear model

$$\mathcal{H}_k : \mathbf{y} = \mathbf{A}_k \mathbf{x}_k + \mathbf{e}. \tag{4.1}$$

Here, $\mathbf{y} \in \mathbb{R}^{N\times 1}$ is the data vector, $\mathbf{A}_k \in \mathbb{R}^{N\times k}$ is a sub-design matrix formed using the first k columns of the full design matrix $\mathbf{A} \in \mathbb{R}^{N\times p}$ (which is known) where $k \leq p < N$. $\mathbf{x}_k \in \mathbb{R}^{k\times 1}$ is the corresponding unknown regression coefficient vector. $\mathbf{e} \in \mathbb{R}^{N\times 1}$ is the associated noise vector, which is assumed to be Gaussian distributed with zero mean and covariance matrix equal to $\sigma_k^2 \mathbf{I}_N$, where σ_k^2 is the unknown noise variance corresponding to hypothesis $\mathcal{H}_k$. For future analysis we use just σ^2 (without the subscript k) to signify the unknown "true" noise variance. $\mathcal{H}_k$ denotes the hypothesis that the data $\mathbf{y}$ is truly generated according to (4.1). The hypotheses $\mathcal{H}_1, \mathcal{H}_2, , \ldots, \mathcal{H}_p$ are assumed to be nested, i.e., $\mathbf{A}_k$ is a sub-block of $\mathbf{A}_{k+1}$ for $k < p$. However, note that this is not true in all cases. The integer subindex $k \in [1, p]$ in (4.1) indicates the order (or dimension) of the model. Let $k_0 (> 0)$ denote the true model order, then $x_k \neq 0$ for $k = 1, \ldots, k_0$ and $x_k = 0$ for $k = k_0 + 1, \ldots, p$. The true order k_0 (which is an integer value) is unknown and the model order selection problem involves the detection or estimation of this parameter.

A popular means of solving this problem is by using information theoretic criteria [14, 43–45]. Such a criterion assigns a score to each candidate model based on some underlying statistical principle and the model with the lowest score is selected as the final model. For the linear model in (4.1), let $p(\mathbf{y}|\boldsymbol{\theta}_k)$ denote the probability density function (pdf) of the data vector $\mathbf{y}$, where $\boldsymbol{\theta}_k$ is the vector consisting of all the unknown parameters associated with the kth candidate model. Then a general information theoretic criterion selects the model that minimizes the following metric

$$f(k) = -2 \ln p\left(\mathbf{y}\middle|\hat{\boldsymbol{\theta}}_k\right) + \mathcal{P}(k); \quad k = 1, \ldots, p \tag{4.2}$$

where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}_k$ and $\mathcal{P}(k)$ is the penalty that compensates for overparameterization. Therefore, the model order is selected as

$$\hat{k} = \underset{k\in\{1,\ldots,p\}}{\arg\min} \Big\{ f(k) \Big\}, \tag{4.3}$$

where $\hat{k}$ denotes the model order estimate. In this regard, a popular criterion for model order selection is the Bayesian information criterion (BIC), which was introduced by Schwarz [47] and is one of the most widely known and ubiquitous tools used in statistical model order selection. The BIC score for a model of order k is given as

$$\mathrm{BIC}(k) = -2 \ln p\left(\mathbf{y}\middle|\hat{\boldsymbol{\theta}}_k\right) + k \ln N, \tag{4.4}$$

where $k \ln N$ is the penalty term, which compensates for overparameterization. BIC gained popularity due to its computational simplicity and consistent performance in various fields. BIC is formulated on the maximum a-posteriori (MAP) framework of

Bayesian estimation and is a consistent criterion i.e., it will almost surely select the true model order as the sample size $N \to \infty$. However, in [67], the authors have added a new consistency requirement in the context of linear regression models, which is, given that N is small and fixed, the probability of selecting the true order should also tend to one as $\sigma^2 \to 0$. Hence, with the inclusion of this new consistency requirement, an order selection criterion is said to be consistent if it satisfies the following conditions:

$$\begin{aligned} \Pr\left\{\hat{k} = k_0\right\} &\to 1 \qquad \text{as} \qquad N \to \infty \text{ with fixed } \sigma^2 \\ \Pr\left\{\hat{k} = k_0\right\} &\to 1 \qquad \text{as} \qquad \sigma^2 \to 0 \text{ with fixed } N. \end{aligned} \tag{4.5}$$

It has been shown in [67] that when N is small and fixed, BIC is an inconsistent estimator of model order in nested model selection as $\sigma^2 \to 0$.

To circumvent this problem, the authors in [16] proposed different high signal-to-noise-ratio (SNR) forms of BIC that guarantees consistency as $\sigma^2 \to 0$. Please note that in [16], the notion of high-SNR or SNR$\to \infty$ is attributed to $\sigma^2 \to 0$. However, the high-SNR forms of the BIC proposed in [16] suffer from a scaling problem. This scaling problem arises due to the fact that the penalty contains a data-dependent term, which leads to different penalization when the data is scaled or whether $\sigma^2 > 1$ or $\sigma^2 < 1$. This is definitely not a desirable property for any model order selection criterion. In this chapter, we investigate this data-scaling problem and propose an alternate form of the BIC that eliminates this problem and at the same time is consistent as SNR $\to \infty$ as well as when $N \to \infty$.

At this point, it is important to highlight that there are other order selection criteria that obey the consistency requirements given in (4.5) and are also devoid of any scaling issues. Examples of such criteria include normalized maximum likelihood (NML) [50], g-maximum description length (gMDL) [49], exponentially embedded family (EEF) [68], penalizing adaptively the likelihood (PAL) [51]. However, our emphasis herein is on BIC partly because of the following reason: In high-dimensional models where $p \gg N$ and assuming a non-nested structure, the order selection problem can be redefined as a subset selection problem where the goal is to estimate the true support set $\mathcal{S} = \{k : x_k \neq 0\}$ in the linear model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$. In such cases, the classical model selection methods including the original BIC are ineffective and prone to overfitting [17,69,70]. However, unlike other existing methods, BIC can be extended, in a relatively straightforward manner to handle such large-p small-N scenarios. Examples of such rules include extended BIC [17] and extended Fisher information criterion [18], which we have briefly mentioned in the previous chapters.

To help the reader, we restate the notations used in this chapter. Boldface letters denote matrices and vectors. The notation $(\cdot)^T$ stands for transpose. $\mathbf{\Pi}_k = \mathbf{A}_k(\mathbf{A}_k^T\mathbf{A}_k)^{-1}\mathbf{A}_k^T$ denotes the orthogonal projection matrix on the span of $\mathbf{A}_k$ and $\mathbf{\Pi}_k^{\perp} = \mathbf{I}_N - \mathbf{\Pi}_k$ denotes the orthogonal projection matrix on the right null space of $\mathbf{A}_k^T$. $\mathbf{I}_N$ is a $N \times N$ identity matrix. The notation $|\mathbf{X}|$ denotes the determinant of the matrix $\mathbf{X}$. $X \sim \mathcal{N}(0,1)$ denotes a normal distributed random variable with mean 0

and variance 1. $X \sim \chi_k^2$ is a central chi-squared distributed random variable with k degrees of freedom, $X \sim \chi_k^2(\lambda)$ is a noncentral chi-squared distributed random variable with k degrees of freedom and non-centrality parameter λ. $X \sim \mathcal{B}(\alpha, \beta)$ is a beta distributed random variable with parameters α and β. The notation $X \xrightarrow{d} Y$ means that the random variable X converges in distribution to Y. The symbol c denotes a constant and will be used as a generic placeholder for all constant terms.

4.2 BIC and its Forms

To motivate the proposed criterion we start with a brief derivation of the original BIC that is based on the MAP estimator [16, 43, 47]. The pdf of the data vector $\mathbf{y}$ in (4.1) under hypothesis $\mathcal{H}_k$ is

$$p\left(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k\right) = \frac{1}{\left(2\pi\sigma_k^2\right)^{N/2}} \exp\left\{ -\frac{\|\mathbf{y} - \mathbf{A}_k\mathbf{x}_k\|_2^2}{2\sigma_k^2} \right\} \tag{4.6}$$

where $\boldsymbol{\theta}_k = [\mathbf{x}_k^T, \sigma_k^2]^T$ and $\|\cdot\|_2$ denotes the Euclidean norm. Under hypothesis $\mathcal{H}_k$, the MLEs of $\mathbf{x}_k$ and σ_k^2 are obtained as [63]

$$\hat{\mathbf{x}}_k = \left(\mathbf{A}_k^T\mathbf{A}_k\right)^{-1}\mathbf{A}_k^T\mathbf{y} \tag{4.7}$$

$$\hat{\sigma}_k^2 = \frac{1}{N}\|\mathbf{y} - \mathbf{A}_k\hat{\mathbf{x}}_k\|_2^2 = \frac{\mathbf{y}^T\boldsymbol{\Pi}_k^{\perp}\mathbf{y}}{N}. \tag{4.8}$$

Hence, $\hat{\boldsymbol{\theta}}_k = [\hat{\mathbf{x}}_k^T, \hat{\sigma}_k^2]^T$. Let $p(\boldsymbol{\theta}_k|\mathcal{H}_k)$ denote the prior pdf of the parameter vector $\boldsymbol{\theta}_k$ under $\mathcal{H}_k$. Then we have the joint density

$$p(\mathbf{y}, \boldsymbol{\theta}_k|\mathcal{H}_k) = p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k)p(\boldsymbol{\theta}_k|\mathcal{H}_k) \tag{4.9}$$

and the marginal distribution is

$$p(\mathbf{y}|\mathcal{H}_k) = \int p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k)p(\boldsymbol{\theta}_k|\mathcal{H}_k)d\boldsymbol{\theta}_k. \tag{4.10}$$

The posterior probability $\Pr(\mathcal{H}_k|\mathbf{y})$ is given by

$$\Pr(\mathcal{H}_k|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{H}_k)\Pr(\mathcal{H}_k)}{p(\mathbf{y})}. \tag{4.11}$$

The MAP estimator chooses the model with the largest posterior probability

$$\hat{k}_{\text{MAP}} = \underset{k\in[1,\ldots,p]}{\arg\max}\ \left\{\Pr(\mathcal{H}_k|\mathbf{y})\right\}. \tag{4.12}$$

Traditionally it is assumed that all hypotheses $\{\mathcal{H}_k\}_{k=1}^p$ of interest are equiprobable such that the prior probability of each hypothesis is $\Pr(\mathcal{H}_k) = \frac{1}{p}$ for all k. This

implies that the MAP estimate is equivalently given by maximizing $p(\mathbf{y}|\mathcal{H}_k)$ as both $\Pr(\mathcal{H}_k)$ and the marginal $p(\mathbf{y})$ are independent of k. To find the MAP estimate we need to compute the integration in (4.10). In this context, this is most commonly evaluated under the assumption that the number of samples N and/or the SNR are large enough and using the so-called Laplace's approximation that involves a second order Taylor expansion of $\ln p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k)$ around the MLE

$$\begin{aligned}\ln p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k) \approx \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{H}_k) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^T \frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k)}{\partial \boldsymbol{\theta}_k}\bigg|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k} \\ + \frac{1}{2}(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^T \left[\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k^T}\bigg|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k}\right](\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k).\end{aligned} \tag{4.13}$$

In (4.13), the first order term is zero when evaluated at the MLE. Moreover, the prior $p(\boldsymbol{\theta}_k|\mathcal{H}_k)$ is assumed to be essentially flat over the "practical support" of $p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k)$. With this we can write (4.10) as

$$p(\mathbf{y}|\mathcal{H}_k) \approx p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{H}_k) p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k) \underbrace{\int \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^T \hat{\mathbf{J}}_k (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)\right\} d\boldsymbol{\theta}_k}_{T} \tag{4.14}$$

where $\hat{\mathbf{J}}_k$ is the sample Fisher information matrix (FIM) under $\mathcal{H}_k$ evaluated at the MLE

$$\hat{\mathbf{J}}_k = -\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k)}{\partial \boldsymbol{\theta}_k \partial \boldsymbol{\theta}_k^T}\bigg|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k}. \tag{4.15}$$

Now, observe that for the linear regression model in (4.1) we have

$$-\ln p(\mathbf{y}|\boldsymbol{\theta}_k, \mathcal{H}_k) = \frac{N}{2}\sigma_k^2 + \frac{\|\mathbf{y} - \mathbf{A}_k \mathbf{x}_k\|_2^2}{2\sigma_k^2} + \frac{N}{2}\ln 2\pi. \tag{4.16}$$

Hence, obtaining the second order partial derivatives and evaluating at the MLE gives

$$\hat{\mathbf{J}}_k = \begin{bmatrix} \frac{1}{\hat{\sigma}_k^2}\mathbf{A}_k^T \mathbf{A}_k & \mathbf{0} \\ \mathbf{0} & \frac{N}{2\hat{\sigma}_k^4} \end{bmatrix}. \tag{4.17}$$

Comparing to the multivariate Gaussian pdf we can express the factor T in (4.14) as

$$T = \frac{(2\pi)^{(k+1)/2}}{\left|\hat{\mathbf{J}}_k\right|^{1/2}} \underbrace{\int \frac{e^{-\frac{1}{2}(\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)^T \hat{\mathbf{J}}_k (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k)}}{(2\pi)^{(k+1)/2}\left|\hat{\mathbf{J}}_k\right|^{-1/2}} d\boldsymbol{\theta}_k}_{=1} = \frac{(2\pi)^{(k+1)/2}}{\left|\hat{\mathbf{J}}_k\right|^{1/2}}, \tag{4.18}$$

where we assume that $\hat{\mathbf{J}}_k$ is non-singular. From (4.14) and (4.18) we get

$$\ln p(\mathbf{y}|\mathcal{H}_k) \approx \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{H}_k) + \ln p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k) + \frac{k+1}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\hat{\mathbf{J}}_k\right|. \tag{4.19}$$

Now, for the linear regression model, the pdf $p(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{H}_k)$ is

$$p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{H}_k\right) = \frac{\exp\left\{-\frac{\|\mathbf{y}-\mathbf{A}_k\hat{\mathbf{x}}_k\|_2^2}{2\hat{\sigma}_k^2}\right\}}{\left(2\pi\hat{\sigma}_k^2\right)^{N/2}} = (\hat{\sigma}_k^2)^{-N/2}(2\pi)^{-N/2}\exp\{-N/2\}$$

$$\implies -2\ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_k, \mathcal{H}_k\right) = N\ln\hat{\sigma}_k^2 + N\ln 2\pi + N. \tag{4.20}$$

Therefore, using (4.20), we can rewrite (4.19) as

$$-2\ln p(\mathbf{y}|\mathcal{H}_k) \approx N\ln\hat{\sigma}_k^2 + \ln\left|\hat{\mathbf{J}}_k\right| - 2\ln p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k) - k\ln 2\pi \underbrace{-\ln 2\pi + N\ln 2\pi + N}_{\text{constant}}. \tag{4.21}$$

The constant terms do not depend on the dimension k and hence can be neglected. Consequently, the MAP based model order estimate of the linear regression model is obtained as

$$\hat{k}_{\text{MAP}} = \underset{k\in[1,\ldots,p]}{\arg\min}\left\{N\ln\hat{\sigma}_k^2 + \ln\left|\hat{\mathbf{J}}_k\right| - 2\ln p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k) - k\ln 2\pi\right\}. \tag{4.22}$$

The transition from the above MAP criterion to BIC involves some further approximations. Firstly, the contribution from the prior term i.e., $\ln p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k)$ can be neglected if we consider it to be flat and uninformative. Traditionally, $p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k)$ is assumed to be independent of N and SNR, hence it behaves as an $\mathcal{O}(1)$ quantity. Secondly, for large sample sizes, the term $k\ln(2\pi)$ can be ignored as it becomes much smaller than the first two terms. However, note that for small sample sizes and low-SNR scenarios this term can have some significant impact on the performance, which we will see later in the proposed criterion. Thus, ignoring the term $\ln p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k)$ as well as $k\ln 2\pi$ from (4.22) we obtain what we refer to as the "fundamental" form of the BIC

$$\text{BIC}_{\text{fund}}(k) = N\ln\hat{\sigma}_k^2 + \ln\left|\hat{\mathbf{J}}_k\right|. \tag{4.23}$$

In the derivation of the original form of the BIC for linear regression models, certain assumptions are made on the sample FIM $\hat{\mathbf{J}}_k$ associated with the model. For large-N the most commonly used assumption is that (see, e.g., [16, 71])

$$\frac{1}{N}\hat{\mathbf{J}}_k = \begin{bmatrix} \frac{1}{N}\frac{\mathbf{A}_k^T\mathbf{A}_k}{\hat{\sigma}_k^2} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\hat{\sigma}_k^4} \end{bmatrix} = \mathcal{O}(1). \tag{4.24}$$

Considering $\hat{\sigma}_k^2$ to be $\mathcal{O}(1)$ for large N, clearly this assumption is equivalent to

$$\lim_{N\to\infty}\left\{\frac{\mathbf{A}_k^T\mathbf{A}_k}{N}\right\} = \mathbf{M}_k, \tag{4.25}$$

where $\mathbf{M}_k$ is a $k \times k$ positive definite matrix that is bounded as $N \to \infty$. The above assumption on the design matrix $\mathbf{A}_k$ is true in many applications but not all (see [16, 72] for more details). However, we stick with (4.25) for further analysis in this paper in order to not distract the readers from the main points of the current contribution. Now, the term $\ln\left|\hat{\mathbf{J}}_k\right|$ in (4.23) can be expanded using (4.17) as

$$\ln\left|\hat{\mathbf{J}}_k\right| = \ln\left[\frac{N}{2(\hat{\sigma}_k^2)^{k+2}}\left|\mathbf{A}_k^T\mathbf{A}_k\right|\right] = \ln(N/2) - (k+2)\ln\hat{\sigma}_k^2 + \ln\left|\mathbf{A}_k^T\mathbf{A}_k\right|. \quad (4.26)$$

Using (4.25), it is possible to show that for large N

$$\ln\left|\mathbf{A}_k^T\mathbf{A}_k\right| = \ln\left|N \cdot \frac{\mathbf{A}_k^T\mathbf{A}_k}{N}\right| = k\ln N + \mathcal{O}(1). \quad (4.27)$$

Substituting (4.26), (4.27) in (4.23) and ignoring the term $\ln(N/2)$, which is independent of k, and the $\mathcal{O}(1)$ terms (which includes $\hat{\sigma}_k^2$ as well) lead to the original and well-known form of the BIC

$$\mathrm{BIC}(k) = N\ln\hat{\sigma}_k^2 + k\ln N. \quad (4.28)$$

This form of the BIC is consistent for large sample sizes, i.e., $\lim_{N\to\infty}\Pr(\hat{k} = k_0) = 1$. However, as mentioned earlier, (4.28) is inconsistent for fixed N and increasing SNR scenarios, i.e., $\lim_{\mathrm{SNR}\to\infty}\Pr(\hat{k} = k_0) \neq 1$. To alleviate this problem, the authors in [16] proposed different high-SNR forms of the BIC, which we will briefly discuss in the next section to set the right context.

4.2.1 High-SNR Forms of BIC

In [16], the authors have argued that when dealing with small-N high-SNR scenarios, (4.28) is not the proper form of BIC. In such cases, it is important to derive the proper forms of BIC in order to correctly detect the true model order. The key assumption in [16] is that the SNR takes on large values due to the fact that the power of the noise in the data is small, where the notion of 'small' is attributed to the following

$$\hat{\sigma}_k^2 \ll 1, \quad \text{for } k \geq k_0. \quad (4.29)$$

Next, normalization of the sample FIM $\hat{\mathbf{J}}_k$ is performed with respect to $\hat{\sigma}_k^2$. To achieve this the authors in [16] have chosen the following matrix

$$\mathbf{L}_{\mathrm{SNR}}^{-1/2} = \begin{bmatrix}\sqrt{\hat{\sigma}_k^2}\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \hat{\sigma}_k^2\end{bmatrix}. \quad (4.30)$$

Now it is possible to show that for the small and fixed N case

$$\left|\mathbf{L}_{\mathrm{SNR}}^{-1/2}\hat{\mathbf{J}}_k\mathbf{L}_{\mathrm{SNR}}^{-1/2}\right| = \begin{vmatrix}\mathbf{A}_k^T\mathbf{A}_k & \mathbf{0} \\ \mathbf{0} & N/2\end{vmatrix} = \mathcal{O}(1), \quad (4.31)$$

where $\mathcal{O}(1)$ is a term independent of $\hat{\sigma}_k^2$. Using (4.31), the $\ln\left|\hat{\mathbf{J}}_k\right|$ term in $\mathrm{BIC}_{\mathrm{fund}}$ (4.23) can be expressed as

$$\begin{aligned}\ln\left|\hat{\mathbf{J}}_k\right| &= \ln\left[\left|\mathbf{L}_{\mathrm{SNR}}\right|\left|\mathbf{L}_{\mathrm{SNR}}^{-1/2}\hat{\mathbf{J}}\mathbf{L}_{\mathrm{SNR}}^{-1/2}\right|\right] \\ &= -(k+2)\ln\hat{\sigma}_k^2 + \mathcal{O}(1). \end{aligned} \tag{4.32}$$

Finally inserting (4.32) in (4.23) leads to the high-SNR form of the BIC as proposed in [16]

$$\mathrm{BIC}_{\mathrm{SNR}}(k) = \begin{cases} N\ln\hat{\sigma}_k^2 - (k+2)\ln\hat{\sigma}_k^2 & \text{if} \quad \hat{\sigma}_k^2 < 1 \\ N\ln\hat{\sigma}_k^2 & \text{otherwise} \end{cases} \tag{4.33}$$

which can be re-written compactly as

$$\mathrm{BIC}_{\mathrm{SNR}}(k) = N\ln\hat{\sigma}_k^2 + \max\{0, -(k+2)\ln\hat{\sigma}_k^2\}. \tag{4.34}$$

$\mathrm{BIC}_{\mathrm{SNR}}$ is consistent as $\sigma^2 \to 0$ for fixed N but inconsistent as $N \to \infty$ and fixed $\sigma^2 > 0$.

4.2.2 Combined Forms of BIC

At this point, it is difficult to ascertain which form of BIC to use for model order selection. If we are dealing with large-N case then we choose the standard BIC (4.28). On the other hand if it is small-N high-SNR scenario we use $\mathrm{BIC}_{\mathrm{SNR}}$ (4.34). However, in most of the real-world data, deciding whether N is large or small or similarly if SNR is high or low is not easy to deduce. In such cases, [16] proposes to choose between BIC and $\mathrm{BIC}_{\mathrm{SNR}}$ by picking the criterion with the largest penalty. This leads to the *combined form of BIC*:

$$\mathrm{BIC}_{N,\mathrm{SNR}}(k) = N\ln\hat{\sigma}_k^2 + \max\{k\ln N, -(k+2)\ln\hat{\sigma}_k^2\}. \tag{4.35}$$

Furthermore, if both N and SNR are high or if N is large but SNR is low then [16] proposes another combined form of BIC where the penalties of the two criteria BIC and $\mathrm{BIC}_{\mathrm{SNR}}$ are added together to produce a sum penalty, which leads to the following *combined form of BIC*:

$$\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}(k) = N\ln\hat{\sigma}_k^2 + k\ln N - (k+2)\ln\hat{\sigma}_k^2. \tag{4.36}$$

The above criterion (4.36) satisfies the consistency requirements in (4.5). In the next section, we discuss the data-scaling problem of the above high-SNR forms of the BIC.

4.3 Data-Scaling Problem

The penalty term in the original BIC (4.28) is independent of $\hat{\sigma}_k^2$, hence it does not suffer from any data-scaling issues. However, in the high-SNR and combined

forms of the BIC, the penalty term is dependent on the data via $\hat{\sigma}_k^2$. As such, their behaviour is irregular under changing signal and noise statistics, or in other words, they are not scale-invariant. To understand the scaling problem let us study the combined BIC criterion given in (4.36). If we scale the data $\mathbf{y}$, the noise variance $\hat{\sigma}_k^2$ will be scaled as well. For the first term (i.e., $N \ln \hat{\sigma}_k^2$) it does not matter as it will only introduce a constant that is independent of k. However, for the third term $(-(k+2) \ln \hat{\sigma}_k^2)$ it will affect the relative penalty for different k. To elaborate this point consider the difference

$$\begin{aligned}\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}(k) - \widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}(k_0) =& N \ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) + (k-k_0) \ln N - (k+2) \ln \hat{\sigma}_k^2 + \\ & (k_0+2) \ln \hat{\sigma}_{k_0}^2 \\ =& (N-2) \ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) + (k-k_0) \ln N - k \ln \hat{\sigma}_k^2 + \\ & k_0 \ln \hat{\sigma}_{k_0}^2. \end{aligned} \tag{4.37}$$

Now, if the data $\mathbf{y}$ is scaled by a constant c, the estimated noise variances are scaled by c^2 and the difference becomes

$$\begin{aligned}\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}(k) - \widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}(k_0) =& (N-2) \ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) + (k-k_0) \ln N - k \ln \hat{\sigma}_k^2 + \\ & k_0 \ln \hat{\sigma}_{k_0}^2 + (k_0-k) \ln c^2, \end{aligned} \tag{4.38}$$

where $\hat{\sigma}_k^2$ and $\hat{\sigma}_{k_0}^2$ are as before, i.e., based on the unscaled data. It is clearly observed that (4.37) and (4.38) are not equal and the difference after scaling contains an additional term $(k_0 - k) \ln c^2$. Thus, this form of the BIC is not immune to scaling issues. Same issue persists for $\mathrm{BIC}_{\mathrm{SNR}}$ and $\mathrm{BIC}_{N,\mathrm{SNR}}$ as well. Moreover, in [16], it is assumed that SNR is high because $\sigma^2 \ll 1$. But this assumption is not true always since we can have higher SNR values due to strong signal power where σ^2 may be greater than 1. The proposed high-SNR forms of the BIC in [16] do not capture this in their penalty term leading to irregular penalization under different signal and noise statistics. For example, consider a high-SNR scenario where $\sigma^2 > 1$. In this case, the penalty term of $\mathrm{BIC}_{\mathrm{SNR}}(k)$(4.34) is zero for values of $k \geq k_0$. It is clearly evident that the outcome is overfitting. A similar problem is encountered in $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}(k)$ (4.36). For high-SNR scenarios with $\sigma^2 > 1$, the $-(k+2) \ln \hat{\sigma}_k^2$ term will be negative leading to decrease in the penalty for values of $k \geq k_0$ resulting in overfitting. Hence, it is evident from the above discussion that we do need proper high-SNR forms of the BIC, which are immune to such changes in signal and noise statistics.

4.4 BIC Robust

In the previous section we saw that in order to achieve consistency as SNR $\rightarrow \infty$, the penalty term of BIC requires a data-dependent term. At this stage, we cannot completely deviate away from this structure of the BIC, but we do need slight modification to eliminate the scaling problem and maintain the consistency

requirements. From the MAP criterion in (4.22), ignoring only the prior probability term $\ln p(\hat{\boldsymbol{\theta}}_k|\mathcal{H}_k)$, we get the basic form of the BIC robust or BIC_R in short

$$\text{BIC}_\text{R}^{\text{basic}}(k) = N \ln \hat{\sigma}_k^2 + \ln \left|\hat{\mathbf{J}}_k\right| - k \ln 2\pi. \tag{4.39}$$

In order to have a scaling-free form of the BIC that is consistent for both large-N and high-SNR, the penalty term should contain quantities which are functions of both N and SNR. In this regard, we consider normalization of $\hat{\mathbf{J}}_k$ for both large-N and high-SNR scenarios. To achieve this consider the following matrix

$$\mathbf{L}^{-1/2} = \begin{bmatrix} \sqrt{\frac{1}{N}}\sqrt{\frac{\hat{\sigma}_k^2}{\hat{\sigma}_0^2}}\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \sqrt{\frac{1}{N}}\frac{\hat{\sigma}_k^2}{\hat{\sigma}_0^2} \end{bmatrix}. \tag{4.40}$$

Here, (4.40) has a similar structure as (4.30) but with an additional factor of $\hat{\sigma}_0^2$ where

$$\hat{\sigma}_0^2 = \frac{\|\mathbf{y}\|_2^2}{N}. \tag{4.41}$$

The factor $\hat{\sigma}_0^2$ in the normalization matrix $\mathbf{L}$ is used in order to eliminate the scaling problem. This is motivated by the fact that given (4.25), when SNR is a constant we have

$$\mathbb{E}[\hat{\sigma}_0^2] \to c \quad \& \quad \text{Var}[\hat{\sigma}_0^2] \to 0 \quad \text{as } N \to \infty. \tag{4.42}$$

Furthermore, from the considered generating linear model (4.1), for fixed sample sizes ($0 < N < \infty$) we have

$$\mathbb{E}[\hat{\sigma}_0^2] \to c \quad \& \quad \text{Var}[\hat{\sigma}_0^2] \to 0 \quad \text{as } \sigma^2 \to 0. \tag{4.43}$$

A detailed analysis of the factor $\hat{\sigma}_0^2$ is provided in Appendix 4.B. Now, using (4.17), (4.25) and (4.40) it is possible to show that

$$\left|\mathbf{L}^{-1/2}\hat{\mathbf{J}}_k\mathbf{L}^{-1/2}\right| = \begin{vmatrix} \frac{1}{\hat{\sigma}_0^2}\frac{\mathbf{A}_k^T\mathbf{A}_k}{N} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\hat{\sigma}_0^4} \end{vmatrix} = \mathcal{O}(1), \tag{4.44}$$

where $\mathcal{O}(1)$ is a term that is bounded as $N \to \infty$ and $\sigma_k^2 \to 0$, and therefore may be discarded with little effect on the criterion. Using (4.44), the $\ln \left|\hat{\mathbf{J}}_k\right|$ term in (4.39) can be expressed as

$$\ln \left|\hat{\mathbf{J}}_k\right| = \ln \left[\left|\mathbf{L}\right|\left|\mathbf{L}^{-1/2}\hat{\mathbf{J}}_k\mathbf{L}^{-1/2}\right|\right] = \ln |\mathbf{L}| + \mathcal{O}(1), \tag{4.45}$$

where

$$\begin{aligned}\ln|\mathbf{L}| &= \ln\begin{vmatrix} N\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_k^2}\right)\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & N\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_k^2}\right)^2 \end{vmatrix} \\ &= (k+1)\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_k^2}\right) \\ &= k\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_k^2}\right) + \ln N. \end{aligned} \tag{4.46}$$

The $\ln N$ term in (4.46) is a constant term independent of the model dimension k and can be ignored. Using (4.45) and (4.46) in (4.39) (after ignoring the constant and $\mathcal{O}(1)$ terms), we get the new form of $\mathrm{BIC_R}$ given as

$$\begin{aligned}\mathrm{BIC_R}(k) &= N\ln\hat{\sigma}_k^2 + k\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_k^2}\right) - k\ln 2\pi \\ &= (N-k-2)\ln\hat{\sigma}_k^2 + k\ln\left(\frac{N}{2\pi}\right) + (k+2)\ln\hat{\sigma}_0^2. \end{aligned} \tag{4.47}$$

The proposed new criterion $\mathrm{BIC_R}$ contains an additional quantity in the penalty term that is $(k+2)\ln\hat{\sigma}_0^2$. In this way, the ratio $\hat{\sigma}_0^2/\hat{\sigma}_k^2$ is: (1) always greater than or equal to one and (2) independent of the scaling of $\mathbf{y}$. The first property is perhaps not necessary but seems reasonable if one views the penalty terms separately (the one dependent on N and the one dependent on $\hat{\sigma}_k^2$) and require them to be non-negative. The scaling independence is clearly desirable. The ratio $\hat{\sigma}_0^2/\hat{\sigma}_k^2$ is $\mathcal{O}(1)$ for $k < k_0$ and $\mathcal{O}(\mathrm{SNR})$ for $k \geq k_0$ if $\mathrm{SNR} \gg 1$. Moreover, note that $\hat{\sigma}_0^2$ is on the order of unity, i.e., the same order as terms we typically have neglected in the case of BIC and high-SNR forms of the BIC. It also leads to a penalty that is a function of SNR and performs no matter whether the SNR is high because the signal is strong or because the noise is weak. There are of course other alternatives possible with at least partially similar properties, e.g., using $\mathbf{y}^T\mathbf{\Pi}_k\mathbf{y}/N$ instead of $\hat{\sigma}_0^2$. This would lead to a penalty involving an SNR measure and be independent of scaling but will not guarantee a positive penalty term associated with the SNR (the ratio of $\mathbf{y}^T\mathbf{\Pi}_k\mathbf{y}/N$ and $\hat{\sigma}_k^2$ may typically take on small values for small k).

Next, we present the following analysis highlighting the signal scaling immunity of $\mathrm{BIC_R}$. Consider the difference

$$\begin{aligned}\mathrm{BIC_R}(k) - \mathrm{BIC_R}(k_0) =& (N-2)\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) + (k-k_0)\ln(N/2\pi) - k\ln\hat{\sigma}_k^2 + \\ & k_0\ln\hat{\sigma}_{k_0}^2 + (k-k_0)\ln\hat{\sigma}_0^2. \end{aligned} \tag{4.48}$$

In a similar fashion, if the data $\mathbf{y}$ is scaled by c, the estimated noise variances ($\hat{\sigma}_k^2$ and $\hat{\sigma}_{k_0}^2$) as well as the term $\hat{\sigma}_0^2$ are scaled by c^2. Therefore, in this case the

difference becomes

$$\begin{aligned}\mathrm{BIC_R}(k) - \mathrm{BIC_R}(k_0) =&(N-2)\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) + (k-k_0)\ln(N/2\pi) - k\ln\hat{\sigma}_k^2 + \\ &k_0\ln\hat{\sigma}_{k_0}^2 + (k-k_0)\ln\hat{\sigma}_0^2 + (k_0-k)\ln c^2 - (k_0-k)\ln c^2 \\ =&(N-2)\ln(\hat{\sigma}_k^2/\hat{\sigma}_{k_0}^2) + (k-k_0)\ln(N/2\pi) - k\ln\hat{\sigma}_k^2 + \\ &k_0\ln\hat{\sigma}_{k_0}^2 + (k-k_0)\ln\hat{\sigma}_0^2. \qquad (4.49)\end{aligned}$$

It is clearly observed that the unscaled (4.48) and scaled (4.49) differences are equal, which underlines the advantage of the proposed criterion.

Finally, note that as compared to the fundamental form of the BIC (4.23), here in $\mathrm{BIC_R}$ we have considered retaining a term that was previously ignored under large sample and/or high-SNR approximations. This term is "$k\ln 2\pi$" whose significance is small when N is large and/or SNR is high. However, in the numerical analysis performed, it is found that this term can play an important role in improving the PCOS in low regions of N and/or SNR. Since for large-N and/or high-SNR scenarios, the effect of this term is negligible, there is no harm in keeping this term in the criterion. Moreover, in real scenarios deciding if the data length is large or if we are dealing with a high-SNR case is not possible to ascertain with precision. Hence, having this term in the criterion can be beneficial in certain cases without hampering performance.

4.5 Proof of Consistency

In this section, we provide the necessary proofs to show that $\mathrm{BIC_R}$ is a consistent criterion, i.e., it satisfies (4.5). First, we show consistency with respect to increasing SNR and then with respect to increasing sample size. Note that $\mathrm{BIC_R}$ falls under the class of methods discussed in [71,73]. Hence, the high-SNR consistency of $\mathrm{BIC_R}$ then follows from the analyses in [71, 73]. However, for completeness, we provide proof of consistency for both high-SNR and large-N cases.

4.5.1 Consistency as $\sigma^2 \to 0$ or SNR $\to \infty$ for fixed N

Here we evaluate the consistency of $\mathrm{BIC_R}$ as the true noise variance σ^2 becomes vanishingly small. We consider two cases. The first is the overfitting scenario where we investigate the probability of $\mathrm{BIC_R}$ choosing more than k_0 variables. Second is the underfitting case where we evaluate the probability of $\mathrm{BIC_R}$ choosing less than k_0 variables as $\sigma^2 \to 0$.

Over-Fitting Case

Let us compute the probability that $\mathrm{BIC_R}$ would prefer a model of order $(k_0 + i)$ where $0 < i \leq p - k_0$, to the model of order k_0 as $\sigma^2 \to 0$, i.e., $\Pr\{\mathrm{BIC_R}(k_0) >$

$\mathrm{BIC_R}(k_0+i)\}$, which after some rewriting gives

$$\Pr\Big\{(N-k_0-2)\ln\hat{\sigma}^2_{k_0}-(N-k_0-i-2)\ln\hat{\sigma}^2_{k_0+i}-i\ln(N/2\pi)-i\ln(\hat{\sigma}^2_0)>0\Big\}. \quad (4.50)$$

Now let us define the random variable

$$X_k=\frac{\hat{\sigma}^2_k}{\sigma^2}, \quad \text{then} \quad N\cdot X_k\sim\mathcal{X}^2_{N-k} \quad \forall k\geq k_0. \quad (4.51)$$

See Appendix 4.C for details. Hence, $\mathbb{E}[X_k]=(N-k)/N$ and $\mathrm{Var}[X_k]=2(N-k)/N^2$. This implies that the variables X_k are independent of σ^2 for $k\geq k_0$. Now, we can express

$$(N-k_0-2)\ln\hat{\sigma}^2_{k_0}=\ln X_{k_0}^{N-k_0-2}+(N-k_0-2)\ln\sigma^2, \quad (4.52)$$

and similarly,

$$(N-k_0-i-2)\ln\hat{\sigma}^2_{k_0+i}=\ln X_{k_0+i}^{N-k_0-i-2}+(N-k_0-i-2)\ln\sigma^2. \quad (4.53)$$

Therefore, we can rewrite (4.50) as

$$\Pr\left\{\ln\left(\frac{X_{k_0}^{N-k_0-2}}{X_{k_0+i}^{N-k_0-i-2}}\right)+\ln(2\pi/N)^i+\ln(1/\hat{\sigma}^2_0)^i>\ln(1/\sigma^2)^i\right\}, \quad (4.54)$$

which after exponentiation gives

$$\Pr\left\{\left(\frac{X_{k_0}^{N-k_0-2}}{X_{k_0+i}^{N-k_0-i-2}}\right)\left(\frac{2\pi}{N}\right)^i\left(\frac{1}{\hat{\sigma}^2_0}\right)^i>\left(\frac{1}{\sigma^2}\right)^i\right\}. \quad (4.55)$$

Let the random variable Y denote the entire left-hand side of (4.55) and let $W=X_{k_0}^{N-k_0-2}/X_{k_0+i}^{N-k_0-i-2}$. From (4.43) we have $\lim_{\sigma^2\to 0}\hat{\sigma}^2_0=c$. This implies that the random variable $Y\xrightarrow{d}cW$ as $\sigma^2\to 0$ where $c>0$ is some constant. The random variable W is independent of σ^2 and the right-hand side of (4.55) grows unbounded as $\sigma^2\to 0$. Hence, the probability of overfitting (4.55) tends to zero as $\sigma^2\to 0$, i.e.,

$$\lim_{\sigma^2\to 0}\Pr\Big\{\mathrm{BIC_R}(k_0)>\mathrm{BIC_R}(k_0+i)\Big\}=0. \quad (4.56)$$

Under-Fitting Case

We compute the probability that $\mathrm{BIC_R}$ would prefer a model of order (k_0-i) where $0<i<k_0$, to the model of order k_0 as $\sigma^2\to 0$, i.e., $\Pr\{\mathrm{BIC_R}(k_0)>\mathrm{BIC_R}(k_0-i)\}$, which after some rewriting gives

$$\Pr\Big\{(N-k_0-2)\ln\hat{\sigma}^2_{k_0}-(N-k_0+i-2)\ln\hat{\sigma}^2_{k_0-i}+i\ln(N/2\pi)+i\ln(\hat{\sigma}^2_0)>0\Big\}. \quad (4.57)$$

Now using (4.51) in (4.57) and after exponentiation we get

$$\Pr\left\{\left(\frac{X_{k_0}^{N-k_0-2}}{\left(\hat{\sigma}^2_{k_0-i}\right)^{N-k_0+i-2}}\right)\left(\frac{N}{2\pi}\right)^{i}\left(\hat{\sigma}_0^2\right)^{i} > \left(\frac{1}{\sigma^2}\right)^{N-k_0-2}\right\}. \tag{4.58}$$

Let the random variable Y denote the entire left-hand side of (4.58). From the properties of least-squares estimates we have (proof is provided in Appendix 4.C)

$$\lim_{\sigma^2\to 0}\mathbb{E}\left[\hat{\sigma}_k^2\right] = c \quad \& \quad \lim_{\sigma^2\to 0}\mathrm{Var}\left[\hat{\sigma}_k^2\right] = 0, \quad \forall k < k_0. \tag{4.59}$$

Also the random variable $X_{k_0}^{N-k_0-2}$ is independent of σ^2 and $\lim_{\sigma^2\to 0}\hat{\sigma}_0^2 = c$ (4.43). This implies that the random variable $Y \xrightarrow{d} cX_{k_0}^{N-k_0-2}$ as $\sigma^2 \to 0$ and the right-hand side of (4.58) grows unbounded as $\sigma^2 \to 0$. Thus, the probability of underfitting (4.58) tends to zero as $\sigma^2 \to 0$, i.e.,

$$\lim_{\sigma^2\to 0}\Pr\left\{\mathrm{BIC_R}(k_0) > \mathrm{BIC_R}(k_0-i)\right\} = 0. \tag{4.60}$$

Thus, from (4.56) and (4.60) we conclude that $\mathrm{BIC_R}$ satisfies $\Pr\{\hat{k} = k_0\} \to 1$ as $\sigma^2 \to 0$.

4.5.2 Consistency as $N \to \infty$ for Fixed σ^2 $(0 < \sigma^2 < \infty)$

In this section, we evaluate the consistency of $\mathrm{BIC_R}$ as the sample size N grows large. As in the previous section, we consider two cases. The first is the overfitting scenario where we investigate the probability of $\mathrm{BIC_R}$ choosing more than k_0 variables. Second, the underfitting case where we evaluate the probability of $\mathrm{BIC_R}$ choosing less than k_0 variables as $N \to \infty$ when the true noise variance σ^2 is fixed.

Over-Fitting Case

We compute the probability that $\mathrm{BIC_R}$ will choose a model of order (k_0+i) to the model of order k_0 as $N \to \infty$, i.e., $\Pr\{\mathrm{BIC_R}(k_0+i) < \mathrm{BIC_R}(k_0)\}$, $0 < i \le p-k_0$, which after some rewriting gives

$$\Pr\left\{(N-k_0-i-2)\ln\left(\frac{\hat{\sigma}^2_{k_0+i}}{\hat{\sigma}^2_{k_0}}\right) + i\ln\left(\frac{N}{2\pi}\right) + i\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2_{k_0}}\right) < 0\right\}. \tag{4.61}$$

Let us denote $X = \left(\frac{\hat{\sigma}^2_{k_0+i}}{\hat{\sigma}^2_{k_0}}\right)$, then $X \sim \mathcal{B}\left(\frac{N-k_0-i}{2}, \frac{i}{2}\right)$ (the proof is given in Lemma 4.3 of Appendix 4.A). Now define another random variable $Z = 1 - X$, then $Z \sim \mathcal{B}\left(\frac{i}{2}, \frac{N-k_0-i}{2}\right)$ (mirror-image symmetry). Hence,

$$\mathbb{E}[Z] = \frac{1}{1+\beta/\alpha} = \mathcal{O}\left(\frac{1}{N}\right) \quad \text{and} \quad \mathrm{Var}[Z] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \mathcal{O}\left(\frac{1}{N^2}\right), \tag{4.62}$$

where $\alpha = \frac{i}{2}$ and $\beta = \frac{N-k_0-i}{2}$. Now, for large N case we have

$$(N - k_0 - i - 2)\ln X \approx N\ln(1 - Z) \approx -NZ, \tag{4.63}$$

where the last approximation follows by linearization of the logarithm for small Z. From above it follows that

$$\lim_{N\to\infty} \mathbb{E}[NZ] = i \quad \text{and} \quad \lim_{N\to\infty} \mathrm{Var}[NZ] = i/2. \tag{4.64}$$

Hence, for large N, the term $(N - k_0 - i - 2)\ln\left(\hat{\sigma}^2_{k_0+i}/\hat{\sigma}^2_{k_0}\right)$ is approximately a random variable with finite mean and variance. Furthermore, $\lim_{N\to\infty} \hat{\sigma}^2_0 = c > 0$ (see Appendix 4.B) and $\lim_{N\to\infty} \hat{\sigma}^2_{k_0} = \sigma^2$ (see Appendix 4.C). Then we get

$$\lim_{N\to\infty} \Pr\left\{(N - k_0 - i - 2)\ln X + i\ln\left(\frac{N}{2\pi}\right) + i\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{k_0}}\right) < 0\right\} = 0, \tag{4.65}$$

because the left-hand side of the inequality grows with $\ln N$. Thus, the probability of overfitting tends to zero as $N \to \infty$, i.e.,

$$\lim_{N\to\infty} \Pr\left\{\mathrm{BIC_R}(k_0 + i) < \mathrm{BIC_R}(k_0)\right\} = 0. \tag{4.66}$$

Under-Fitting Case

Let us compute the probability that $\mathrm{BIC_R}$ would prefer a model of order $(k_0 - i)$ to the model of order k_0 as $N \to \infty$, i.e., $\Pr\{\mathrm{BIC_R}(k_0 - i) < \mathrm{BIC_R}(k_0)\}$ where $0 < i < k_0$. After some rewriting, we get

$$\Pr\left\{(N - k_0 + i - 2)\ln\left(\frac{\hat{\sigma}^2_{k_0-i}}{\hat{\sigma}^2_{k_0}}\right) - i\ln\left(\frac{N}{2\pi}\right) - i\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{k_0}}\right) < 0\right\}. \tag{4.67}$$

From Appendix 4.B and 4.C we have $\lim_{N\to\infty} \hat{\sigma}^2_0 = c$ and $\lim_{N\to\infty} \hat{\sigma}^2_{k_0} = \sigma^2$ respectively. Therefore, the random variable $\hat{\sigma}^2_0/\hat{\sigma}^2_{k_0} \xrightarrow{d} c > 0$ as $N \to \infty$. Furthermore, consider the sub-matrix $\mathbf{A}_i$ such that $[\mathbf{A}_{k_0-i},\ \mathbf{A}_i] = \mathbf{A}_{k_0}$. Then, we have

$$\begin{aligned}
\frac{\mathbf{y}^T\mathbf{\Pi}^{\perp}_{k_0}\mathbf{y}}{N} &= \frac{\mathbf{y}^T\left(\mathbf{I}_N - \mathbf{\Pi}_{k_0-i} - \mathbf{\Pi}_{\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i}\right)\mathbf{y}}{N} \\
&= \frac{\mathbf{y}^T\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{y}}{N} - \frac{\mathbf{y}^T\mathbf{\Pi}_{\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i}\mathbf{y}}{N} \\
\implies \hat{\sigma}^2_{k_0-i} &= \hat{\sigma}^2_{k_0} + C,
\end{aligned} \tag{4.68}$$

where $C = \frac{\mathbf{y}^T\mathbf{\Pi}_{\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i}\mathbf{y}}{N} > 0$ and $\mathbf{\Pi}_{\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i}$ is the projection on to the space spanned by $\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i$ defined as

$$\mathbf{\Pi}_{\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i} = \mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i\left(\mathbf{A}_i^T\mathbf{\Pi}^{\perp}_{k_0-i}\mathbf{A}_i\right)^{-1}\mathbf{A}_i^T\mathbf{\Pi}^{\perp}_{k_0-i} \tag{4.69}$$

and $\mathbf{\Pi}^{\perp}_{k_0-i}$ is the orthogonal projection matrix on the right null space of $\mathbf{A}^T_{k_0-i}$. Hence, from (4.68) we can say that $\hat{\sigma}^2_{k_0-i} > \hat{\sigma}^2_{k_0}$ for $0 < i < k_0$. Thus,

$$\ln\left(\frac{\hat{\sigma}^2_{k_0-i}}{\hat{\sigma}^2_{k_0}}\right) = \ln\left(\frac{\hat{\sigma}^2_{k_0}+C}{\hat{\sigma}^2_{k_0}}\right) = \ln\left(1+C'\right), \tag{4.70}$$

where $C' = C/\hat{\sigma}^2_{k_0} > 0$. This implies that

$$\lim_{N\to\infty} \Pr\left\{ N\ln(1+C') - i\ln\left(\frac{N}{2\pi}\right) - i\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{k_0}}\right) < 0 \right\} = 0, \tag{4.71}$$

since $N\ln(1+C')$ is the dominating term here as it tends to infinity much faster than the $\ln N$ term. Hence, the probability of underfitting for $\mathrm{BIC_R}$ tends to zero as the sample size grows large, i.e.,

$$\lim_{N\to\infty} \Pr\left\{ \mathrm{BIC_R}(k_0-i) < \mathrm{BIC_R}(k_0) \right\} = 0. \tag{4.72}$$

Therefore, from (4.66) and (4.72) we can conclude that $\mathrm{BIC_R}$ satisfies $\Pr\{\hat{k} = k_0\} \to 1$ as $N \to \infty$.

4.6 Simulation Results

In this section, we provide numerical simulation results to analyze the behaviour of the proposed scale-invariant consistent criterion $\mathrm{BIC_R}$ and compare its statistical performance with the other forms of BIC as well as with NML, gMDL, and PAL. Next, we briefly present these criteria before moving to the details of the simulations and results.

4.6.1 Existing Popular High-SNR Criteria for Order Selection

1. The **NML** criterion, derived in [50], and it is given by

$$\mathrm{NML}(k) = (N-k)\ln(\hat{\sigma}^2_k) + k\ln(\hat{R}_k) + (N-k-1)\ln\left(\frac{N}{N-k}\right) - (k+1)\ln k \tag{4.73}$$

where

$$\hat{R}_k = \mathbf{y}^T\mathbf{y} - N\hat{\sigma}^2_k = \mathbf{y}^T\mathbf{\Pi}_k\mathbf{y} \tag{4.74}$$

is the fitted sum of squares. It is shown in [71] that NML is consistent when $\sigma^2 \to 0$.

2. The **gMDL** criterion, developed by Hansen and Yu [49], is based on the Bayesian mixture form of MDL. It is called gMDL for its use of the g-prior

and is given by

$$\text{gMDL}(k) = \left(\frac{N-k}{2}\right) \ln\left(\frac{N\hat{\sigma}_k^2}{N-k}\right) + \frac{k}{2}\ln\left(\frac{\hat{R}_k}{k}\right) + \ln N, \tag{4.75}$$

where $\hat{R}_k$ is as given in (4.74). It has been shown in [71] that gMDL is consistent as $\sigma^2 \to 0$.

3. The **PAL** criterion was developed by Stoica and Babu [51] and is given by

$$\text{PAL}(k) = N \ln \hat{\sigma}_k^2 + k \ln(p) \frac{\ln(r_k + 1)}{\ln(\rho_k + 1)} \tag{4.76}$$

where

$$r_k = N \ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{k-1}^2}\right) \quad ; \quad \rho_k = N \ln\left(\frac{\hat{\sigma}_{k-1}^2}{\hat{\sigma}_p^2}\right), \tag{4.77}$$

and $\hat{\sigma}_0^2$ as given in (4.41). The primary motivation behind the design of the PAL rule is to achieve the following properties for the penalty term in both large-N and high-SNR case:

   a) for $k \leq k_0$ the penalty term should be small enough so that the function in (4.76) decreases with increasing k, and
   b) for $k > k_0$ the penalty term should be sufficiently large such that the function in (4.76) increases with increasing k.

4.6.2 General Simulation Setup

In the simulations, we consider the linear regression model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where the design matrix $\mathbf{A} \in \mathbb{R}^{N \times p}$ is generated with independent entries following normal distribution $\mathcal{N}(0, 1)$. The SNR in dB is SNR (dB) $= 10 \log_{10}(\sigma_s^2/\sigma^2)$, where σ_s^2 and σ^2 denote signal and true noise power, respectively. The signal power is computed as $\sigma_s^2 = ||\mathbf{A}_{k_0}\mathbf{x}_{k_0}||_2^2/N$. Based on σ_s^2 and the chosen SNR (dB), the noise power is set as $\sigma^2 = \sigma_s^2/10^{\text{SNR (dB)}/10}$. Using this σ^2, the noise vector $\mathbf{e}$ is generated following $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. The probability of correct order selection (PCOS) is estimated over 5000 Monte Carlo trials. To maintain randomness in the data, a new design matrix $\mathbf{A}$ is generated at each Monte Carlo trial.

4.6.3 Model Order Selection versus SNR

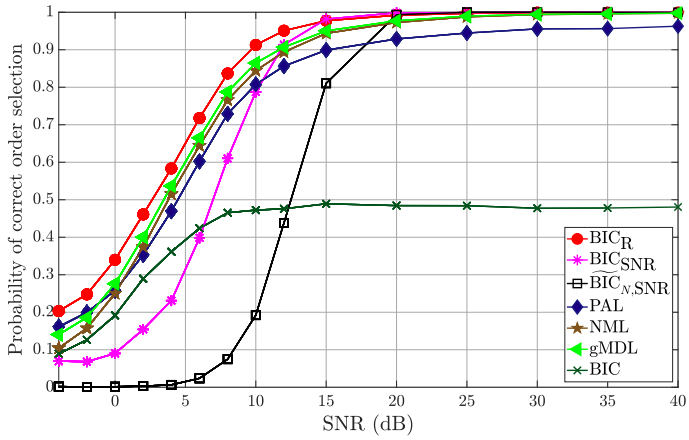
The first simulation is analyzing the behavior of the methods as a function of SNR when N is small and fixed. To particularly highlight the data-scaling problem and visualize the behaviour of other forms of BIC with respect to BIC_R, we consider two scenarios. In the first scenario we assume the true regression coefficient vector to be $\mathbf{x}_{k_0} = [0.1, 0.1, 0.1, 0.1, 0.1]^T$. Thus, the data is generated as

$$\mathbf{y} = 0.1\mathbf{a}_1 + 0.1\mathbf{a}_2 + 0.1\mathbf{a}_3 + 0.1\mathbf{a}_4 + 0.1\mathbf{a}_5 + \mathbf{e}. \tag{4.78}$$

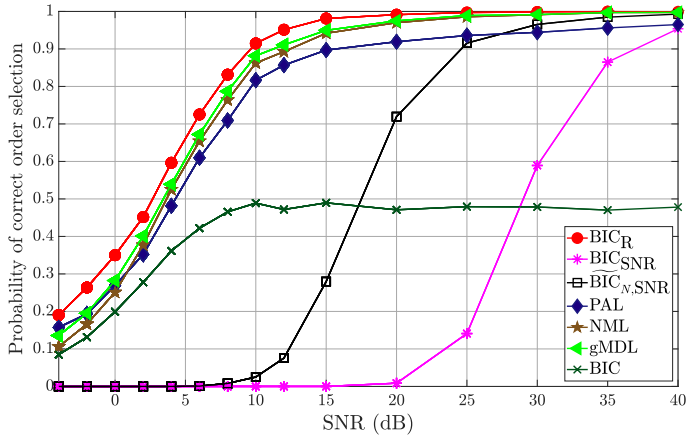
In the second scenario, we assume the true regression coefficient vector to be $\mathbf{x}_{k_0} = [10, 10, 10, 10, 10]^T$. Thus, in this case the data is generated as

$$\mathbf{y} = 10\mathbf{a}_1 + 10\mathbf{a}_2 + 10\mathbf{a}_3 + 10\mathbf{a}_4 + 10\mathbf{a}_5 + \mathbf{e}, \tag{4.79}$$

where $\mathbf{e}$ is the Gaussian noise vector (independent of $\mathbf{a}_k$) with zero mean and covariance matrix $\sigma^2\mathbf{I}_N$. Note that in the simulations we compute σ^2 based on the chosen SNR level and the current signal power value $\sigma_s^2 = \|\mathbf{A}_{k_0}\mathbf{x}_{k_0}\|_2^2/N$. Hence, both models (4.78) and (4.79) will have the same SNR. Furthermore, for both the models, $k_0 = 5$, $p = 10$ and the regressor coefficients $x_k = 0$ for all $k > k_0$.



(a) $\mathbf{x}_{k_0} = [0.1, 0.1, 0.1, 0.1, 0.1]^T$



(b) $\mathbf{x}_{k_0} = [10, 10, 10, 10, 10]^T$

Figure 4.1: The PCOS versus SNR (dB) for $N = 15$, $p = 10$ and $k_0 = 5$.

Fig. 4.1 illustrates the PCOS versus SNR (dB) for $N = 15$. This is a small-N increasing SNR scenario. Fig. 4.1a corresponds to the first case where $\mathbf{x}_{k_0} = [0.1, 0.1, 0.1, 0.1, 0.1]^T$ and Fig. 4.1b corresponds to the second case with $\mathbf{x}_{k_0} = [10, 10, 10, 10, 10]^T$. Comparing Fig. 4.1a and Fig. 4.1b, the first major observation is that the behavior of $\mathrm{BIC_R}$, PAL, NML and gMDL is robust to the scaling and immune to such changing signal and noise statistics as compared to the other high-SNR forms of BIC viz. $\mathrm{BIC_{SNR}}$ and $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$. PAL is expected to perform better than BIC, which can be observed here, but its convergence rate (order of $\ln\ln(\mathrm{SNR})$) is very slow when N is low and SNR $\to \infty$. Except for BIC, which we know is inconsistent as SNR $\to \infty$, and PAL, which has a slow convergence rate, the PCOS for all other criteria approaches one as SNR increases. The poor performance of $\mathrm{BIC_{SNR}}$ and $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$ compared to the other methods in Fig. 4.1a and Fig. 4.1b can be explained as follows: The entries in the design matrix $\mathbf{A}$ are drawn from $\mathcal{N}(0, 1)$. When $\mathbf{x}_{k_0} = [0.1, 0.1, 0.1, 0.1, 0.1]$, it is more likely that the signal power $\sigma_s^2 = \|\mathbf{A}_{k_0}\mathbf{x}_{k_0}\|_2^2/N < 1$. This implies that for SNR > 0 dB, $\sigma^2 \ll 1$ and as such the $-(k+2)\ln\hat{\sigma}_k^2$ term in the penalty of both $\mathrm{BIC_{SNR}}$ and $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$ is positive and may become quite large in value, thus producing a much bigger overall penalty, which leads to underfitting issues. On the contrary, for $\mathbf{x}_{k_0} = [10, 10, 10, 10, 10]^T$, it is more likely that $\sigma_s^2 \gg 1$ and therefore $\sigma^2 > 1$ for a wider range of SNR ≥ 0 dB. This implies that the penalty of $\mathrm{BIC_{SNR}}$ is 0 and the $-(k+2)\ln\hat{\sigma}_k^2$ term of $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$ is negative for values of SNR ≥ 0 dB. This lowers the overall penalty of $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$ and leads to overfitting issues, whose effect we see in Fig. 4.1b.

4.6.4 Model Order Selection versus N

In analyzing the performance of the different criteria as a function of N, we consider a relatively harder case where the regression coefficients are unequal and in decreasing order of amplitude. Generally in such situations, there is a chance that most of the order selection rules will have a higher tendency to underfit. In the first scenario, we assume the true regression coefficient vector to be $\mathbf{x}_{k_0} = [0.5, 0.4, 0.3, 0.2, 0.1]^T$ and the data generation is as follows

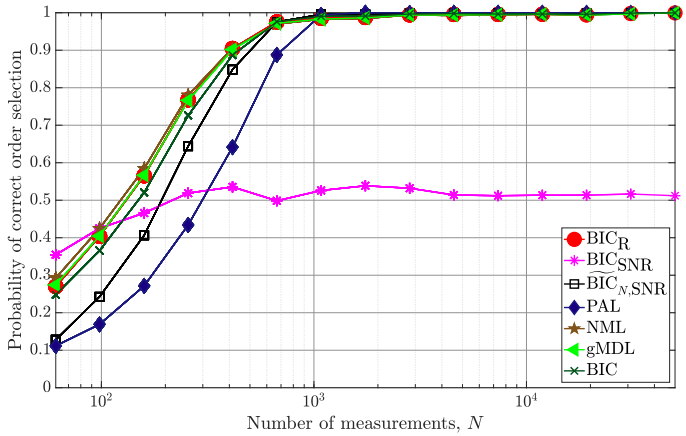
$$\mathbf{y} = 0.5\mathbf{a}_1 + 0.4\mathbf{a}_2 + 0.3\mathbf{a}_3 + 0.2\mathbf{a}_4 + 0.1\mathbf{a}_5 + \mathbf{e}. \tag{4.80}$$

In the second scenario, we assume the true regression coefficient vector to be $\mathbf{x}_{k_0} = [50, 40, 30, 20, 10]^T$. Thus, in this case the data is generated as

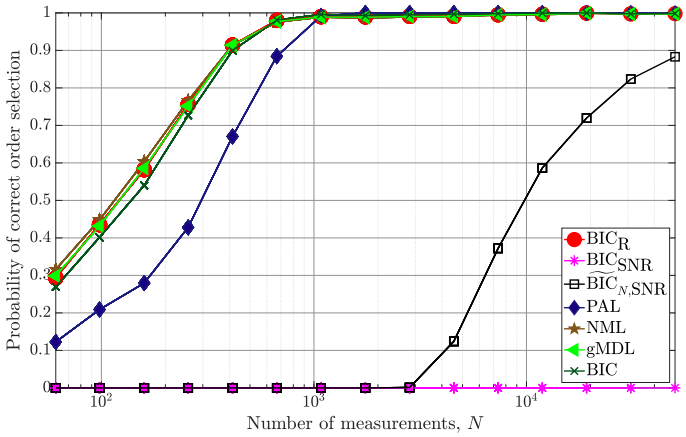
$$\mathbf{y} = 50\mathbf{a}_1 + 40\mathbf{a}_2 + 30\mathbf{a}_3 + 20\mathbf{a}_4 + 10\mathbf{a}_5 + \mathbf{e}. \tag{4.81}$$

Similarly, for both the models $p = 10$, $k_0 = 5$ and $x_k = 0$ for all $k > k_0$.

Fig. 4.2 shows the PCOS vs N for a fixed SNR of 3 dB. This is a low-SNR increasing N scenario. Fig. 4.2a corresponds to $\mathbf{x}_{k_0} = [0.5, 0.4, 0.3, 0.2, 0.1]^T$ and Fig. 4.2b corresponds to $\mathbf{x}_{k_0} = [50, 40, 30, 20, 10]^T$. Comparing both the figures, the first clear observation is that $\mathrm{BIC_{SNR}}$ is inconsistent, which is obvious since it is



(a) $\mathbf{x}_{k_0} = [0.5, 0.4, 0.3, 0.2, 0.1]^T$



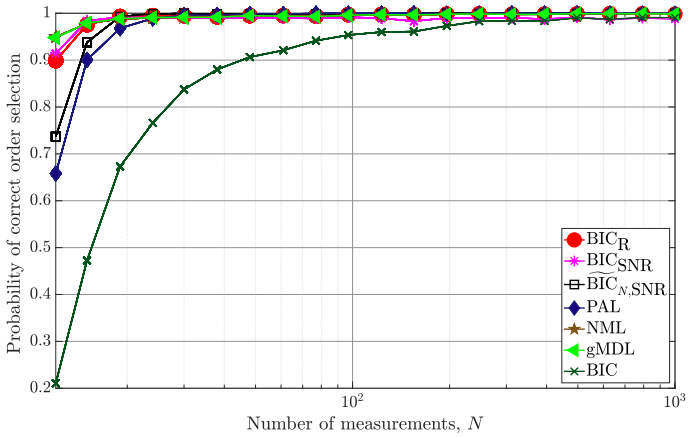
(b) $\mathbf{x}_{k_0} = [50, 40, 30, 20, 10]^T$

Figure 4.2: The PCOS versus N for SNR = 3 dB, $p = 10$ and $k_0 = 5$.

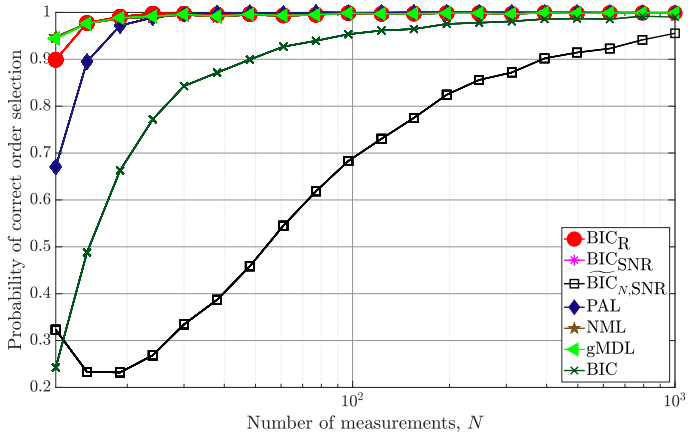
not designed to handle large sample sizes under low and fixed SNR cases. Secondly, the PCOS for BIC$_{\mathrm{R}}$, PAL, NML, gMDL, and BIC approaches one as N increases. Furthermore, performances of BIC$_{\mathrm{R}}$, NML, and gMDL are quite similar for this case and they undergo the minimum underfitting loss compared to the other methods. Also, observe that the behaviour of BIC$_{\mathrm{R}}$ and BIC are quite close. This is because for low-SNR conditions (3 dB in this case), the $(k+2)\ln(\hat{\sigma}_0^2/\hat{\sigma}_k^2)$ term of BIC$_{\mathrm{R}}$ behaves very close to a $\mathcal{O}(1)$ quantity for $k \geq k_0$. This implies that the $k \ln N$ term dominates the penalty, which is the same for BIC as well and as such the behaviour of BIC and BIC$_{\mathrm{R}}$ are very similar here. The performance of $\widetilde{\text{BIC}}_{N,\mathrm{SNR}}$ is at par with

the other methods for the case when $\mathbf{x}_{k_0} = [0.5, 0.4, 0.3, 0.2, 0.1]^T$. However, due to its inability to handle the signal scaling problem, $\widetilde{\text{BIC}}_{N,\text{SNR}}$ struggles to achieve convergence when $\mathbf{x}_{k_0} = [50, 40, 30, 20, 10]^T$ and requires quite a large sample size to reach PCOS = 1.

As mentioned before, Fig. 4.1 and Fig. 4.2 showcased the scenarios corresponding to small-N increasing SNR and low-SNR increasing N, respectively. In the final figure, Fig. 4.3, we present a high-SNR increasing N scenario where the plot shows the PCOS versus N for a fixed SNR of 25 dB. All other parameters are the same as in Fig. 4.2. Comparing Fig. 4.3a and Fig. 4.3b , the first clear observation is



(a) $\mathbf{x}_{k_0} = [0.5, 0.4, 0.3, 0.2, 0.1]^T$



(b) $\mathbf{x}_{k_0} = [50, 40, 30, 20, 10]^T$

Figure 4.3: The PCOS versus N for SNR = 25 dB, $p = 10$ and $k_0 = 5$.

that when the SNR is relatively high, BIC_R provides a better convergence rate with increasing N as compared to BIC in particular and in some cases of the high-SNR forms of the BIC (Fig. 4.3b). Secondly, the behaviour of BIC_R, gMDL, and NML are more or less similar in this scenario (as well as in the previous scenarios see Fig. 4.1 and Fig. 4.2) and as such no conclusion can be made as to who is the winner among them. However, as mentioned before, when it comes to high-dimensional data with large parameter dimension, the BIC framework can prove to be a major advantage compared to the existing methods of model selection.

4.6.5 Remarks from Simulation Results

The main points of discussion from the simulation results are as follows:

- BIC_R is consistent estimator of the true model order in nested model selection when N is fixed and $\text{SNR} \to \infty$ as well as when SNR is constant and $N \to \infty$.
- BIC_R completely eliminates the data-scaling problem.
- In case of the proposed high-SNR forms of BIC in [16], to know which form of the BIC to apply for model selection we require knowledge of whether we are dealing with a small-N, high-SNR scenario, or large-N, low-SNR scenario. This information is hard to extract from the data which makes it difficult to choose the right BIC criterion. On the contrary, BIC_R mitigates this problem. From the numerical simulations, it is observed that the average performance of BIC_R is pretty robust in small-N, high-SNR as well as in large-N, low-SNR scenarios. Hence, BIC_R does not require such prior information or the need to extract such information from the data. This makes BIC_R a more versatile model order selection criterion.
- In terms of performance, the proposed BIC_R clearly surpasses all the other forms of BIC considered in the analysis. However, as compared to gMDL and NML (which are derived based on other frameworks/arguments) the difference in performance is small. Nevertheless, with the current upgradation of the BIC framework, it is now able to compete with state-of-the-art methods.

4.7 Summary

In this chapter, we discussed the data-scaling problem present in the high-SNR forms of the BIC in the context of order selection for linear regression models. These high-SNR forms are not scale-invariant due to the data-dependent penalty design, leading to unstable behaviour under different signal and noise statistics. To resolve this scaling issue, we proposed a new form of the BIC named as BIC_R (where the subscript R stands for robust) by modifying the existing high-SNR forms of the BIC. BIC_R is scale-invariant and resilient to the dynamics of signal and noise statistics. Numerical simulations with synthetic data verified that BIC_R is robust to scaling

and clear performance benefits are observed as compared to the earlier proposed high-SNR forms of the BIC. Moreover, its performance is similar or slightly better than existing state-of-the-art high-SNR scale-invariant order selection criteria such as NML, gMDL, and PAL. Additionally, we analytically examined the consistency behaviour of $\mathrm{BIC}_{\mathrm{R}}$ as $\sigma^2 \to 0$ as well as when $N \to \infty$. In both instances, we have shown that the probability of $\mathrm{BIC}_{\mathrm{R}}$ selecting the true model order approaches one. However, note that the proposed modification is semi ad-hoc in nature and as such a deeper motivation for the rule is desired. Nevertheless, the primary objective was to design a scale-invariant robust form of BIC that guarantees consistent performance in large-N and high-SNR cases. As future work, we intend to extend the proposed $\mathrm{BIC}_{\mathrm{R}}$ to model selection in high-dimensional data settings ($p \gg N$) employing greedy or shrinkage methods for regressor selection.

Appendix

4.A Lemmas

Lemma 4.1 *Let* $\mathbf{y}$ *be a* $N \times 1$ *dimensional vector following* $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_N)$ *and* $\mathbf{Z}$ *be a* $N \times N$ *symmetric, idempotent matrix with rank*$(\mathbf{Z}) = r$*. Then the ratio* $\mathbf{y}^T\mathbf{Z}\mathbf{y}/\sigma^2$ *has a non-central chi-square distribution* $\chi^2_r(\lambda)$ *with* r *degrees of freedom and non-centrality parameter* $\lambda = \boldsymbol{\mu}^T\mathbf{Z}\boldsymbol{\mu}/\sigma^2$ *(See, e.g., Chapter 5 of [74]).*

Lemma 4.2 *Let* V_1 *and* V_2 *be two subspaces of* $\mathbb{R}^N$ *with orthogonal projection matrices* $\boldsymbol{\Pi}_1$ *and* $\boldsymbol{\Pi}_2$*, respectively. Then the following three statements are equivalent (Theorem 2.11, [75]).*

1. $V_1 \subset V_2$.
2. $\boldsymbol{\Pi}_2 - \boldsymbol{\Pi}_1$ *is the orthogonal projector onto* $V_2 \cap V_1^{\perp}$.
3. $\boldsymbol{\Pi}_1\boldsymbol{\Pi}_2 = \boldsymbol{\Pi}_2\boldsymbol{\Pi}_1 = \boldsymbol{\Pi}_1$.

Lemma 4.3 *For the general linear model* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$ *where* $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$*, given the assumption that the candidate models are nested and* k_0 *is the true model order such that* $x_k \neq 0$ *for* $k = 1, \ldots, k_0$ *and* $x_k = 0$ *for* $k = k_0 + 1, \ldots, p$*, then the ratio* $\left(\frac{\hat{\sigma}^2_{k_0+i}}{\hat{\sigma}^2_{k_0}}\right)$ *for* $i = 1, \ldots, p - k_0$ *follows a Beta distribution with parameters* $\alpha = \frac{N-k_0-i}{2}$ *and* $\beta = \frac{i}{2}$.

Proof: Given the ratio $X = \left(\hat{\sigma}^2_{k_0+i}/\hat{\sigma}^2_{k_0}\right)$, we can expand it as follows

$$X = \left(\frac{\hat{\sigma}^2_{k_0+i}}{\hat{\sigma}^2_{k_0}}\right) = \frac{\mathbf{y}^T\boldsymbol{\Pi}^{\perp}_{k_0+i}\mathbf{y}}{\mathbf{y}^T\boldsymbol{\Pi}^{\perp}_{k_0}\mathbf{y}} = \frac{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{k_0+i}\mathbf{e}}{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{k_0}\mathbf{e}} \quad \text{(Using 4.93)} \tag{4.82}$$

$$= \frac{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{k_0+i}\mathbf{e}}{\mathbf{e}^T\left(\mathbf{I}_N - \boldsymbol{\Pi}_{k_0+i} + \boldsymbol{\Pi}_{k_0+i} - \boldsymbol{\Pi}_{k_0}\right)\mathbf{e}} \tag{4.83}$$

$$= \frac{\left\{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{k_0+i}\mathbf{e}\right\}/\sigma^2}{\left\{\mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{k_0+i}\mathbf{e} + \mathbf{e}^T\left(\boldsymbol{\Pi}_{k_0+i} - \boldsymbol{\Pi}_{k_0}\right)\mathbf{e}\right\}/\sigma^2} = \frac{X_1}{X_1 + X_2} \tag{4.84}$$

where $X_1 = \mathbf{e}^T\boldsymbol{\Pi}^{\perp}_{k_0+i}\mathbf{e}/\sigma^2$ and $X_2 = \mathbf{e}^T\left(\boldsymbol{\Pi}_{k_0+i} - \boldsymbol{\Pi}_{k_0}\right)\mathbf{e}/\sigma^2$. Here, $span(\mathbf{A}_{k_0}) \subset span(\mathbf{A}_{k_0+i})$, therefore using Lemma 4.2 we have $\boldsymbol{\Pi}_{k_0}\boldsymbol{\Pi}_{k_0+i} = \boldsymbol{\Pi}_{k_0+i}\boldsymbol{\Pi}_{k_0} = \boldsymbol{\Pi}_{k_0}$. Now, we can show that the product $\boldsymbol{\Pi}^{\perp}_{k_0+i}(\boldsymbol{\Pi}_{k_0+i} - \boldsymbol{\Pi}_{k_0}) = \mathbf{0}$, which implies that the random variables X_1 and X_2 are statistically independent. Furthermore, from Lemma 4.2 we can say that $(\boldsymbol{\Pi}_{k_0+i} - \boldsymbol{\Pi}_{k_0})$ is a projection matrix projecting on the column space of $span(\mathbf{A}_{k_0+i}) \cap span(\mathbf{A}_{k_0})^{\perp}$ of dimension $k_0 + i - k_0 = i$. Hence, $X_2 \sim \chi^2_i$, and from results as in Appendix 4.C we have $X_1 \sim \chi^2_{N-k_0-i}$. It is well

known that if $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ are two independent random variables then the ratio $\frac{X_1}{X_1+X_2} \sim \mathcal{B}\left(\frac{k_1}{2}, \frac{k_2}{2}\right)$ [76]. Hence, we can conclude

$$X \sim \mathcal{B}\left(\frac{N-k_0-i}{2}, \frac{i}{2}\right). \tag{4.85}$$

4.B Statistical Analysis of $\hat{\sigma}_0^2$

From the generating model (4.1), the true data vector follows $\mathbf{y} \sim \mathcal{N}\left(\mathbf{A}_{k_0}\mathbf{x}_{k_0}, \sigma^2\mathbf{I}_N\right)$. Consider $\hat{\sigma}_0^2$ defined in (4.41)

$$\hat{\sigma}_0^2 = \frac{\|\mathbf{y}\|_2^2}{N} = \left(\frac{\sigma^2}{N}\right)\frac{\mathbf{y}^T\mathbf{I}_N\mathbf{y}}{\sigma^2}. \tag{4.86}$$

From Lemma 4.1 we have

$$\frac{\mathbf{y}^T\mathbf{I}_N\mathbf{y}}{\sigma^2} \sim \chi^2_N(\lambda) \text{ where } \lambda = \frac{\|\mathbf{A}_{k_0}\mathbf{x}_{k_0}\|_2^2}{\sigma^2}. \tag{4.87}$$

This implies that

$$\left(\frac{N}{\sigma^2}\right)\hat{\sigma}_0^2 \sim \chi^2_N(\lambda). \tag{4.88}$$

Therefore, the mean and variance of $\hat{\sigma}_0^2$ are:

$$\begin{aligned} \mathbb{E}[\hat{\sigma}_0^2] &= \frac{\sigma^2}{N}(N+\lambda) = \sigma^2 + \frac{\|\mathbf{A}_{k_0}\mathbf{x}_{k_0}\|_2^2}{N} \\ \mathrm{Var}[\hat{\sigma}_0^2] &= 2\frac{\sigma^4}{N^2}(N+2\lambda) = 2\frac{\sigma^4}{N} + 4\frac{\sigma^2}{N^2}\|\mathbf{A}_{k_0}\mathbf{x}_{k_0}\|_2^2. \end{aligned} \tag{4.89}$$

Hence, for a fixed N,

$$\lim_{\sigma^2\to 0}\mathbb{E}[\hat{\sigma}_0^2] = \frac{\|\mathbf{A}_{k_0}\mathbf{x}_{k_0}\|_2^2}{N} \quad \& \quad \lim_{\sigma^2\to 0}\mathrm{Var}[\hat{\sigma}_0^2] = 0. \tag{4.90}$$

Further, when SNR or σ^2 is fixed, using the assumption $\lim_{N\to\infty}\left\{\frac{\mathbf{A}_k^T\mathbf{A}_k}{N}\right\} = \mathbf{M}_k$ we get

$$\lim_{N\to\infty}\mathbb{E}[\hat{\sigma}_0^2] = \sigma^2 + \mathbf{x}_{k_0}^T\mathbf{M}_{k_0}\mathbf{x}_{k_0} \quad \& \quad \lim_{N\to\infty}\mathrm{Var}[\hat{\sigma}_0^2] = 0, \tag{4.91}$$

where $\mathbf{M}_{k_0}$ is a bounded positive definite matrix.

4.C Statistical Analysis of $\hat{\sigma}_k^2$

The noise variance estimate under hypothesis $\mathcal{H}_k$ can be rewritten as

$$\hat{\sigma}_k^2 = \left(\frac{\sigma^2}{N}\right)\frac{\mathbf{y}^T\mathbf{\Pi}_k^{\perp}\mathbf{y}}{\sigma^2}. \tag{4.92}$$

Case $k \geq k_0$:
The true model $\mathbf{t}^* = \mathbf{A}_{k_0}\mathbf{x}_{k_0}$ lies in a linear subspace spanned by the columns of $\mathbf{A}_{k_0}$. Consequently, because of the assumed nested structure of $\mathbf{A}$, we have $\mathbf{\Pi}_k^{\perp}\mathbf{t}^* = \mathbf{0}$ for all $k \geq k_0$. This implies that $\mathbf{y}^T\mathbf{\Pi}_k^{\perp}\mathbf{y} = \mathbf{e}^T\mathbf{\Pi}_k^{\perp}\mathbf{e}$ for all $k \geq k_0$. Thus we have,

$$\frac{\mathbf{y}^T\mathbf{\Pi}_k^{\perp}\mathbf{y}}{\sigma^2} = \frac{\mathbf{e}^T\mathbf{\Pi}_k^{\perp}\mathbf{e}}{\sigma^2} \sim \chi^2_{N-k} \text{ (Using Lemma 4.1)}, \quad \forall\ k \geq k_0. \tag{4.93}$$

This implies that

$$\left(\frac{N}{\sigma^2}\right)\hat{\sigma}_k^2 \sim \chi^2_{N-k}, \quad \forall\ k \geq k_0. \tag{4.94}$$

Therefore, the mean and variance of $\hat{\sigma}_k^2$ for $k \geq k_0$ are:

$$\mathbb{E}[\hat{\sigma}_k^2] = \frac{\sigma^2}{N}(N-k) \quad \& \quad \text{Var}[\hat{\sigma}_k^2] = 2\frac{\sigma^4}{N^2}(N-k), \quad \forall\ k \geq k_0. \tag{4.95}$$

Hence, for a fixed N,

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\hat{\sigma}_k^2] = 0 \quad \& \quad \lim_{\sigma^2 \to 0} \text{Var}[\hat{\sigma}_k^2] = 0, \quad \forall\ k \geq k_0. \tag{4.96}$$

Moreover, when σ^2 is constant,

$$\lim_{N \to \infty} \mathbb{E}[\hat{\sigma}_k^2] = \sigma^2 \quad \& \quad \lim_{N \to \infty} \text{Var}[\hat{\sigma}_k^2] = 0, \quad \forall\ k \geq k_0. \tag{4.97}$$

Case $k < k_0$:
For $k < k_0$ the random variable $\hat{\sigma}_k^2$ follows a noncentral chi-square distribution (cf. Lemma 4.1)

$$\left(\frac{N}{\sigma^2}\right)\hat{\sigma}_k^2 \sim \chi^2_{N-k}(\lambda) \quad \text{where } \lambda = \frac{(\mathbf{A}_{k_0}\mathbf{x}_{k_0})^T\mathbf{\Pi}_k^{\perp}(\mathbf{A}_{k_0}\mathbf{x}_{k_0})}{\sigma^2}, \quad \forall\ k < k_0. \tag{4.98}$$

Therefore, the mean and variance of $\hat{\sigma}_k^2$ for $k < k_0$ are:

$$\mathbb{E}[\hat{\sigma}_k^2] = \frac{\sigma^2}{N}(N-k+\lambda) \quad \& \quad \text{Var}[\hat{\sigma}_k^2] = 2\frac{\sigma^4}{N^2}(N-k+2\lambda), \quad \forall\ k < k_0. \tag{4.99}$$

Hence, for a fixed N,

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\hat{\sigma}_k^2] = \frac{(\mathbf{A}_{k_0}\mathbf{x}_{k_0})^T\mathbf{\Pi}_k^{\perp}(\mathbf{A}_{k_0}\mathbf{x}_{k_0})}{N} \quad \& \quad \lim_{\sigma^2 \to 0} \text{Var}[\hat{\sigma}_k^2] = 0, \quad \forall\ k < k_0. \tag{4.100}$$

Furthermore, we can show that

$$\frac{(\mathbf{A}_{k_0}\mathbf{x}_{k_0})^T\mathbf{\Pi}_k^{\perp}(\mathbf{A}_{k_0}\mathbf{x}_{k_0})}{N} = \mathbf{x}_{k_0}^T\left[\frac{\mathbf{A}_{k_0}^T\mathbf{U}\mathbf{U}^T\mathbf{A}_{k_0}}{N}\right]\mathbf{x}_{k_0} = \mathbf{x}_{k_0}^T\left[\frac{\tilde{\mathbf{A}}_{k_0}^T\tilde{\mathbf{A}}_{k_0}}{N}\right]\mathbf{x}_{k_0}, \tag{4.101}$$

where $\mathbf{\Pi}_k^{\perp} = \mathbf{U}\mathbf{U}^T$, $\tilde{\mathbf{A}}_{k_0} = \mathbf{U}^T\mathbf{A}_{k_0} \in \mathbb{R}^{(N-k)\times k_0}$ and $\mathbf{U} \in \mathbb{R}^{N\times(N-k)}$ is a semi-orthogonal matrix whose columns span the null space of $\mathbf{A}_k$. Therefore, under the assumption (4.25), when σ^2 is fixed we have

$$\lim_{N\to\infty} \mathbb{E}[\hat{\sigma}_k^2] = \sigma^2 + \mathbf{x}_{k_0}^T \widetilde{\mathbf{M}}_{k_0} \mathbf{x}_{k_0} \quad \& \quad \lim_{N\to\infty} \mathrm{Var}[\hat{\sigma}_k^2] = 0, \quad \forall\ k < k_0 \tag{4.102}$$

where $\widetilde{\mathbf{M}}_{\mathbf{k_0}} = \frac{\tilde{\mathbf{A}}_{k_0}^T \tilde{\mathbf{A}}_{k_0}}{N}$ is a bounded positive definite matrix.

Chapter 5

Extended Bayesian Information Criterion-Robust

> "The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge."
>
> —*Stephan Hawking (1942–2018)*

Model selection in linear regression models is a major challenge when dealing with high-dimensional data where the number of available measurements (sample size) is much smaller than the dimension of the parameter space. Traditional methods for model selection such as Akaike information criterion, Bayesian information criterion (BIC), and minimum description length are heavily prone to overfitting in the high-dimensional setting. In this regard, extended BIC (EBIC), which is an extended version of the original BIC, and extended Fisher information criterion (EFIC), which is a combination of EBIC and Fisher information criterion, are consistent estimators of the true model as the number of measurements grows very large. However, EBIC is not consistent in high signal-to-noise-ratio (SNR) scenarios where the sample size is fixed and EFIC is not invariant to data-scaling resulting in unstable behaviour. In this chapter, we present a new form of the EBIC criterion called EBIC-Robust (or EBIC_{R} in short), which is invariant to data-scaling and consistent in both large sample sizes and high-SNR scenarios. Analytical proofs are presented to guarantee its consistency. Simulation results indicate that the performance of EBIC_{R} is quite superior to that of both EBIC and EFIC.

5.1 Introduction

In Chapter 4, we proposed a robust form of the classical BIC to mitigate the data-scaling problem in the high-SNR forms of the BIC. In this chapter, we expand this concept to handle model selection in high-dimensional settings where $p \gg N$. Our

primary focus is on model selection in high-dimensional linear regression models associated with the maximum likelihood (ML) method of parameter estimation. Consider the linear model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}, \tag{5.1}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the measurement vector and $\mathbf{A} \in \mathbb{R}^{N \times p}$ is the known design matrix. We are considering a high-dimensional setting, hence $p > N$. Also, we consider the case where p is not fixed but it can grow with N. Here, we link p to N as $p = N^d$, where $d > 0$ is a real value. This is a common setting in the model selection literature [17, 18, 77]. $\mathbf{e} \in \mathbb{R}^N$ is the associated noise vector whose elements are assumed to be i.i.d. following a Gaussian distribution, i.e., $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_N)$ where σ^2 is the unknown true noise power. $\mathbf{x} \in \mathbb{R}^p$ is the unknown parameter vector. Here, $\mathbf{x}$ is assumed to be sparse, which implies that very few of the elements of $\mathbf{x}$ are non-zero. We denote $\mathcal{S}$ as the true support of $\mathbf{x}$, i.e., $\mathcal{S} = \{i : x_i \neq 0\}$ having cardinality $card(\mathcal{S}) = k_0 \ll N$ and $\mathbf{A}_\mathcal{S}$ as the set of columns of $\mathbf{A}$ corresponding to the support $\mathcal{S}$. The goal of model selection is estimating $\mathcal{S}$ given $\mathbf{y}$ and $\mathbf{A}$.

Among the classical methods of model selection, BIC has been quite successful due to its simplicity and consistent performance in many fields. BIC is asymptotically consistent in selecting the true model as N grows very large given that p and the true noise variance σ^2 is fixed. However, its performance in high-dimensional settings when $p > N$ is not satisfactory and it has a tendency to select more co-variates than required, thus overfitting the model [17]. To handle the large-p small-N scenario, the authors in [17] proposed a novel extension to the original BIC called extended BIC (EBIC), that takes into account both the number of unknown parameters and the complexity of the model space. EBIC adds dynamic prior model probabilities to each of the models under consideration that is inversely proportional to the model set dimension. This eliminates the earlier assumption of assigning uniform prior to all models irrespective of their sizes, which goes against the principle of parsimony. Under a suitable asymptotic identifiability condition, EBIC is consistent such that it selects the true model as N tends to infinity [17]. However, the consistent behaviour of EBIC fails when N is small and fixed and σ^2 tends to zero [18]. This new consistency requirement was first introduced in [68], where the authors highlighted that the original BIC is also inconsistent for fixed N and decreasing noise variance scenarios where $N > p$.

To overcome the drawbacks of EBIC, the authors in [18] proposed a novel criterion called extended Fisher information criterion (EFIC) that is inspired by EBIC and the model selection criteria with Fisher information [54]. The authors analyzed the performance of EFIC in the high-dimensional setting for two key cases: (1) when σ^2 is fixed and N tends to infinity; (2) when N is fixed and σ^2 tends to zero. In each case, it was shown that EFIC selects the true model with a probability approaching one. However, as indicated in our simulations, EFIC is not invariant to data-scaling and it tends to suffer from overfitting issues (and sometimes underfitting) in practical sizes of N when the data is scaled. This scaling problem is a

result of the data dependent penalty design that may blow the penalty to extremely small or large values depending on how the data is scaled.

Apart from the criteria mentioned above, there are other non-information theoretic methods available for model selection. One such popular method is cross-validation (CV) [78]. However, the performance of CV is quite poor in sample scarce scenarios with large parameter dimensions and even though CV is unbiased, it can have high variance [59]. Recent additions to the list of model selection methods for high-dimensional data are Residual Ratio Thresholding (RRT) [65] and Multi-Beta-Test (MBT) [70]. Both are non-information theoretic methods based on hypothesis testing using a test statistic. They operate along with a greedy variable selection method such as orthogonal matching pursuit (OMP) [37] and involve a tuning parameter $\in [0, 1]$, that controls the false-discovery rate. However, there is no optimal way to set it, and as such their behaviour may tend to overfit or underfit depending on the chosen tuning parameter value. Moreover, in their current form, they can only be used with algorithms that generate monotonic sequences of support estimates such as OMP, which restricts their usability.

In this chapter, we present a modified criterion for model selection in high-dimensional linear regression models called EBIC-Robust or EBIC_{R} in short, where the subscript R stands for robust. EBIC_{R} is a scale-invariant and consistent criterion. To guarantee the consistency, we provide analytical proofs to show that under a suitable asymptotic identifiability condition, EBIC_{R} selects the true model with a probability approaching one as $N \to \infty$ as well as when $\sigma^2 \to 0$.

The notations used here are restated for convenience. Boldface letters denote matrices and vectors. The notation $(\cdot)^T$ stands for transpose. $\mathbf{A}_{\mathcal{I}}$ denotes a submatrix of the full matrix $\mathbf{A}$ formed using the columns indexed by the support set $\mathcal{I}$. $\mathbf{\Pi}_{\mathcal{I}} = \mathbf{A}_{\mathcal{I}}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})^{-1}\mathbf{A}_{\mathcal{I}}^T$ denotes the orthogonal projection matrix on the span of $\mathbf{A}_{\mathcal{I}}$ and $\mathbf{\Pi}_{\mathcal{I}}^{\perp} = \mathbf{I}_N - \mathbf{\Pi}_{\mathcal{I}}$ denotes the orthogonal projection matrix on the null space of $\mathbf{A}_{\mathcal{I}}^T$. The notation $\left|\mathbf{X}\right|$ denotes the determinant of the matrix $\mathbf{X}$ and $\|\cdot\|_2$ denotes the Euclidean norm. $X \sim \mathcal{N}(0, 1)$ denotes a normal distributed random variable with mean 0 and variance 1. $X \sim \chi_k^2$ is a central chi-squared distributed random variable with k degrees of freedom, whereas $X \sim \chi_k^2(\lambda)$ is a noncentral chi-squared distributed random variable with k degrees of freedom and non-centrality parameter λ.

5.2 Background

Given the linear model (5.1), the entire process of model selection or in other words estimating the true support set $\mathcal{S}$ involves two major steps: (i) Predictor/subset selection, which includes finding a competent set of candidate models out of all the $(2^p - 1)$ possible models. In our work, we consider the set of competing models as the collection of all plausible combinatorial models up to a maximum cardinality K, under the assumption that $k_0 \leq K \ll N$; (ii) estimating the true model among the candidate models using a suitable model selection criterion.

Now, for any arbitrary candidate model with support $\mathcal{I}$ having cardinality $card(\mathcal{I}) = k$, the linear model in (5.1) can be reformulated as follows

$$\mathcal{H}_{\mathcal{I}} : \mathbf{y} = \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}} + \mathbf{e}_{\mathcal{I}}, \tag{5.2}$$

where $\mathcal{H}_{\mathcal{I}}$ denotes the hypothesis that the data $\mathbf{y}$ is truly generated according to (5.2), $\mathbf{A}_{\mathcal{I}} \in \mathbb{R}^{N\times k}$ is the sub-design matrix consisting of columns from the known design matrix $\mathbf{A}$ with support $\mathcal{I}$, $\mathbf{x}_{\mathcal{I}} \in \mathbb{R}^k$ is the corresponding unknown parameter vector and $\mathbf{e}_{\mathcal{I}} \in \mathbb{R}^N$ is the associated noise vector following $\mathbf{e}_{\mathcal{I}} \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathcal{I}}^2\mathbf{I}_N)$ where $\sigma_{\mathcal{I}}^2$ is the unknown noise variance corresponding to the hypothesis $\mathcal{H}_{\mathcal{I}}$.

5.2.1 Bayesian Framework for Model Selection

To motivate the proposed criterion we start by describing the Bayesian framework that leads to the maximum a-posteriori (MAP) estimator, which in turn forms the backbone for deriving BIC and its extended versions, viz., EBIC, EFIC, as well as the proposed criterion EBIC_{R}. Now, for the considered model in (5.2), the probability density function (pdf) of the data vector $\mathbf{y}$ is given as

$$p(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}}) = \frac{\exp\{-\|\mathbf{y} - \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}\|_2^2/2\sigma_{\mathcal{I}}^2\}}{(2\pi\sigma_{\mathcal{I}}^2)^{N/2}}, \tag{5.3}$$

where $\boldsymbol{\theta}_{\mathcal{I}} = [\mathbf{x}_{\mathcal{I}}^T, \sigma_{\mathcal{I}}^2]^T$ comprises of all the parameters of the model. Under hypothesis $\mathcal{H}_{\mathcal{I}}$, the maximum likelihood estimates (MLEs) of $\hat{\boldsymbol{\theta}}_{\mathcal{I}} = [\hat{\mathbf{x}}_{\mathcal{I}}^T, \hat{\sigma}_{\mathcal{I}}^2]^T$ are obtained as [63]

$$\hat{\mathbf{x}}_{\mathcal{I}} = \left(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right)^{-1}\mathbf{A}_{\mathcal{I}}^T\mathbf{y} \quad \& \quad \hat{\sigma}_{\mathcal{I}}^2 = \frac{\mathbf{y}^T\boldsymbol{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}}{N}. \tag{5.4}$$

Let $p(\boldsymbol{\theta}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ denote the prior pdf of the parameter vector $\boldsymbol{\theta}_{\mathcal{I}}$ under $\mathcal{H}_{\mathcal{I}}$. Then we have the joint probability

$$p(\mathbf{y}, \boldsymbol{\theta}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) = p(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}})p(\boldsymbol{\theta}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) \tag{5.5}$$

and the marginal distribution of $\mathbf{y}$ is

$$p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) = \int p(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}})p(\boldsymbol{\theta}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})d\boldsymbol{\theta}_{\mathcal{I}}. \tag{5.6}$$

The posterior probability $\Pr(\mathcal{H}_{\mathcal{I}}|\mathbf{y})$ is given by

$$\Pr(\mathcal{H}_{\mathcal{I}}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})\Pr(\mathcal{H}_{\mathcal{I}})}{p(\mathbf{y})}, \tag{5.7}$$

where $\Pr(\mathcal{H}_{\mathcal{I}})$ is the prior probability of the model with support $\mathcal{I}$. The MAP estimator picks the model with the largest posterior probability $\Pr(\mathcal{H}_{\mathcal{I}}|\mathbf{y})$. Ignoring

$p(\mathbf{y})$ which is a normalizing factor and independent of $\mathcal{I}$, the MAP estimate of $\mathcal{S}$ is equivalently given by

$$\hat{\mathcal{S}}_{\text{MAP}} = \arg\max_{\mathcal{I}} \Big\{ \ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) + \ln \Pr(\mathcal{H}_{\mathcal{I}}) \Big\}. \tag{5.8}$$

To compute the MAP estimate, we need to evaluate the integral in (5.6). Traditionally, under the assumption that N and/or SNR are large, we can obtain an approximation of $\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}})$ using a second order Taylor series expansion, which gives (see [16, 79] for details)

$$\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) \approx \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}}) + \ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) + \frac{k+1}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right|, \tag{5.9}$$

where $k = card(\mathcal{I})$ and $\hat{\mathbf{F}}_{\mathcal{I}}$ is the sample Fisher information matrix under $\mathcal{H}_{\mathcal{I}}$ given as [63]

$$\hat{\mathbf{F}}_{\mathcal{I}} = -\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}})}{\partial \boldsymbol{\theta}_{\mathcal{I}} \partial \boldsymbol{\theta}_{\mathcal{I}}^T}\bigg|_{\boldsymbol{\theta}_{\mathcal{I}}=\hat{\boldsymbol{\theta}}_{\mathcal{I}}}. \tag{5.10}$$

Evaluating (5.10) using (5.3) and (5.4) we get [16]

$$\hat{\mathbf{F}}_{\mathcal{I}} = \begin{bmatrix} \frac{1}{\hat{\sigma}^2_{\mathcal{I}}}\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}} & \mathbf{0} \\ \mathbf{0} & \frac{N}{2\hat{\sigma}^4_{\mathcal{I}}} \end{bmatrix}. \tag{5.11}$$

Now, for the considered linear model we have

$$-2\ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\mathcal{I}}, \mathcal{H}_{\mathcal{I}}) = N\ln\hat{\sigma}^2_{\mathcal{I}} + \text{const.} \tag{5.12}$$

Therefore, using (5.12), we can rewrite (5.9) as

$$-2\ln p(\mathbf{y}|\mathcal{H}_{\mathcal{I}}) \approx N\ln\hat{\sigma}^2_{\mathcal{I}} + \ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| - 2\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}}) - k\ln 2\pi + \text{const.} \tag{5.13}$$

Furthermore, it is assumed that the prior term in (5.9), i.e., $\ln p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ is flat and uninformative, and hence disregarded from the analysis. Thus, dropping the constants and the terms independent of the model dimension k, we can equivalently reformulate the MAP based model estimate as

$$\hat{\mathcal{S}}_{\text{MAP}} = \arg\min_{\mathcal{I}} \Big\{ N\ln\hat{\sigma}^2_{\mathcal{I}} + \ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| - k\ln 2\pi - 2\ln\Pr(\mathcal{H}_{\mathcal{I}}) \Big\}. \tag{5.14}$$

5.2.2 BIC

The BIC can be obtained from the MAP estimator in (5.14). The term $-k\ln 2\pi$ is ignored as it weakly depends on the model dimension k and hence is typically much smaller than the dominating terms. Moreover, the prior probability of each candidate model is assumed to be equiprobable. Hence, the $-2\ln\Pr(\mathcal{H}_{\mathcal{I}})$ term is dropped as well. Now, expanding the $\ln|\hat{\mathbf{F}}_{\mathcal{I}}|$ term of (5.14) using (5.11) we have

$$\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| = \ln(N/2) - (k+2)\ln\hat{\sigma}^2_{\mathcal{I}} + \ln\left|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right|. \tag{5.15}$$

Here, the following property of the design matrix $\mathbf{A}$ is assumed [16, 71]

$$\lim_{N\to\infty} \left\{N^{-1}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})\right\} = \mathbf{M}_{\mathcal{I}} = \mathcal{O}(1), \tag{5.16}$$

where $\mathbf{M}_{\mathcal{I}}$ is a $k \times k$ positive definite matrix and bounded as $N \to \infty$. The assumption in (5.16) is true in many applications but not all (see [72] for more details). Using (5.16), it is possible to show that for large N

$$\ln\left|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right| = \ln\left|N\cdot N^{-1}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})\right| = k\ln N + \mathcal{O}(1). \tag{5.17}$$

Furthermore, $\hat{\sigma}_{\mathcal{I}}^2$ is considered to be of $\mathcal{O}(1)$ as well since it does not grow with N. As such, the $\mathcal{O}(1)$ term, $(k+2)\ln\hat{\sigma}_{\mathcal{I}}^2$ and $\ln(N/2)$ (a constant) are ignored from (5.15). This leads to the final form of the BIC

$$\mathrm{BIC}(\mathcal{I}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + k\ln N. \tag{5.18}$$

BIC is consistent when p is fixed and $N \to \infty$. However, it is inconsistent when N is fixed and $\sigma^2 \to 0$ [67, 79] as well as when $p > N$ and p grows with N [17].

5.2.3 EBIC

The authors in [17] proposed an extended version of the BIC, i.e., EBIC, to mitigate the drawbacks of BIC for large-p small-N scenarios. EBIC can be derived from the MAP estimator in (5.14), using the same assumptions as in BIC, except for the prior probability term $\Pr(\mathcal{H}_{\mathcal{I}})$. In EBIC, the idea of equiprobable models is discredited, and instead, a prior probability is assigned that is inversely proportional to the size of the model space. Thus, a model with dimension k is assigned prior probability of $\Pr(\mathcal{H}_{\mathcal{I}}) \propto \binom{p}{k}^{-\gamma}$, where $0 \le \gamma \le 1$ is a tuning parameter. Thus, the EBIC is

$$\mathrm{EBIC}(\mathcal{I}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + k\ln N + 2\gamma\ln\binom{p}{k}. \tag{5.19}$$

When $\gamma = 0$, EBIC boils down to BIC (5.18). Moreover, unlike BIC, EBIC is consistent in selecting the true model for $p \gg N$ cases where p grows with N. However, empirical experiments performed in [18] show that in situations when N is small and fixed, EBIC is inconsistent as $\sigma^2 \to 0$.

5.2.4 EFIC

To circumvent the shortcomings of EBIC in high-SNR cases, the authors in [18] proposed EFIC. In EFIC, the assumptions imposed on the sample FIM (5.15) are removed and the entire structure is included as it is in the criterion except for the constant term $\ln(N/2)$. Some further simplifications are involved:

$$N\ln\hat{\sigma}_{\mathcal{I}}^2 = N\ln\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2 - N\ln N \tag{5.20}$$

$$(k+2)\ln\hat{\sigma}_{\mathcal{I}}^2 = (k+2)\left[\ln\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2 - \ln N\right]. \tag{5.21}$$

The $-N\ln N$ and $-2\ln N$ term of (5.20) and (5.21) respectively are independent of the model dimension k and hence ignored. Similar to EBIC the prior probability term is assumed to be proportional to the model space, hence $\Pr(\mathcal{H}_{\mathcal{I}}) \propto \binom{p}{k}^{-c}$, where $c > 0$ is a tuning parameter. Furthermore, under the large-p approximation and since $k \leq K \ll p$, the $\ln\binom{p}{k}$ term is approximated as

$$\ln\binom{p}{k} = \sum_{i=0}^{k-1} \ln(p-i) - \ln(k!) \approx k\ln p. \tag{5.22}$$

Hence, for large-p case, it is possible to set $-2\ln p(\mathcal{H}_{\mathcal{I}}) \approx 2ck\ln p$. Thus, the EFIC is given as

$$\mathrm{EFIC}(\mathcal{I}) = N\ln\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2 + k\ln N + \ln\left|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right| - (k+2)\ln\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2 + 2ck\ln p. \tag{5.23}$$

If we replace $p = N^d$, then the last term in (5.23) can be written as $2ckd\ln N$. EFIC is consistent in both large-N and high-SNR scenarios [18]. However, EFIC suffers from a scaling problem due to the inclusion of the data-dependent penalty term (i.e., $-(k+2)\ln\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2$) and as such, the performance of EFIC is not invariant to data-scaling. This point will be further discussed in Section 5.3.1.

5.3 Proposed Criterion: EBIC-Robust (EBIC$_{\mathrm{R}}$)

In this section, we present the necessary steps for deriving EBIC$_{\mathrm{R}}$. EBIC$_{\mathrm{R}}$ can be seen as a natural extension of BIC$_{\mathrm{R}}$ for performing model selection in large-p small-N scenarios. Below, we provide a detailed derivation and establish the connection to BIC$_{\mathrm{R}}$. The initial steps are identical to that in Chapter 4, however, we present them here for completeness. We perform normalization of $\hat{\mathbf{F}}_{\mathcal{I}}$ considering large-N and high-SNR scenario. For this we factorize the $\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right|$ term in (5.14) in the following manner

$$\begin{aligned}\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| &= \ln\left[\left|\mathbf{L}\right|\left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{L}^{-1/2}\right|\right] \\ &= \ln|\mathbf{L}| + \ln\underbrace{\left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{L}^{-1/2}\right|}_{\mathrm{T}}.\end{aligned} \tag{5.24}$$

The goal here is to choose a suitable $\mathbf{L}$ matrix that normalizes the sample FIM $\hat{\mathbf{F}}_{\mathcal{I}}$ such that the T term in (5.24) is $\mathcal{O}(1)$, i.e., in this case T should be bounded as $N \to \infty$ and/or $\sigma^2 \to 0$. To accomplish this objective, we choose the following $\mathbf{L}^{-1/2}$ matrix

$$\mathbf{L}^{-1/2} = \begin{bmatrix} \sqrt{\frac{1}{N}}\sqrt{\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_0^2}}\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \sqrt{\frac{1}{N}}\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_0^2} \end{bmatrix}, \tag{5.25}$$

where $\hat{\sigma}_0^2 = \|\mathbf{y}\|_2^2/N$. The factor, $\hat{\sigma}_0^2$, is used in $\mathbf{L}^{-1/2}$ in order to neutralize the data-scaling problem and is motivated by the fact that given (5.16), when the SNR is a constant, we have

$$\mathbb{E}[\hat{\sigma}_0^2] \to \text{const.} \quad \& \quad \text{Var}[\hat{\sigma}_0^2] \to 0 \tag{5.26}$$

as $N \to \infty$. Furthermore, from the considered generating model in (5.1), when N is fixed, (5.26) is also satisfied as $\sigma^2 \to 0$ (see Appendix 5.B for details on $\hat{\sigma}_0^2$). Now using (5.11), (5.25) and the assumptions in (5.16), (5.26), it is possible to show that

$$\left|\mathbf{L}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{L}^{-1/2}\right| = \begin{vmatrix} \frac{1}{\hat{\sigma}_0^2}\frac{\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}}{N} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\hat{\sigma}_0^4} \end{vmatrix} = \mathcal{O}(1), \tag{5.27}$$

as $N \to \infty$ and/or $\sigma^2 \to 0$ and therefore may be discarded without much effect on the criterion. Furthermore, the $\ln|\mathbf{L}|$ term can be expanded as follows

$$\begin{aligned} \ln|\mathbf{L}| &= \ln \begin{vmatrix} N\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right)\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & N\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right)^2 \end{vmatrix} \\ &= (k+1)\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right). \end{aligned} \tag{5.28}$$

Therefore, using (5.27) and (5.28) we can rewrite (5.24) as

$$\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| = k\ln N + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right) + \mathcal{O}(1) + \ln N. \tag{5.29}$$

Next, for the model prior probability term $-2\ln\Pr(\mathcal{H}_{\mathcal{I}})$ in (5.14), a similar proposition is taken as in EBIC such that $\Pr(\mathcal{H}_{\mathcal{I}}) \propto \binom{p}{k}^{-\zeta}$, where $\zeta \geq 0$ is a tuning parameter. For large-p, we follow a similar approach as in EFIC by employing the following approximation $\ln\binom{p}{k} \approx k\ln p$. This gives

$$-2\ln\Pr(\mathcal{H}_{\mathcal{I}}) = 2\zeta k\ln p + \text{const.} \tag{5.30}$$

Now, substituting (5.29), (5.30) in (5.14) and dropping the $\mathcal{O}(1)$, the $\ln N$ term (independent of k), the constant and the $p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ term we arrive at the EBIC_{R}:

$$\text{EBIC}_{\text{R}}(\mathcal{I}) = N\ln\hat{\sigma}_{\mathcal{I}}^2 + k\ln\left(\frac{N}{2\pi}\right) + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right) + 2k\zeta\ln p. \tag{5.31}$$

If we use the relation $p = N^d$, the last term in (5.31) can be replaced with $2k\zeta d\ln N$. The true model is estimated as

$$\hat{\mathcal{S}}_{\text{EBIC}_{\text{R}}} = \arg\min_{\mathcal{I}}\big\{\text{EBIC}_{\text{R}}(\mathcal{I})\big\}. \tag{5.32}$$

It can be observed from (5.31) that the penalty of EBIC$_{\mathrm{R}}$ is a function of the number of measurements N, the ratio $(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2)$ and the parameter dimension p. Notice that the ratio $(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2)$ is always greater than one and independent of the scaling of $\mathbf{y}$. Furthermore, when $\mathcal{S} \not\subset \mathcal{I}$, the ratio $(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2) \approx \mathcal{O}(1)$ and for $\mathcal{S} \subset \mathcal{I}$ we have $(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2) \approx \mathcal{O}(\mathrm{SNR}+1)$. Hence, the behaviour of the penalty can be summarized as follows: (i) For fixed p and SNR, as $N \to \infty$ the penalty grows as $\mathcal{O}(\ln N)$; (ii) If N and p are constant, as SNR $\to \infty$, the penalty grows approximately as $\mathcal{O}\left(\ln(\mathrm{SNR}+1)\right)$ for all $\mathcal{I} \supset \mathcal{S}$; (iii) when SNR is a constant and given that p grows with N, then as $N \to \infty$ the penalty grows as $\mathcal{O}(\ln N) + \mathcal{O}(\ln p)$.

5.3.1 Scaling Robustness as Compared to EFIC

In this section, we elaborately discuss the data-scaling problem. Ideally, any model selection criterion should be invariant to data-scaling, which means that if $\mathbf{y}$ is scaled by any arbitrary constant $C > 0$, the equivalent penalty for each of the models $\mathcal{I}$ should not change. This property is necessary because otherwise the behaviour of the model selection criterion will be unreliable and may suffer from overfitting or underfitting issues when the data is scaled. As mentioned before, the penalty of EFIC is not invariant to data-scaling. This can be observed from the following analysis. Let $\Delta = card(\mathcal{I}) - card(\mathcal{S})$. Now, consider the difference assuming $\mathcal{I} \neq \mathcal{S}$

$$\begin{aligned}\mathrm{EFIC}(\mathcal{I}) - \mathrm{EFIC}(\mathcal{S}) &= (N-2)\ln\frac{\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2}{\left\|\mathbf{\Pi}_{\mathcal{S}}^{\perp}\mathbf{y}\right\|_2^2} + \ln\frac{\left|\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\right|}{\left|\mathbf{A}_{\mathcal{S}}^T\mathbf{A}_{\mathcal{S}}\right|} - k\ln\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\right\|_2^2 \\ &\quad + k_0\ln\left\|\mathbf{\Pi}_{\mathcal{S}}^{\perp}\mathbf{y}\right\|_2^2 + \Delta\left(\ln N + 2c\ln p\right) \\ &= D_{\mathrm{EFIC}}\ (\text{say}). \end{aligned} \tag{5.33}$$

Ideally, for correct model selection, $D_{\mathrm{EFIC}} > 0$ for all $\mathcal{I} \neq \mathcal{S}$. Now, if we scale the data $\mathbf{y}$ by a constant $C > 0$, the data dependent term becomes $\ln\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}C\mathbf{y}\|_2^2 = \ln C^2 + \ln\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2$ and the difference becomes

$$\mathrm{EFIC}(\mathcal{I}) - \mathrm{EFIC}(\mathcal{S}) = D_{\mathrm{EFIC}} - \Delta\ln C^2. \tag{5.34}$$

It is evident that (5.33) and (5.34) are unequal and the difference after scaling contains an additional term $-\Delta\ln C^2$. This implies that scaling changes the EFIC score difference between any arbitrary model $\mathcal{I}$ and the true model $\mathcal{S}$. Hence, depending on the C value ($C < 1$ or $C \geq 1$) and $\Delta > 0$ or $\Delta < 0$, the difference in (5.34) may become negative leading to a false model selection. Thus, EFIC is not invariant to data-scaling. On the contrary, consider the difference for EBIC$_{\mathrm{R}}$,

$$\begin{aligned}&\mathrm{EBIC}_{\mathrm{R}}(\mathcal{I}) - \mathrm{EBIC}_{\mathrm{R}}(\mathcal{S}) \\ &= (N-2)\ln\left(\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_{\mathcal{S}}^2}\right) - k\ln\hat{\sigma}_{\mathcal{I}}^2 + k_0\ln\hat{\sigma}_{\mathcal{S}}^2 + \Delta\ln\hat{\sigma}_0^2 + \Delta\left[\ln\left(\frac{N}{2\pi}\right) + 2\zeta\ln p\right] \\ &= D_{\mathrm{EBIC}_{\mathrm{R}}}(\text{say}) \end{aligned} \tag{5.35}$$

Now, scaling $\mathbf{y}$ by C, scales the noise variance estimates $\hat{\sigma}^2_{\mathcal{I}}$, $\hat{\sigma}^2_{\mathcal{S}}$ and $\hat{\sigma}^2_0$ by C^2, however, the difference remains the same, i.e., $D_{\mathrm{EBIC_R}}$. This is because in this case the $-\Delta \ln C^2$ term is cancelled by $+\Delta \ln C^2$ generated by $\Delta \ln \hat{\sigma}^2_0$. Hence, $\mathrm{EBIC_R}$ is invariant to data-scaling, which is a desired property of any model selection criterion.

5.4 Consistency of $\mathrm{EBIC_R}$

In this section, we provide the necessary proofs to show that $\mathrm{EBIC_R}$ is a consistent criterion. Generally speaking, a model selection criterion with $\hat{\mathcal{S}}$ as its estimate of the true model $\mathcal{S}$ is consistent if it satisfies the following conditions [18]

$$\lim_{\sigma^2 \to 0} \Pr\{\hat{\mathcal{S}} = \mathcal{S}\} = 1 \quad \& \quad \lim_{N \to \infty} \Pr\{\hat{\mathcal{S}} = \mathcal{S}\} = 1. \tag{5.36}$$

Let us define the set of all overfitted models of dimension k as

$$\mathcal{I}^k_o = \{\mathcal{I} : card(\mathcal{I}) = k, \mathcal{S} \subset \mathcal{I}\}, \tag{5.37}$$

and the set of all misfitted models of dimension k as

$$\mathcal{I}^k_m = \{\mathcal{I} : card(\mathcal{I}) = k, \mathcal{S} \not\subset \mathcal{I}\}. \tag{5.38}$$

Furthermore, let $\mathbb{O}$ denote the set of all $\mathcal{I}^k_o$ for $k = k_0 + 1, \ldots, K$, and let $\mathbb{M}$ denote the set of all $\mathcal{I}^k_m$ for $k = 1, \ldots, K$, i.e.,

$$\mathbb{O} = \bigcup_{k=k_0+1}^{K} \mathcal{I}^k_o \quad \text{and} \quad \mathbb{M} = \bigcup_{k=1}^{K} \mathcal{I}^k_m \ , \tag{5.39}$$

where K is some upper bound for k_0 and $k_0 \leq K \ll N$. In practice, $\mathrm{EBIC_R}$ picks the true model $\mathcal{S}$, if the following conditions are satisfied:

$$\mathcal{C}_1 : \mathrm{EBIC_R}(\mathcal{S}) < \mathrm{EBIC_R}(\mathcal{I}) \quad \forall\, \mathcal{I} \in \mathbb{O} \tag{5.40}$$

$$\mathcal{C}_2 : \mathrm{EBIC_R}(\mathcal{S}) < \mathrm{EBIC_R}(\mathcal{I}) \quad \forall\, \mathcal{I} \in \mathbb{M}. \tag{5.41}$$

5.4.1 Asymptotic Identifiability of the Model

In general, the model is identifiable if no model of comparable size other than the true submodel can predict the noise free response almost equally well [17]. In the context of linear regression, this is equivalent to say $\mathbf{y} = \mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} \neq \mathbf{A}_{\mathcal{I}}\mathbf{x}_{\mathcal{I}}$ for $\left\{\mathcal{I} : card(\mathcal{I}) \leq card(\mathcal{S}), \mathcal{I} \neq \mathcal{S}\right\}$. The identifiability of the true model in the high-dimensional linear regression setup is uniformly maintained if the minimal eigenvalue of all restricted sub-matrices, $\mathbf{A}^T_{\mathcal{I}}\mathbf{A}_{\mathcal{I}}$ for $\{\mathcal{I} : card(\mathcal{I}) \leq 2K\}$, is bounded away from zero [18]. A sufficient assumption on the design matrix $\mathbf{A}$ to prove the consistency of $\mathrm{EBIC_R}$ is the sparse Riesz condition [39]:

$$\lim_{N \to \infty} \left\{N^{-1}\left(\mathbf{A}^T_{\mathcal{I}}\mathbf{A}_{\mathcal{I}}\right)\right\} = \mathbf{M}_{\mathcal{I}}, \quad \forall\ card(\mathcal{I}) \leq 2K, \tag{5.42}$$

where $\mathbf{M}_{\mathcal{I}}$ denotes a bounded positive definite matrix.

5.4.2 Consistency as $\sigma^2 \to 0$ or SNR $\to \infty$ for Fixed N

In this subsection, we examine whether EBIC$_{\mathrm{R}}$ selects the true model $\mathcal{S}$ as σ^2 goes vanishingly small (or equivalently SNR$\to \infty$) under the assumption that N is fixed. We formulate this into a theorem as follows:

Theorem 5.1 *Assume that N and p are fixed and the matrix $\mathbf{A}$ satisfies the condition given by (5.42). If $K \geq k_0$, then $\Pr\{EBIC_R(\mathcal{S}) < EBIC_R(\mathcal{I})\} \to 1$ as $\sigma^2 \to 0$ for all $\mathcal{I} \neq \mathcal{S}$ and $card(\mathcal{I}) = 1, \ldots, K$.*

Proof. The proof consists of two parts. In part (a) we show that the probability of overfitting ($\mathcal{S} \subset \hat{\mathcal{S}}_{\mathrm{EBIC_R}}$) tends to 0 as $\sigma^2 \to 0$, which in this case is equivalent to showing $\lim_{\sigma^2 \to 0} \Pr(\mathcal{C}_1) = 1$, cf. (5.40). In part (b) we show that the probability of misfitting ($\mathcal{S} \not\subset \hat{\mathcal{S}}_{\mathrm{EBIC_R}}$) also tends to 0 as $\sigma^2 \to 0$, which is equivalent to $\lim_{\sigma^2 \to 0} \Pr(\mathcal{C}_2) = 1$, cf. (5.41).

(*a*) *Over-fitting case* ($\mathcal{S} \subset \hat{\mathcal{S}}_{\mathrm{EBIC_R}}$): Consider the set of overfitted subsets having cardinality k, which we have denoted as $\mathcal{I}_o^k$. Let $\mathcal{I}_j$ denote the jth subset in the set $\mathcal{I}_o^k$. The total number of subsets in $\mathcal{I}_o^k$ is $\binom{p-k_0}{\Delta}$ where $\Delta = k - k_0$. For any overfitted subset $\mathcal{I}_j \in \mathcal{I}_o^k$, consider the following inequality

$$\mathrm{EBIC_R}(\mathcal{S}) < \mathrm{EBIC_R}(\mathcal{I}_j), \quad \mathcal{I}_j \in \mathcal{I}_o^k, \tag{5.43}$$

where $j = 1, \ldots, \binom{p-k_0}{\Delta}$. Using the relation $p = N^d$ and after some straightforward rearrangement of (5.43) we get

$$(N - k_0 - 2)\ln\hat{\sigma}_{\mathcal{S}}^2 - (N - k - 2)\ln\hat{\sigma}_{\mathcal{I}_j}^2 - \Delta(1 + 2\zeta d)\ln N - \Delta\ln\hat{\sigma}_0^2 + \Delta\ln 2\pi < 0. \tag{5.44}$$

Let us define a random variable $X_{\mathcal{I}_j} = \hat{\sigma}_{\mathcal{I}_j}^2/\sigma^2$, then

$$N \cdot X_{\mathcal{I}_j} \sim \chi_{N-k}^2, \quad \forall\ \mathcal{I}_j \in \mathcal{I}_o^k. \tag{5.45}$$

This implies that the variables $X_{\mathcal{I}_j}$ are independent of σ^2. Now, we can express

$$(N - k - 2)\ln\hat{\sigma}_{\mathcal{I}_j}^2 = \ln X_{\mathcal{I}_j}^{N-k-2} + (N - k - 2)\ln\sigma^2, \tag{5.46}$$

and similarly by defining $X_{\mathcal{S}} = \hat{\sigma}_{\mathcal{S}}^2/\sigma^2$ we get

$$(N - k_0 - 2)\ln\hat{\sigma}_{\mathcal{S}}^2 = \ln X_{\mathcal{S}}^{N-k_0-2} + (N - k_0 - 2)\ln\sigma^2. \tag{5.47}$$

Using (5.46) and (5.47) in (5.44) and after exponentiation we get

$$\left(\frac{X_{\mathcal{S}}^{N-k_0-2}}{X_{\mathcal{I}_j}^{N-k-2}}\right)\left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)}\left(\frac{2\pi}{\hat{\sigma}_0^2}\right)^{\Delta} < \left(\frac{1}{\sigma^2}\right)^{\Delta}. \tag{5.48}$$

Let $E^k_{\mathcal{I}_j}$ denote the entire left hand-side and let η_k denote the right-hand side of the inequality in (5.48). Let $\mathcal{I}^* \in \mathcal{I}^k_o$ denote the subset that produces the maximum value of $E^k_{\mathcal{I}_j}$ among all such subsets $\mathcal{I}_j \in \mathcal{I}^k_o$. Then, let us denote

$$E^k_{\mathcal{I}^*} = \max_{\mathcal{I}_j \in \mathcal{I}^k_o} \left\{ E^k_{\mathcal{I}_j} \right\}, \quad j = 1, 2, \ldots, \binom{p-k_0}{\Delta}. \tag{5.49}$$

The condition $\mathcal{C}_1$ in (5.40) is satisfied as $\sigma^2 \to 0$ under the event $E^k_{\mathcal{I}^*} < \eta_k$, for all $k = k_0 + 1, \ldots, K$. Now, we can express the probability that $E^k_{\mathcal{I}^*} < \eta_k$ as follows

$$\begin{aligned}
\Pr\left(E^k_{\mathcal{I}^*} < \eta_k\right) &= \Pr\left\{ \bigcap_{j=1}^{\binom{p-k_0}{\Delta}} \left(E^k_{\mathcal{I}_j} < \eta_k\right) \right\} \\
&= 1 - \Pr\left\{ \bigcup_{j=1}^{\binom{p-k_0}{\Delta}} \left(E^k_{\mathcal{I}_j} > \eta_k\right) \right\} \\
&\geq 1 - \binom{p-k_0}{\Delta} \Pr\left(E^k_{\mathcal{I}_j} > \eta_k\right) \\
\implies \Pr\left(E^k_{\mathcal{I}^*} > \eta_k\right) &\leq \binom{p-k_0}{\Delta} \Pr\left(E^k_{\mathcal{I}_j} > \eta_k\right),
\end{aligned} \tag{5.50}$$

where the inequality follows from the union bound. Now consider the following probability $\Pr\left\{E^k_{\mathcal{I}_j} > \eta_k\right\}$ for any arbitrary subset $\mathcal{I}_j \in \mathcal{I}^k_o$, which can be expressed as

$$\Pr\left\{ \left(\frac{X^{N-k_0-2}_{\mathcal{S}}}{X^{N-k-2}_{\mathcal{I}_j}} \right) \left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)} \left(\frac{2\pi}{\hat{\sigma}^2_0}\right)^{\Delta} > \left(\frac{1}{\sigma^2}\right)^{\Delta} \right\}. \tag{5.51}$$

Let $W = X^{N-k_0-2}_{\mathcal{S}} / X^{N-k-2}_{\mathcal{I}_j}$. Notice that the random variable W is independent of the noise variance σ^2 and since N is fixed W is bounded as $\sigma^2 \to 0$. Furthermore, $\lim_{\sigma^2 \to 0} \hat{\sigma}^2_0 = c$ (see Appendix 5.B) and the right-hand side of the inequality in (5.51) grows unbounded as $\sigma^2 \to 0$. Thus, we have

$$\lim_{\sigma^2 \to 0} \Pr\left\{E^k_{\mathcal{I}_j} > \eta_k\right\} = 0. \tag{5.52}$$

Therefore, using (5.50) and the result in (5.52), we have

$$\lim_{\sigma^2 \to 0} \Pr\left(E^k_{\mathcal{I}^*} > \eta_k\right) = 0, \qquad \forall\ k = k_0 + 1, \ldots, K. \tag{5.53}$$

Finally, using the union bound, and the result in (5.53), we get

$$\begin{aligned}\Pr\{\mathcal{C}_1\} &= \Pr\left\{\bigcap_{k=k_0+1}^{K} E^k_{\mathcal{I}^*} < \eta_k\right\} \\ &\geq 1 - \sum_{k=k_0+1}^{K} \Pr\left\{E^k_{\mathcal{I}^*} > \eta_k\right\} \to 1, \end{aligned} \tag{5.54}$$

as $\sigma^2 \to 0$.

(*b*) *Misfitting case* ($\mathcal{S} \not\subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$): Let $\mathcal{I}_j$ be any arbitrary jth subset belonging to the set of misfitted subsets of dimension k, i.e., $\mathcal{I}^k_m$. We consider the following inequality

$$\text{EBIC}_\text{R}(\mathcal{S}) < \text{EBIC}_\text{R}(\mathcal{I}_j), \quad \mathcal{I}_j \in \mathcal{I}^k_m, \tag{5.55}$$

where $j = 1, \ldots, t$. Here, t denotes the total number of subsets in the set $\mathcal{I}^k_m$ and $t = \binom{p}{k}$ if $k < k_0$, otherwise $t = \binom{p}{k} - \binom{p-k_0}{\Delta}$ if $k \geq k_0$, where $\Delta = k - k_0$. Denoting $X_\mathcal{S} = \hat{\sigma}^2_\mathcal{S}/\sigma^2$, rearranging and applying exponentiation we can express (5.55) as

$$\left(\frac{X^{N-k_0-2}_\mathcal{S}}{(\hat{\sigma}^2_{\mathcal{I}_j})^{N-k-2}}\right)\left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)}\left(\frac{2\pi}{\hat{\sigma}^2_0}\right)^{\Delta} < \left(\frac{1}{\sigma^2}\right)^{N-k_0-2}. \tag{5.56}$$

Similar to the overfitting case, let $E^k_{\mathcal{I}_j}$ denote the entire left-hand side and η the right-hand side of (5.56). Also, let $E^k_{\mathcal{I}^*} = \max\limits_{\mathcal{I}_j \in \mathcal{I}^k_m}\left\{E^k_{\mathcal{I}_j}\right\}$ for $j = 1, \ldots, t$, where $\mathcal{I}^*$ is the subset that leads to the maximum value of $E^k_{\mathcal{I}_j}$ among all such subsets of dimension k. The condition $\mathcal{C}_2$ in (5.41) is satisfied as $\sigma^2 \to 0$ under the event $E^k_{\mathcal{I}^*} < \eta$, for all $k = 1, \ldots, K$. Now, we can express the probability that $E^k_{\mathcal{I}^*} < \eta$ as

$$\begin{aligned}\Pr\left(E^k_{\mathcal{I}^*} < \eta\right) &= \Pr\left\{\bigcap_{j=1}^{t}\left(E^k_{\mathcal{I}_j} < \eta\right)\right\} \\ \implies \Pr\left(E^k_{\mathcal{I}^*} > \eta\right) &\leq t\Pr\left(E^k_{\mathcal{I}_j} > \eta\right), \end{aligned} \tag{5.57}$$

where the inequality follows from the union bound. Now consider the following probability for any arbitrary subset $\mathcal{I}_j \in \mathcal{I}^k_m$

$$\Pr\left(E^k_{\mathcal{I}_j} > \eta\right) = \Pr\left\{\left(\frac{X^{N-k_0-2}_\mathcal{S}}{(\hat{\sigma}^2_{\mathcal{I}_j})^{N-k-2}}\right)\left(\frac{1}{N}\right)^{\Delta(1+2\zeta d)}\left(\frac{2\pi}{\hat{\sigma}^2_0}\right)^{\Delta} > \left(\frac{1}{\sigma^2}\right)^{N-k_0-2}\right\}. \tag{5.58}$$

Here, $X^{N-k_0-2}_\mathcal{S}$ is independent of σ^2 and N is fixed, therefore $X^{N-k_0-2}_\mathcal{S}$ is bounded as $\sigma^2 \to 0$. Also $\hat{\sigma}^2_{\mathcal{I}_j} \to \|\mathbf{\Pi}^\perp_{\mathcal{I}_j}\mathbf{A}_\mathcal{S}\mathbf{x}_\mathcal{S}\|^2_2/N$ in probability as $\sigma^2 \to 0$ and since we are in the misfitting scenario, from Lemma 5.4 in Appendix 5.A we have

$\|\mathbf{\Pi}_{\mathcal{I}_j}^{\perp}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2/N > 0$. Furthermore, $\lim_{\sigma^2\to 0}\hat{\sigma}_0^2 = \text{const.}$ (see Appendix 5.B) and the right-hand side of the inequality in (5.58) grows unbounded as $\sigma^2 \to 0$. Hence,

$$\lim_{\sigma^2\to 0}\Pr\left\{E_{\mathcal{I}_j}^k > \eta\right\} = 0. \tag{5.59}$$

Using (5.57) and the result in (5.59) we get

$$\lim_{\sigma^2\to 0}\Pr\left\{E_{\mathcal{I}^*}^k > \eta\right\} = 0, \quad \forall\ k = 1,\ldots,K. \tag{5.60}$$

Finally, using the union bound and the result in (5.60), we get

$$\Pr\{\mathcal{C}_2\} \geq 1 - \sum_{k=1}^{K}\Pr\left\{E_{\mathcal{I}^*}^k > \eta\right\} \to 1 \quad \text{as} \quad \sigma^2 \to 0. \tag{5.61}$$

From (5.54) and (5.61) we can conclude that EBIC_R is consistent as $\sigma^2 \to 0$, which proves Theorem 1.

5.4.3 Consistency as $N \to \infty$ when σ^2 is Fixed $(0 < \sigma^2 < \infty)$

In this section, we prove the consistency of EBIC_R as the sample size $N \to \infty$ given that σ^2 is fixed and under the setting $p = N^d$ for some $d > 0$. This leads to the following theorem.

Theorem 5.2 *Assume that $p = N^d$ for some constant $d > 0$, the SNR is fixed and the matrix $\mathbf{A}$ satisfies (5.42). If $K \geq k_0$, then $\Pr\{EBIC_R(\mathcal{S}) < EBIC_R(\mathcal{I})\} \to 1$ as $N \to \infty$ for all $\mathcal{I} \neq \mathcal{S}$ and $card(\mathcal{I}) = 1,\ldots,K$ under the condition $\zeta > 1 - 1/2d$.*

Proof. As in the previous section, we have two parts of the proof. Part (a) is the overfitting case where we show that $\Pr(\mathcal{C}_1) \to 1$ as $N \to \infty$ and part (b) is the misfitting case where we show that $\Pr(\mathcal{C}_2) \to 1$ as $N \to \infty$.
(a) *Overfitting case* $(\mathcal{S} \subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}})$: Let $\mathcal{I}_j \in \mathcal{I}_o^k$ be any overfitted subset of dimension k. Consider the following inequality

$$\text{EBIC}_\text{R}(\mathcal{I}_j) > \text{EBIC}_\text{R}(\mathcal{S}), \quad \mathcal{I}_j \in \mathcal{I}_o^k. \tag{5.62}$$

Denoting $\Delta = k - k_0$ and rearranging (5.62) we get

$$(N-k-2)\ln\left(\frac{\hat{\sigma}_{\mathcal{I}_j}^2}{\hat{\sigma}_{\mathcal{S}}^2}\right) + \Delta(1+2\zeta d)\ln N + \Delta\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{S}}^2}\right) - \Delta\ln 2\pi > 0. \tag{5.63}$$

Let $E_{\mathcal{I}_j}^k$ denote the entire left side of the inequality (5.63) and $\mathcal{I}^*$ denote the subset that leads to the minimum value of $E_{\mathcal{I}_j}^k$ among all such subsets of dimension k. Hence,

$$E_{\mathcal{I}^*}^k = \min_{\mathcal{I}_j\in\mathcal{I}_o^k}\left\{E_{\mathcal{I}_j}^k\right\}, \quad j = 1,2,\ldots,\binom{p-k_0}{\Delta}. \tag{5.64}$$

The condition $\mathcal{C}_1$ in (5.40) is satisfied as $N \to \infty$ under the event $E^k_{\mathcal{I}^*} > 0$, for all $k = k_0 + 1, \ldots, K$. Expanding the ratio we have

$$\begin{aligned}
\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right) &= \ln\left(\frac{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}}{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right) \\
&= \ln\left[\frac{\mathbf{e}^T\left(\mathbf{I} - \mathbf{\Pi}_{\mathcal{I}_j} + \mathbf{\Pi}_{\mathcal{S}} - \mathbf{\Pi}_{\mathcal{S}}\right)\mathbf{e}}{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right] \\
&= \ln\left(\frac{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e} - \mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j\setminus\mathcal{S}}\mathbf{e}}{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right) \\
&= \ln\left(1 - \frac{\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j\setminus\mathcal{S}}}{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}\right), \qquad (5.65)
\end{aligned}$$

where $\mathbf{\Pi}_{\mathcal{I}_j\setminus\mathcal{S}} = \mathbf{\Pi}_{\mathcal{I}_j} - \mathbf{\Pi}_{\mathcal{S}}$. Now we can write

$$\min_{1\le j\le T}\left\{(N-k-2)\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right)\right\} = (N-k-2)\ln\left[1 - \frac{\max\limits_{1\le j\le T}\left\{\left(\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j\setminus\mathcal{S}}\mathbf{e}\right)/\sigma^2\right\}}{\left(\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}\right)/\sigma^2}\right], \qquad (5.66)$$

where $T = \binom{p-k_0}{\Delta}$. Now the term, $(\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j\setminus\mathcal{S}}\mathbf{e})/\sigma^2 \sim \chi^2_{\Delta}$ (see Appendix 5.C). Then from Lemma 5.2 in Appendix 5.A we have the following upper bound

$$\max_{1\le j\le T}\left\{(\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j\setminus\mathcal{S}}\mathbf{e})/\sigma^2\right\} \le \Delta + 2\sqrt{\Delta\psi\ln T} + 2\psi\ln T, \qquad (5.67)$$

with probability approaching one as $N \to \infty$ if $\psi > 1$. Now, for sufficiently large $p = N^d$ we can write $\ln T = \ln\binom{p-k_0}{\Delta} \approx \Delta d\ln N$. This gives

$$\begin{aligned}
\max_{1\le j\le T}\left\{(\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j\setminus\mathcal{S}}\mathbf{e})/\sigma^2\right\} &\le \Delta + 2\Delta\sqrt{\psi d\ln N} + 2\psi\Delta d\ln N \\
&= 2\psi\Delta d\ln N\left(1 + \frac{1}{\sqrt{\psi d\ln N}} + \frac{1}{2\psi d\ln N}\right) \\
&\approx 2\psi\Delta d\ln N, \qquad (5.68)
\end{aligned}$$

as N grows large. Furthermore, the term in the denominator in (5.66), $(\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e})/\sigma^2 \sim \chi^2_{N-k_0}$ and based on the law of large numbers tends to $N - k_0 \approx N$. Therefore, using (5.68) in (5.66) and $(N-k-2) \approx N$ under the large-N approximation we get

$$\begin{aligned}
\min_{1\le j\le T}\left\{N\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right)\right\} &\ge N\ln\left(1 - \frac{2\Delta\psi d\ln N}{N}\right) \\
&\approx -2\Delta\psi d\ln N, \qquad (5.69)
\end{aligned}$$

where the last approximation follows by linearization of the logarithm for small $2\Delta\psi d \ln N/N$ value. Thus, we can write

$$E^k_{\mathcal{I}^*} \geq -2\Delta\psi d \ln N + \Delta(1+2\zeta d)\ln N + \Delta \ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{S}}}\right) - \Delta \ln 2\pi$$
$$= \Delta\left(1+2\zeta d - 2\psi d\right)\ln N + \Delta \ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{S}}}\right) - \Delta \ln 2\pi. \tag{5.70}$$

Since $\lim_{N\to\infty} \hat{\sigma}^2_0 = \text{const.} > 0$ (see Appendix 5.B) and $\lim_{N\to\infty} \hat{\sigma}^2_{\mathcal{S}} = \sigma^2$ (see Appendix 5.C), $E^k_{\mathcal{I}^*} \to \infty$ as $N \to \infty$ for all $k = k_0+1, \ldots, K$ under the condition $1 + 2\zeta d - 2\psi d > 0$ for any $\psi > 1$. Hence, the lower bound on ζ becomes

$$\boxed{\zeta > 1 - \frac{1}{2d}}. \tag{5.71}$$

From the above analysis, we can say that

$$\lim_{N\to\infty} \Pr\left\{E^k_{\mathcal{I}^*} < 0\right\} = 0, \quad \forall\ k = k_0+1, \ldots, K. \tag{5.72}$$

Finally, using the union bound and the result in (5.72) we can express the probability of $\mathcal{C}_1$ (5.40) happening as

$$\Pr\{\mathcal{C}_1\} = \Pr\left\{\bigcap_{k=k_0+1}^{K} E^k_{\mathcal{I}^*} > 0\right\}$$
$$\geq 1 - \sum_{k=k_0+1}^{K} \Pr\left\{E^k_{\mathcal{I}^*} < 0\right\} \to 1 \tag{5.73}$$

as $N \to \infty$.

(*b*) *Misfitting case* ($\mathcal{S} \not\subset \hat{\mathcal{S}}_{\text{EBIC}_\text{R}}$): Let $\mathcal{I}_j \in \mathcal{I}^k_m$ be any misfitted subset of dimension k. Consider the following inequality

$$\text{EBIC}_\text{R}(\mathcal{I}_j) > \text{EBIC}_\text{R}(\mathcal{S}), \quad \mathcal{I}_j \in \mathcal{I}^k_m. \tag{5.74}$$

Denoting $\Delta = k - k_0$ and rearranging (5.74) we get

$$(N-k-2)\ln\left(\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}\right) + (1+2\zeta d)\Delta \ln N + \Delta \ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{S}}}\right) + \Delta \ln\left(\frac{1}{2\pi}\right) > 0. \tag{5.75}$$

Let $E^k_{\mathcal{I}_j}$ denote the entire left hand side of the inequality in (5.75) and $\mathcal{I}^*$ denote the subset that generates the minimum value of $E^k_{\mathcal{I}_j}$ among all such subsets of dimension k. Then we have

$$E^k_{\mathcal{I}^*} = \min_{\mathcal{I}_j \in \mathcal{I}^k_m} \left\{E^k_{\mathcal{I}_j}\right\}, \qquad j = 1, 2, \ldots, T, \tag{5.76}$$

where $T = \binom{p}{k}$ if $k < k_0$ otherwise $T = \binom{p}{k} - \binom{p-k_0}{k}$ if $k \geq k_0$. The condition $\mathcal{C}_2$ in (5.41) is satisfied as $N \to \infty$ under the event $E^k_{\mathcal{I}^*} > 0$, for all $k = 1, \ldots, K$. Now, let $\mathbf{u} = \mathbb{E}[\mathbf{y}] = \mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}$. Using this, the ratio $\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}}$ can be expanded as

$$\begin{aligned}\frac{\hat{\sigma}^2_{\mathcal{I}_j}}{\hat{\sigma}^2_{\mathcal{S}}} &= \frac{\mathbf{y}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{y}}{\mathbf{y}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{y}} = \frac{(\mathbf{u}+\mathbf{e})^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}(\mathbf{u}+\mathbf{e})}{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}} \\ &= \frac{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u} + 2\sigma\sqrt{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}}\cdot Z_j + \mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}}{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}}, \end{aligned} \tag{5.77}$$

where

$$Z_j = \frac{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}}{\sigma\sqrt{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}}} \sim \mathcal{N}(0,1). \tag{5.78}$$

Now

$$\begin{aligned}\min_{1\leq j\leq T}\left\{\hat{\sigma}^2_{\mathcal{I}_j}/\hat{\sigma}^2_{\mathcal{S}}\right\} &= \min_{1\leq j\leq T}\left\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u} + 2\sigma\sqrt{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}}\cdot Z_j + \mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}\right\}\Big/\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e} \\ &\geq \left[\min_{1\leq j\leq T}\left\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\right\} + \sigma^2\min_{1\leq j\leq T}\left\{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}/\sigma^2\right\}\right. \\ &\quad \left. - 2\sigma\sqrt{\max_{1\leq j\leq T}\left\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\right\}}\cdot\max_{1\leq j\leq T}\left\{Z_j\right\}\right]\Big/\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e}. \end{aligned} \tag{5.79}$$

In the misfitting scenario we have two cases: (i) $k < k_0$ (ii) $k \geq k_0$. We consider the case (i) in our further analysis, which also encapsulates case (ii). For $k < k_0$ we have $\ln T = \ln\binom{p}{k} \approx kd\ln N$. Therefore, using the result in Lemma 5.2 we have the following lower bound under large-N approximation

$$\begin{aligned}\min_{1\leq j\leq T}\left\{\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{e}/\sigma^2\right\} &= \mathbf{e}^T\mathbf{e}/\sigma^2 - \max_{1\leq j\leq T}\left\{\mathbf{e}^T\mathbf{\Pi}_{\mathcal{I}_j}\mathbf{e}/\sigma^2\right\} \\ &\geq N - 2\psi' kd\ln N, \end{aligned} \tag{5.80}$$

where $\psi' > 1$ and $\mathbf{e}^T\mathbf{e}/\sigma^2 \approx N$ for large-N. Furthermore, from the result in Lemma 5.3 we have the following upper bound

$$\max_{1\leq j\leq T}\{Z_j\} \leq \sqrt{2\psi' kd\ln N}, \tag{5.81}$$

where $\psi' > 1$. Now, let $C_{\min} = \min_{1\leq j\leq T}\left\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\right\}$ and $C_{\max} = \max_{1\leq j\leq T}\left\{\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}_j}\mathbf{u}\right\}$. Also as $N \to \infty$ we can approximate $(N-k-2) \approx N$ and $\mathbf{e}^T\mathbf{\Pi}^{\perp}_{\mathcal{S}}\mathbf{e} \approx \sigma^2 N$. Using

this, and the results in (5.80) and (5.81) we get

$$\begin{aligned}\min_{1\le j\le T}\left\{N\ln\left(\frac{\hat\sigma^2_{\mathcal{I}_j}}{\hat\sigma^2_{\mathcal{S}}}\right)\right\} &= N\ln\left[\min_{1\le j\le T}\left\{\frac{\hat\sigma^2_{\mathcal{I}_j}}{\hat\sigma^2_{\mathcal{S}}}\right\}\right]\\ &\ge N\ln\Bigg[\Bigg\{C_{\min}-2\sigma\sqrt{C_{\max}}\cdot\sqrt{2\psi' kd\ln N}\\ &\qquad +\sigma^2\left(N-2\psi' kd\ln N\right)\Bigg\}\Big/\sigma^2 N\Bigg]. \end{aligned}\tag{5.82}$$

Now, observe that $C_{\min}=\mathbf{u}^T\mathbf{\Pi}^{\perp}_{\mathcal{I}^*}\mathbf{u}=\mathbf{x}^T_{\mathcal{S}}\mathbf{A}^T_{\mathcal{S}}\mathbf{\Pi}^{\perp}_{\mathcal{I}^*}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}$. Since, we are in the misfitting scenario, from Lemma 5.4, in Appendix 5.A, we can express $C_{\min}=Nb_{\min}$ where $b_{\min}=\mathcal{O}(1)>0$. Similarly, $C_{\max}=Nb_{\max}$ where $b_{\max}=\mathcal{O}(1)>0$ and $0<b_{\min}\le b_{\max}$. Hence, we can rewrite (5.82) as

$$\begin{aligned}&\min_{1\le j\le T}\left\{N\ln\left(\frac{\hat\sigma^2_{\mathcal{I}_j}}{\hat\sigma^2_{\mathcal{S}}}\right)\right\}\ge\\ &N\ln\left(1+\frac{b_{\min}}{\sigma^2}-\frac{2\sqrt{b_{\max}}}{\sigma}\sqrt{\frac{2\psi' kd\ln N}{N}}-\frac{2\psi' kd\ln N}{N}\right)\\ &\approx N\ln\left(1+\frac{b_{\min}}{\sigma^2}\right)\end{aligned}\tag{5.83}$$

as N grows large. For $k<k_0$, we get $\Delta<0$, therefore, in this case we have

$$E^k_{\mathcal{I}^*}\ge N\ln\left(1+\frac{b_{\min}}{\sigma^2}\right)-|\Delta|(1+2\zeta d)\ln N-|\Delta|\ln\left(\frac{\hat\sigma^2_0}{2\pi\hat\sigma^2_{\mathcal{S}}}\right)\to\infty\tag{5.84}$$

as $N\to\infty$ for all $k=1,\ldots,K$, since $N\ln(1+b_{\min}/\sigma^2)$ is the dominating term as it tends to infinity much faster than the $\ln N$ term and $\lim_{N\to\infty}\hat\sigma^2_0=\text{const.}>0$ (see Appendix 5.B) and $\lim_{N\to\infty}\hat\sigma^2_{\mathcal{S}}=\sigma^2$ (see Appendix 5.C). From the above analysis we can say that

$$\lim_{N\to\infty}\Pr\left\{E^k_{\mathcal{I}^*}<0\right\}=0,\quad\forall\ k=1,\ldots,K.\tag{5.85}$$

Finally, using the union bound and the result in (5.85) we can express the probability of $\mathcal{C}_2$ (5.41) happening as

$$\begin{aligned}\Pr\{\mathcal{C}_2\}&=\Pr\left\{\bigcap_{k=1}^{K}E^k_{\mathcal{I}^*}>0\right\}\\ &\ge 1-\sum_{k=1}^{K}\Pr\left\{E^k_{\mathcal{I}^*}<0\right\}\to 1\quad\text{as}\quad N\to\infty.\end{aligned}\tag{5.86}$$

From (5.73) and (5.86) we can conclude that EBIC_{R} is consistent as $N\to\infty$, which proves Theorem 2.

5.4.4 Discussion on the Hyperparameter ζ

If ζ is too large, it will lead to underfitting issues. This is evident from (5.84) where a large value of ζ may force the overall sum to become negative, especially for smaller N values. On the contrary, if ζ is too small, it will lead to overfitting issues as the penalty may not be sufficiently large to compensate for the overparameterization due to large parameter space. Choosing a near-optimal value of ζ is quite crucial in order to select the true model or at least a model very close to the true one. The lower bound on ζ given in (5.71), provides some guideline on the range of possible ζ value to pick from. For example, for a given data, if $d \approx 1.5$, then $\zeta > 0.667$. However, since there is no upper bound on ζ, it can take any value greater than 0.667. Ideally, a rule of thumb is not to set $\zeta > 1$. As a future direction, it will be interesting to investigate novel ways to choose ζ in a more data-driven fashion.

5.5 Predictor Selection Algorithms

In the high-dimensional scenario, when p is large, it is infeasible to perform model selection in the conventional manner. For a design matrix with parameter dimension p, the number of possible candidate models is $2^p - 1$. Hence, the candidate model space grows exponentially with p and we cannot afford to calculate model score for all possible models. Therefore, to perform model selection, we combine a model selection criterion with a predictor selection (support recovery) algorithm such as OMP or LASSO (least absolute shrinkage and selection operator) [28]. The goal of predictor selection is to pick a subset of important predictors from the entire set of p predictors. In this context, the most important predictors refer to the positions of the nonzero elements of the input signal $\mathbf{x}$. Thus, predictor selection reduces the cardinality of the candidate model space to some upper bound K such that $k_0 \leq K \ll N$ under the assumption of a sparse parameter vector. This enables us to apply the model selection criterion on the smaller set of candidate models to pick the best model. The OMP algorithm is shown in Algorithm 5.1. To perform model selection, we combine OMP with EBIC_R as shown in Algorithm 5.2.

Algorithm 5.1 OMP with K iterations

1: **Inputs:** Design matrix $\mathbf{A}$, measurement vector $\mathbf{y}$.
2: **Initialization:** $\|\mathbf{a}_j\|_2 = 1\ \forall j$, $\mathbf{r}^0 = \mathbf{y}$, $\mathcal{S}^0_{\text{OMP}} = \emptyset$
3: **for** $i = 1$ to K **do**
4: Find next column index: $d^i = \arg\max_j \left|\mathbf{a}_j^T \mathbf{r}^{i-1}\right|$
5: Add current index: $\mathcal{S}^i_{\text{OMP}} = \mathcal{S}^{i-1}_{\text{OMP}} \cup \{d^i\}$
6: Update residual: $\mathbf{r}^i = \left(\mathbf{I}_N - \mathbf{\Pi}_{\mathcal{S}^i_{\text{OMP}}}\right)\mathbf{y}$
7: **end for**
8: **Output:** OMP generated index sequence $\mathcal{S}^K_{\text{OMP}}$

Algorithm 5.2 Model selection combining EBIC_R with OMP

1: Run OMP for K iterations to obtain $\mathcal{S}_{\text{OMP}}^{K}$
2: **for** $k = 1$ to K **do**
3: $\quad \mathcal{I} = \mathcal{S}_{\text{OMP}}^{k}$
4: $\quad$ Compute $\text{EBIC}_\text{R}(\mathcal{I})$
5: **end for**
6: Estimated true support: $\hat{\mathcal{S}}_{\text{EBIC}_\text{R}} = \arg\min_{\mathcal{I}} \{\text{EBIC}_\text{R}(\mathcal{I})\}$

LASSO is a shrinkage method for variable selection/estimation in linear regression models developed by Tibshirani [28]. Given the linear model in (5.1), the LASSO solution for $\mathbf{x}$ for a particular choice of the regularization parameter $\lambda \geq 0$ is obtained as

$$\hat{\mathbf{x}}_{\text{lasso}}(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \frac{1}{2N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \right\}, \tag{5.87}$$

where $\|\cdot\|_1$ denotes the l_1 norm. The parameter λ determines the level of sparsity. When $\lambda \to \infty$ the objective function in (5.87) attains the minimum with $\hat{\mathbf{x}}_{\text{lasso}}(\lambda)$ being a zero vector. As we gradually lower the λ value, the number of non-zero components in $\hat{\mathbf{x}}_{\text{lasso}}(\lambda)$ starts increasing. Model selection combining LASSO and EBIC_R can be performed as shown in Algorithm 5.3. Gradually decrease λ from a high value so that the number of non-zero components in $\hat{\mathbf{x}}_{\text{lasso}}(\lambda)$ gradually increases. Therefore, for each decreasing unique value of λ say λ_i, we acquire a different solution $\hat{\mathbf{x}}_{\text{lasso}}(\lambda_i)$, with increasing support and thus obtaining a sequence of candidate models with maximum cardinality K. The value of EBIC_R is computed for each of the candidate models and the model corresponding to the smallest EBIC_R score is selected as the final model. A most useful method for solving LASSO in our context is the (modified) least angle regression (LARS) algorithm [30], since it also provides the required sequence of regularization parameters for which the support changes.

Algorithm 5.3 Model selection combining EBIC_R with LASSO

1: Compute LASSO estimates $\{\hat{\mathbf{x}}_{\text{lasso}}(\lambda_1), \ldots, \hat{\mathbf{x}}_{\text{lasso}}(\lambda_{K_{\max}})\}$ where $card(supp\,(\hat{\mathbf{x}}_{\text{lasso}}(\lambda_{K_{\max}}))) = K$
2: **for** $i = 1$ to $K_{\max}$ **do**
3: $\quad \mathcal{I} = supp\,(\hat{\mathbf{x}}_{\text{lasso}}(\lambda_i))$
4: $\quad$ Compute $\text{EBIC}_\text{R}(\mathcal{I})$
5: **end for**
6: Estimated true support: $\hat{\mathcal{S}}_{\text{EBIC}_\text{R}} = \arg\min_{\mathcal{I}} \{\text{EBIC}_\text{R}(\mathcal{I})\}$

5.6 Simulation Results

In this section, we provide numerical simulation results to illustrate the empirical performance of EBIC_R. The performance of EBIC_R is compared with the 'oracle', EBIC, EFIC and MBT. However, the performance comparison with the RRT [65] method is dropped since it behaves quite similar to MBT (see [70] for details). The 'oracle' criterion assumes *a priori* knowledge of the true cardinality k_0. Thus, the model selection performance of the 'oracle' provides the upper bound on the maximum model selection performance that can be achieved using a particular predictor selection algorithm and for a given set of data settings. Additionally, we also provide simulation results to highlight the drawbacks of classical methods for model selection in high-dimensional linear regression models with a sparse parameter vector.

5.6.1 General Simulation Setup

In the simulations, we consider the model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where the design matrix $\mathbf{A} \in \mathbb{R}^{N \times p}$ is generated with independent entries following normal distribution $\mathcal{N}(0, 1)$. Since $\mathbf{x}$ is assumed to be sparse, we choose $k_0 = 5$. Furthermore, without loss of generality, we assume that the true support is $\mathcal{S} = [1, 2, 3, 4, 5]$, therefore, $\mathbf{x}_\mathcal{S} = [x_1, x_2, x_3, x_4, x_5]^T$ and $\mathbf{A}_\mathcal{S} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{a}_5]$. This implies that the elements of $\mathbf{x}$ follows $x_k \neq 0$ for $k = 1, \ldots, k_0$ and $x_k = 0$ for $k > k_0$. The SNR in dB is $\text{SNR (dB)} = 10 \log_{10}(\sigma_s^2/\sigma^2)$, where σ_s^2 and σ^2 denote signal and true noise power, respectively. The signal power is computed as $\sigma_s^2 = ||\mathbf{A}_\mathcal{S}\mathbf{x}_\mathcal{S}||_2^2/N$. Based on σ_s^2 and the chosen SNR (dB), the noise power is set as $\sigma^2 = \sigma_s^2/10^{\text{SNR (dB)}/10}$. Using this σ^2, the noise vector $\mathbf{e}$ is generated following $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. The probability of correct model selection (PCMS) is estimated over 1000 Monte Carlo trials. To maintain randomness in the data, a new design matrix $\mathbf{A}$ is generated at each Monte Carlo trial. OMP is used for predictor selection for its simplicity and wider range of applicability.

5.6.2 Tuning Parameter Selection

An important step in model selection is the choice of the tuning parameter. As mentioned earlier, too small or large values of the tuning parameter can cause severe performance degradation in certain scenarios. Fig. 5.1 shows a performance comparison of EBIC_R for four different values of ζ (0.4, 0.6, 1, and 2). Here, we set $p = N^d$ where $d = 1.1$. Hence, from Theorem 5.2 we require $\zeta > 1 - 1/2d = 0.55$ to achieve consistency. From the figure, we see that for $\zeta = 0.4$, the performance of EBIC_R degrades after a certain point with increasing N, which justifies the theory. For all other $\zeta > 0.55$, the performances improve with increasing N. For $\zeta = 0.6$, which is very close to the lower bound, the convergence to PCMS = 1 is slow and will require a very large sample size. For, $\zeta = 2$, the performance suffers (due to underfitting) in the low N regime, but does achieve perfect selection as

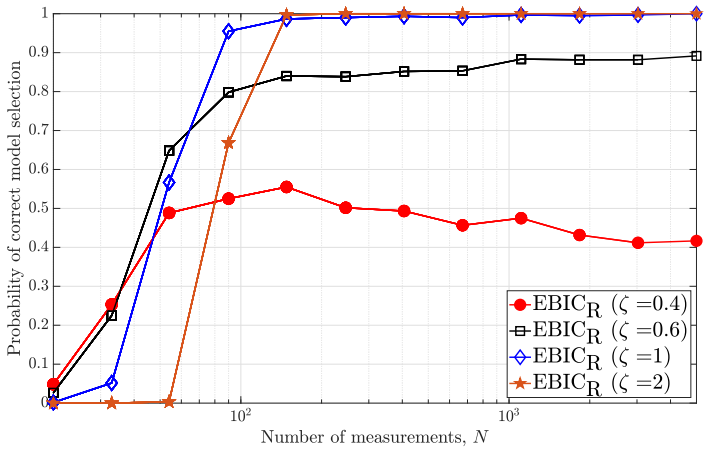


Figure 5.1: PCMS vs N with $\mathbf{x}_{\mathcal{S}} = [1, 1, 1, 1, 1]$, SNR = 5 dB, $p = N^d$ and $d = 1.1$.

N increases. In this case, $\zeta = 1$ provides a much better overall performance for a broader range of N. A similar trend as in EBIC_R is observed even in EBIC and EFIC for different choices of γ and c. Hence, to maintain fairness, the following tuning parameter settings are considered for further analysis: $\zeta = 1$ (EBIC_R), $c = 1$ (EFIC) and $\gamma = 1$ (EBIC). For MBT [70], $\lim_{N\to\infty} \text{PCMS} \to 1$ as $\beta \to 1$. Hence, we choose $\beta = 0.999$.

5.6.3 Model Selection with Classical Methods in High-Dimensional Setting

This section presents simulation results for model selection using classical methods in high-dimensional linear regression models and compares their performances with EBIC_R. The purpose of these results is to highlight the limitations of the classical methods in dealing with large-p small-N scenarios. The classical methods used here are BIC [47], $\widetilde{\text{BIC}}_{N,\text{SNR}}$ [16], BIC_R [79], gMDL [49], and PAL [51]. The classical methods used in the simulations are described below.

- **BIC**: The BIC was developed by Schwarz [47]. The BIC score for a model $\mathcal{I}$ is given by (5.18).

- **$\widetilde{\text{BIC}}_{N,\text{SNR}}$**: This combined high-SNR form of the BIC was proposed by Stoica and Babu [16] as a means to solve the high-SNR consistency requirement of BIC. For a model $\mathcal{I}$, the $\widetilde{\text{BIC}}_{N,\text{SNR}}$ score is

$$\widetilde{\text{BIC}}_{N,\text{SNR}}(\mathcal{I}) = N \ln \hat{\sigma}^2_{\mathcal{I}} + k \ln N - (k+2) \ln \hat{\sigma}^2_{\mathcal{I}} \tag{5.88}$$

- **BIC_R**: The BIC_R score for a model $\mathcal{I}$ is given as [79]

$$\text{BIC}_\text{R}(\mathcal{I}) = N \ln \hat{\sigma}_\mathcal{I}^2 + k \ln\left(\frac{N}{2\pi}\right) + (k+2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_\mathcal{I}^2}\right). \tag{5.89}$$

- **gMDL:** The criterion gMDL was developed by Hansen and Yu [49] and is based on the Bayesian mixture form of MDL. It is called gMDL for its use of the g-prior and is given by

$$\text{gMDL}(\mathcal{I}) = \left(\frac{N-k}{2}\right)\ln\left(\frac{N\hat{\sigma}_\mathcal{I}^2}{N-k}\right) + \frac{k}{2}\ln\left(\frac{\hat{R}_\mathcal{I}}{k}\right) + \ln N \tag{5.90}$$

  where $\hat{R}_\mathcal{I} = \mathbf{y}^T\mathbf{y} - N\hat{\sigma}_\mathcal{I}^2 = \mathbf{y}^T\mathbf{\Pi}_\mathcal{I}\mathbf{y}$ is the fitted sum of squares.

- **PAL:** The PAL criterion was developed by Stoica and Babu [51]. The PAL score for a model with support $\mathcal{I}$ is evaluated as

$$\text{PAL}(\mathcal{I}) = N \ln \hat{\sigma}_\mathcal{I}^2 + k \ln(p) \frac{\ln(r_\mathcal{I}+1)}{\ln(\rho_\mathcal{I}+1)}. \tag{5.91}$$

  Here $r_\mathcal{I} = N \ln\left(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}-1}^2\right)$ and $\rho_\mathcal{I} = N\ln\left(\hat{\sigma}_{\mathcal{I}-1}^2/\hat{\sigma}_K^2\right)$ where $\hat{\sigma}_{\mathcal{I}-1}^2$ and $\hat{\sigma}_K^2$ denotes the noise variance estimate for the previous and the last candidate model, respectively.

In the simulation, we consider the true parameter vector to be $\mathbf{x}_\mathcal{S} = [5,4,3,2,1]^T$. Fig. 5.2 illustrates the PCMS versus SNR in dB for fixed $N = 100$ and $p = 500$. This gives $d = \log(p)/\log(N) \approx 1.35$, hence, $\zeta > 1 - 1/2d \approx 0.63$. The first major observation from the figure is that EBIC_R ($\zeta = 1$) clearly outperforms all the

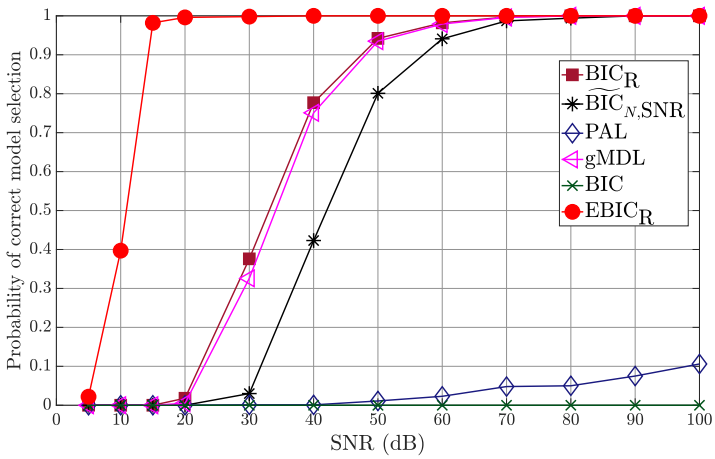
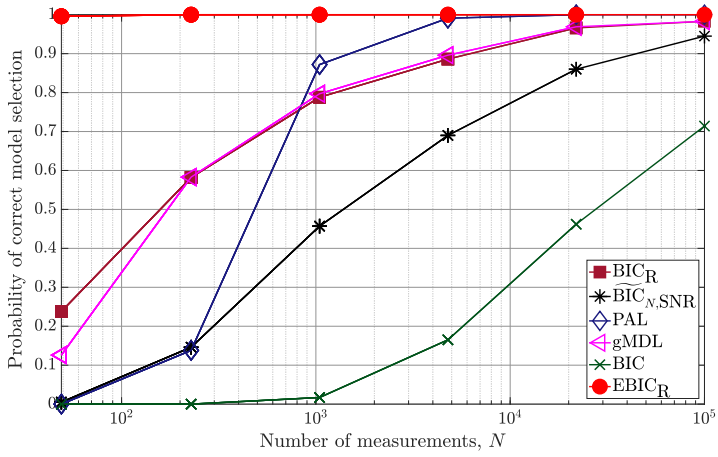
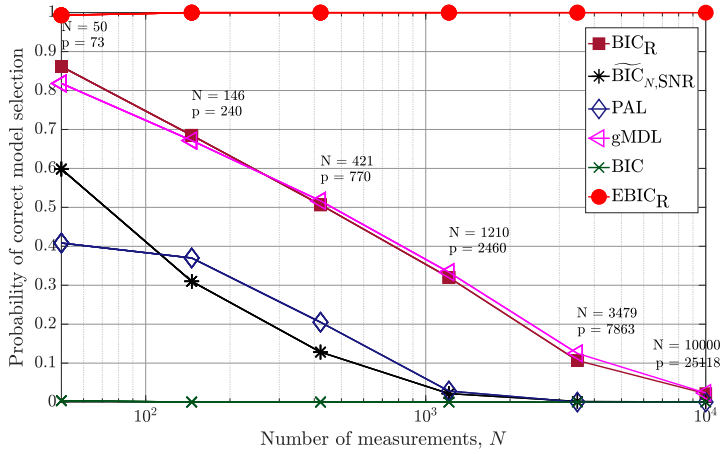


Figure 5.2: PCMS versus SNR (dB) for $N = 100$, $p = 500$ and $\mathbf{x}_\mathcal{S} = [5,4,3,2,1]$.



(a) $p = 500$



(b) $p = N^d$ where $d = 1.1$

Figure 5.3: PCMS versus N for SNR = 30 dB with $\mathbf{x}_{\mathcal{S}} = [5, 4, 3, 2, 1]$.

classical methods by a huge margin. Secondly, for the considered setting, the performances of $\mathrm{BIC_R}$ and gMDL are quite similar followed by $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$. The criteria $\mathrm{BIC_R}$, gMDL and $\widetilde{\mathrm{BIC}}_{N,\mathrm{SNR}}$ do achieve convergence to detection probability one but at the expense of very high values of SNR. The performances of PAL and BIC are extremely poor in this case, even in the high-SNR regions.

Fig. 5.3 presents the plot for PCMS versus N for two different settings of p. Fig. 5.3a corresponds to a fixed $p = 500$ and Fig. 5.3b to a varying $p = N^d$ where $d = 1.1$. The figures show that $\mathrm{EBIC_R}$ ($\zeta = 1$) clearly surpasses the classical methods

with huge differences in performance. Furthermore, both the figures clearly show distinctive behaviour of the classical methods that highly differ from each other. For the fixed p case, the PCMS for all the classical methods approaches one as N grows large. This is in tune with the fact that the classical methods are consistent when p is fixed and $N \to \infty$. On the contrary, when p is varying and grows with N, the consistency attribute does not hold any longer, hence, we see the decreasing performance trend in Fig. 5.3. However, this is not the case for EBIC_R and it achieves consistency in both cases. Another key point that we can deduce from this analysis is that when p is not much smaller than N, the convergence rate for the classical methods is much slower than EBIC_R even when $N > p$ (see Fig. 5.3a for $N > 500$). Thus, we can say that methods like EBIC_R are important and very necessary in scenarios where $N > p$ but p is not sufficiently smaller than N.

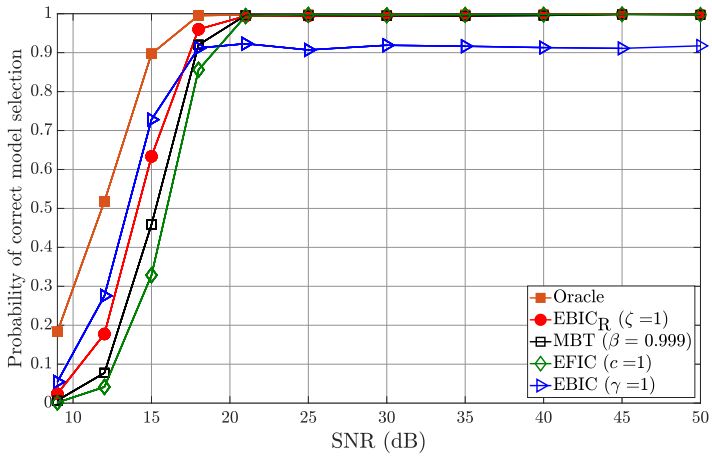
5.6.4 Model Selection with the Latest Methods in High-Dimensional Setting

In the previous section, we highlighted the drawbacks of classical methods in model selection under the high-dimensional setting. We observed that the performance of the classical methods collapses when p grows with N and the consistency property breaks down. In this section, we present simulation results for model selection comparing EBIC_R to the existing state-of-the-art methods, designed to deal with the large-p small-N scenarios.

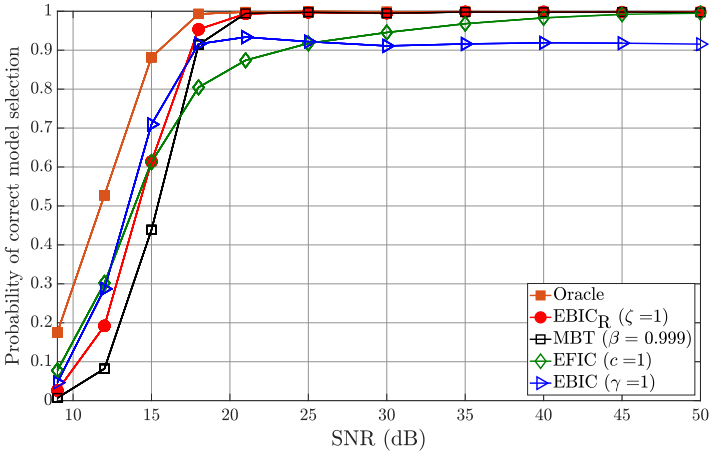
Model Selection versus SNR

To emphasize the scale-invariant and consistent behaviour of EBIC_R, we consider two scenarios. In the first scenario, we assume the true parameter vector to be $\mathbf{x}_\mathcal{S} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$ and in the second scenario, we assume $\mathbf{x}_\mathcal{S} = [50, 40, 30, 20, 10]^T$. Note that in the simulations we compute the noise variance σ^2 based on the chosen SNR level and the current signal power value $\sigma_s^2 = \|\mathbf{A}_\mathcal{S}\mathbf{x}_\mathcal{S}\|_2^2/N$. To simulate the PCMS versus SNR in a high-dimensional setting we fixed $N = 55$ and $p = 1000$. This gives $d = \log(p)/\log(N) \approx 1.724$, hence, $\zeta > 1 - 1/2d \approx 0.71$.

Fig. 5.4 shows the empirical PCMS versus SNR (dB). Fig. 5.4a and Fig. 5.4b correspond to $\mathbf{x}_\mathcal{S} = [0.05, 0.04, 0.03, 0.02, 0.01]$ and $\mathbf{x}_\mathcal{S} = [50, 40, 30, 20, 10]$, respectively. Both the figures depict fixed N increasing SNR scenario. Comparing the figures, the first clear observation is that unlike the other criteria, the behaviour of EFIC is not identical for the two different $\mathbf{x}_\mathcal{S}$ given that the other parameters viz, N, p and k_0 are constant and the performance is evaluated for the same SNR range. This illustrates the scaling problem present in EFIC that leads to either high underfitting or overfitting issues. This behavior or EFIC can be explained as follows. The data dependent penalty term (DDPT) of EFIC is $\text{DDPT} = -(k+2)\ln\|\mathbf{\Pi}_\mathcal{I}^\perp\mathbf{y}\|_2^2$, whose overall value depends on the value $\|\mathbf{\Pi}_\mathcal{I}^\perp\mathbf{y}\|_2^2$, which in turn is influenced by the signal and noise powers σ_s^2 and σ^2, respectively. If $\|\mathbf{\Pi}_\mathcal{I}^\perp\mathbf{y}\|_2^2 \ll 1$, then $\text{DDPT} \gg 0$,



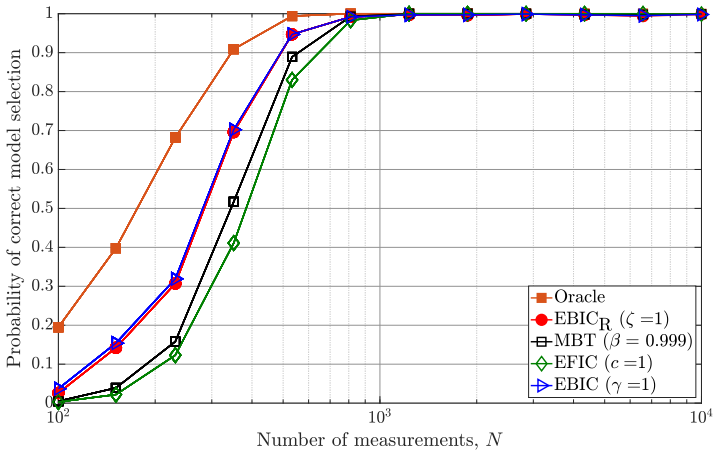
(a) $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$



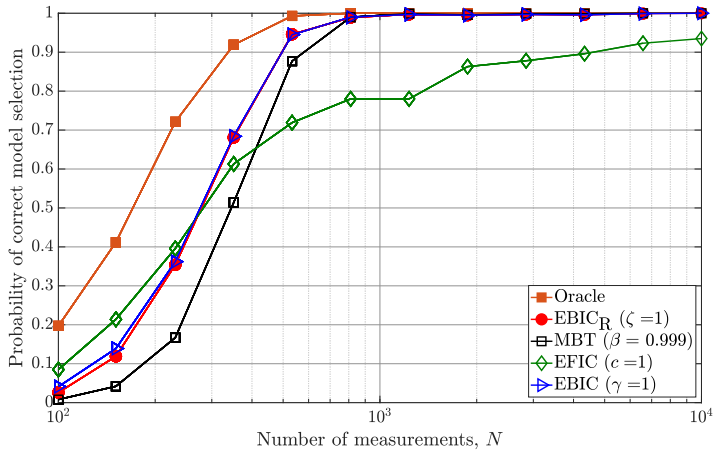
(b) $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$

Figure 5.4: PCMS versus SNR (dB) for $N = 55$ and $p = 1000$.

which may blow the overall penalty to a large value leading to underfitting issues. This is most likely the case when $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$ (Fig. 5.4a). On the contrary if $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2 \gg 1$, then DDPT $\ll 0$, thus lowering the overall penalty leading to overfitting issues (when $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$, Fig. 5.4b). The second major observation is that EBIC is inconsistent when SNR is high but N is small and fixed. This behaviour of EBIC is already reported in [18]. In general, EFIC, MBT (for $\beta \to 1$), and EBIC$_{\mathrm{R}}$ are consistent for increasing SNR scenarios with N fixed, but while EBIC$_{\mathrm{R}}$ and MBT are invariant to data-scaling EFIC is not.



(a) $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$



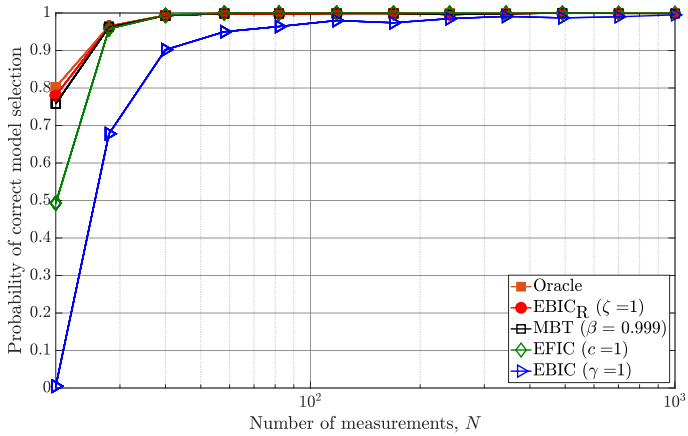
(b) $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$

Figure 5.5: PCMS versus N for SNR = 6 dB and $p = 1000$.

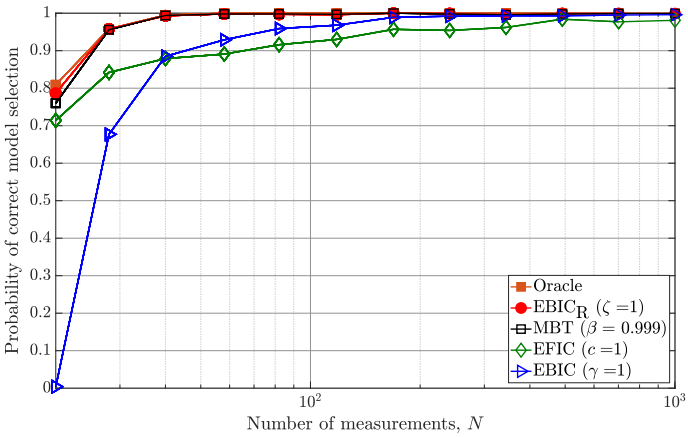
Model Selection versus N

In this section, we present the simulation results for model selection as a function of the sample size N. Similar to the previous section, we consider two different choices of $\mathbf{x}_{\mathcal{S}}$. Fig. 5.5 illustrates the empirical PCMS versus N for SNR = 6 dB, $p = 1000$ and $k_0 = 5$. It depicts a fairly low-SNR increasing N scenario. Fig. 5.5a and Fig. 5.5b corresponds to $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$ and $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$, respectively. Comparing Fig. 5.5a and Fig. 5.5b, we

notice that the performance of EFIC varies a lot compared to the other criteria. When $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$, EFIC slightly suffers from underfitting for low-SNR region due to high penalty value arising from the $-(k+2)\ln\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{y}\|_2^2$ term of the EFIC. On the contrary, when $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$ it is clearly seen that compared to the other criteria, EFIC suffers from the scaling issue and requires a large sample size to achieve PCMS = 1. Among all the criteria, the performance of EBIC and EBIC_{R} are closest to the oracle. Furthermore, observe that the performance of EBIC_{R} and EBIC are more or less alike for the current setting. This is primarily because the SNR is low (6 dB) hence the $(k+2)\ln(\hat{\sigma}_0^2/\hat{\sigma}_{\mathcal{I}}^2)$



(a) $\mathbf{x}_{\mathcal{S}} = [0.05, 0.04, 0.03, 0.02, 0.01]^T$



(b) $\mathbf{x}_{\mathcal{S}} = [50, 40, 30, 20, 10]^T$

Figure 5.6: PCMS versus N (20 to 10^3) for SNR = 25 dB, $p = N^d$ where $d = 1.3$.

term of EBIC_R behaves very close to a $\mathcal{O}(1)$ quantity for $k \geq k_0$. Thus, for low SNR scenarios, the penalties of EBIC and EBIC_R are similar, and as such the behaviour of these two criteria overlaps in this case. However, note that this is not true in the high-SNR cases, which will be evident from the discussion following Fig. 5.6.

The plots shown in Fig. 5.4 and Fig. 5.5 represent fixed-N increasing-SNR and low-SNR increasing-N scenarios, respectively. In Fig. 5.6, we present a high-SNR increasing-N case where SNR = 25 dB. Here, we consider a varying parameter space such that $p = N^d$ where $d = 1.3$. It is clearly observed that for high-SNR scenarios, EBIC_R and MBT provide much faster convergence to oracle behaviour as compared to EBIC which requires a larger sample size to achieve detection probability one. Furthermore, we also notice that EFIC suffers from a higher false selection error and performs worse than EBIC in a certain region of the sample size. This clearly shows the effects of scaling on the behaviour of EFIC.

5.6.5 Remarks from the Simulation Results

Key points from the simulation results are as follows:

- Classical methods struggle to handle large-p small-N cases. Convergence to true selection probability one requires more measurement samples when p is large but fixed and they fail miserably when p grows with N.
- Even for $N > p$ case but p sufficiently close to N, the extended versions offer faster convergence to oracle property compared to the classical methods.
- EBIC can handle large-p small-N cases. It is a consistent estimator of the true model as $N \to \infty$ even if p grows with N. However, the consistency property does not hold when N is fixed (and small) and $\text{SNR} \to \infty$.
- EFIC is a consistent criterion in the high-dimensional regime for both cases when $N \to \infty$ and/or $\text{SNR} \to \infty$. However, it is not invariant to data-scaling and its performance is unstable under changing signal and noise statistics.
- EBIC_R solves the data-scaling problem in EFIC. In addition, EBIC_R is a consistent criterion for both large-N and high-SNR cases and offers stable performance in changing noise and signal statistics.

5.7 Summary

In this chapter, we provided a new criterion, which is an extension of BIC_R, to handle model selection in sparse high-dimensional linear regression models employing greedy methods for predictor selection. The extended version is named EBIC_R, where the subscript 'R' stands for robust and it is a scale-invariant and consistent model selection criterion. Additionally, we analytically examined the behaviour of EBIC_R as $\sigma^2 \to 0$ and as $N \to \infty$. In both cases, it is shown that the probability of

detecting the true model approaches one. We further highlighted the data-scaling issue present in EFIC, which is a consistent criterion for both large sample sizes and high-SNR scenarios. Extensive simulation results show that the performance of EBIC_{R} is either similar to or superior to that of EBIC, EFIC, and MBT.

Appendix

5.A Lemmas

Lemma 5.1 *Let* $\mathbf{y}$ *be a* $N \times 1$ *dimensional vector following* $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_N)$ *and* $\boldsymbol{\Pi}$ *be a* $N \times N$ *symmetric, idempotent matrix with rank*$(\boldsymbol{\Pi}) = r$*. Then the ratio* $\mathbf{y}^T \boldsymbol{\Pi} \mathbf{y}/\sigma^2$ *has a non-central chi-square distribution* $\chi_r^2(\lambda)$ *with* r *degrees of freedom and non-centrality parameter* $\lambda = \boldsymbol{\mu}^T \boldsymbol{\Pi} \boldsymbol{\mu}/\sigma^2$ *(see, e.g., Chapter 5 of [74]).*

Lemma 5.2 *Let* $Z_{max} = \max_i \{Z_i\}_{i=1}^m$ *where* $Z_1, Z_2, \ldots, Z_m$ *is a sequence of identically distributed random variables (not necessarily independent) having a Chi-square distribution with* k *degrees of freedom where* $k < m$*. Then* $Z_{max} \leq k + 2\sqrt{k\psi \ln m} + 2\psi \ln m$ *for some constant* $\psi > 1$ *with probability approaching one as* $m \to \infty$*.*

Proof: From the union bound we have

$$\Pr\left(Z_{\max} \leq \eta\right) \geq 1 - m \Pr\left(Z_i \geq \eta\right). \tag{5.92}$$

Since $Z_i \sim \chi_k^2$, then from the Chi-square tail bound (Lemma 1 of [80]) we have the following result

$$\Pr\left(Z_i \geq k + 2\sqrt{kt} + 2t\right) \leq e^{-t}. \tag{5.93}$$

Setting $t = \psi \ln m$ in (5.93) where $\psi > 1$ we get

$$\Pr\left(Z_i \geq k + 2\sqrt{k\psi \ln m} + 2\psi \ln m\right) \leq e^{-\psi \ln m} = m^{-\psi}. \tag{5.94}$$

Using (5.94) in (5.92) we get

$$\Pr\left(Z_{\max} \leq k + 2\sqrt{k\psi \ln m} + 2\psi \ln m\right) \geq 1 - \frac{1}{m^{\psi - 1}}. \tag{5.95}$$

Therefore, $Z_{\max} \leq k + 2\sqrt{k\psi \ln m} + 2\psi \ln m$ with probability approaching one as $m \to \infty$ if $\psi > 1$.

Lemma 5.3 *Let* $X_{max} = \max_i \{X_i\}_{i=1}^m$ *where* $X_1, X_2, \ldots, X_m$ *is a sequence of identically distributed random variables (not necessarily independent) having a Gaussian distribution with zero mean and variance one. Then* $X_{max} \leq \sqrt{2 \ln m}$ *with probability approaching one as* $m \to \infty$*.*

Proof: From the union bound we have

$$\Pr\left(X_{\max} \leq \eta\right) \geq 1 - m \Pr\left(X_i \geq \eta\right). \tag{5.96}$$

Since $X_i \sim \mathcal{N}(0, 1)$, from the Gaussian tail bound we have

$$\Pr\left(X_i \geq \eta\right) \leq \frac{1}{\eta} \frac{e^{-\eta^2/2}}{\sqrt{2\pi}}, \tag{5.97}$$

for all $\eta > 0$. Setting $\eta = \sqrt{2\ln m}$ in (5.97) we get

$$\Pr\left(X_i \geq \sqrt{2\ln m}\right) \leq \frac{m^{-1}}{2\sqrt{\pi \ln m}}. \tag{5.98}$$

Using (5.98) in (5.96) we get

$$\Pr\left(X_{\max} \leq \sqrt{2\ln m}\right) \geq 1 - \frac{1}{2\sqrt{\pi \ln m}}. \tag{5.99}$$

Therefore, $X_{\max} \leq \sqrt{2\ln m}$ with probability approaching one as $m \to \infty$.

Lemma 5.4 *For any arbitrary support $\mathcal{I} \in \mathcal{I}_m^k \in \mathbb{M}$, under the asymptotic identifiability condition in (5.42) the following inequality holds*

$$\left\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\right\|_2^2 > 0.$$

Proof: Let $\mathcal{S}' = \{\mathcal{S} \setminus \mathcal{I}\}$. The true support $\mathcal{S}$ can be split into two disjoint subsets as $\mathcal{S} = \{\mathcal{S} \cap \mathcal{I}\} \cup \{\mathcal{S} \setminus \mathcal{I}\}$. Since $\text{span}(\mathbf{A}_{\mathcal{S}\cap\mathcal{I}}) \subset \text{span}(\mathbf{A}_{\mathcal{I}})$ we have

$$\begin{aligned}\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2 =& \|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}\mathbf{x}_{\mathcal{S}'}\|_2^2 \\ =& N\mathbf{x}_{\mathcal{S}'}^T\left(N^{-1}\mathbf{A}_{\mathcal{S}'}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}\right)\mathbf{x}_{\mathcal{S}'}.\end{aligned}$$

Now, consider the matrix $\mathbf{M} = \begin{bmatrix}\mathbf{A}_{\mathcal{S}'} & \mathbf{A}_{\mathcal{I}}\end{bmatrix}$ where $card(\mathcal{S}') \leq K$ and $card(\mathcal{I}) \leq K$, such that $card(\mathcal{S}' \cup \mathcal{I}) \leq 2K$. Under the assumption (5.42)

$$N^{-1}\mathbf{M}^T\mathbf{M} = N^{-1}\begin{bmatrix}\mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{S}'} & \mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{I}} \\ \mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{S}'} & \mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}\end{bmatrix} \tag{5.100}$$

is a bounded positive definite matrix. Then the Schur complement of the block matrix $\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}}$ is

$$\begin{aligned}& N^{-1}\left[\mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{S}'} - \mathbf{A}_{\mathcal{S}'}^T\mathbf{A}_{\mathcal{I}}(\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{I}})^{-1}\mathbf{A}_{\mathcal{I}}^T\mathbf{A}_{\mathcal{S}'}\right] \\ =& N^{-1}\mathbf{A}_{\mathcal{S}'}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}\end{aligned}$$

is also positive definite and bounded as $N \to \infty$. Let $\widetilde{\mathbf{M}} = N^{-1}\mathbf{A}_{\mathcal{S}'}^T\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}'}$, then, $\mathbf{x}_{\mathcal{S}'}^T\widetilde{\mathbf{M}}\mathbf{x}_{\mathcal{S}'} = b$ (say) $= \mathcal{O}(1) > 0$. Hence, $\|\mathbf{\Pi}_{\mathcal{I}}^{\perp}\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2 = Nb > 0$ for all $\mathcal{I} \in \mathcal{I}_m^k \in \mathbb{M}$.

5.B Statistical Analysis of $\hat{\sigma}_0^2$

From the generating model (5.1), the true data vector follows $\mathbf{y} \sim \mathcal{N}\left(\mathbf{A}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}, \sigma^2\mathbf{I}_N\right)$. Consider the factor $\hat{\sigma}_0^2$, which is defined as

$$\hat{\sigma}_0^2 = \frac{\|\mathbf{y}\|_2^2}{N} = \left(\frac{\sigma^2}{N}\right)\frac{\mathbf{y}^T\mathbf{I}_N\mathbf{y}}{\sigma^2}. \tag{5.101}$$

From Lemma 5.1 we have

$$\frac{\mathbf{y}^T \mathbf{I}_N \mathbf{y}}{\sigma^2} \sim \chi^2_N(\lambda) \text{ where } \lambda = \frac{\|\mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\|_2^2}{\sigma^2}. \tag{5.102}$$

This implies that $\left(\frac{N}{\sigma^2}\right) \hat{\sigma}^2_0 \sim \chi^2_N(\lambda)$. Therefore, the mean and variance of $\hat{\sigma}^2_0$ are:

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2_0] &= \frac{\sigma^2}{N}(N+\lambda) = \sigma^2 + \frac{\|\mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\|_2^2}{N} \\ \text{Var}[\hat{\sigma}^2_0] &= 2\frac{\sigma^4}{N^2}(N+2\lambda) = 2\frac{\sigma^4}{N} + 4\frac{\sigma^2}{N^2}\|\mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\|_2^2. \end{aligned} \tag{5.103}$$

Hence, for a fixed N,

$$\lim_{\sigma^2 \to 0} \mathbb{E}[\hat{\sigma}^2_0] = \frac{\|\mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}\|_2^2}{N} \quad \& \quad \lim_{\sigma^2 \to 0} \text{Var}[\hat{\sigma}^2_0] = 0. \tag{5.104}$$

Further, when SNR or σ^2 is fixed, using the assumption $\lim_{N\to\infty} \left\{ \frac{\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}}}{N} \right\} = \mathbf{M}_{\mathcal{S}}$ we get

$$\lim_{N \to \infty} \mathbb{E}[\hat{\sigma}^2_0] = \sigma^2 + \mathbf{x}_{\mathcal{S}}^T \mathbf{M}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}} \quad \& \quad \lim_{N \to \infty} \text{Var}[\hat{\sigma}^2_0] = 0, \tag{5.105}$$

where $\mathbf{M}_{\mathcal{S}}$ is a bounded positive definite matrix and as such $\mathbf{x}_{\mathcal{S}}^T \mathbf{M}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}} = \mathcal{O}(1)$ as N grows large.

5.C Statistical Analysis of $\hat{\sigma}^2_{\mathcal{I}}$ when $\mathcal{S} \subseteq \mathcal{I}$

The noise variance estimate under hypothesis $\mathcal{H}_{\mathcal{I}}$ can be rewritten as

$$\hat{\sigma}^2_{\mathcal{I}} = \left(\frac{\sigma^2}{N}\right) \frac{\mathbf{y}^T \mathbf{\Pi}^{\perp}_{\mathcal{I}} \mathbf{y}}{\sigma^2}. \tag{5.106}$$

The true model $\mathbf{u} = \mathbf{A}_{\mathcal{S}} \mathbf{x}_{\mathcal{S}}$ lies in a linear subspace spanned by the columns of $\mathbf{A}_{\mathcal{S}}$. Consequently, for $\mathcal{I} \supseteq \mathcal{S}$ we have $\mathbf{\Pi}^{\perp}_{\mathcal{I}} \mathbf{u} = \mathbf{0}$. This implies that $\mathbf{y}^T \mathbf{\Pi}^{\perp}_{\mathcal{I}} \mathbf{y} = \mathbf{e}^T \mathbf{\Pi}^{\perp}_{\mathcal{I}} \mathbf{e}$. Thus we have,

$$\frac{\mathbf{y}^T \mathbf{\Pi}^{\perp}_{\mathcal{I}} \mathbf{y}}{\sigma^2} = \frac{\mathbf{e}^T \mathbf{\Pi}^{\perp}_{\mathcal{I}} \mathbf{e}}{\sigma^2} \sim \chi^2_{N-k} \text{ (Using Lemma 5.1)}, \tag{5.107}$$

where $k = card(\mathcal{I}) \geq k_0$. This implies that $\left(\frac{N}{\sigma^2}\right) \hat{\sigma}^2_{\mathcal{I}} \sim \chi^2_{N-k}$. Therefore, the mean and variance of $\hat{\sigma}^2_{\mathcal{I}}$ for $\mathcal{I} \supseteq \mathcal{S}$ are:

$$\mathbb{E}[\hat{\sigma}^2_{\mathcal{I}}] = \frac{\sigma^2}{N}(N-k) \quad \& \quad \text{Var}[\hat{\sigma}^2_{\mathcal{I}}] = 2\frac{\sigma^4}{N^2}(N-k). \tag{5.108}$$

Hence, when σ^2 is a constant,

$$\lim_{N \to \infty} \mathbb{E}[\hat{\sigma}^2_{\mathcal{I}}] = \sigma^2 \quad \& \quad \lim_{N \to \infty} \text{Var}[\hat{\sigma}^2_{\mathcal{I}}] = 0. \tag{5.109}$$

Chapter 6

Model Selection in Block-Sparse Linear Regression

"What we know is a drop, what we don't know is an ocean."
—*Sir Issac Newton (1643–1729)*

In this chapter, we tackle the problem of model selection in a general linear regression model where the parameter matrix possesses a block-sparse structure, i.e, the non-zeros entries occur in clusters or blocks and the number of such non-zero blocks is very small compared to the parameter dimension. Furthermore, a high-dimensional setting is considered where the parameter dimension is quite large compared to the available measurements. To perform model selection in this setting, we present an information criterion that is motivated by the Extended Bayesian Information Criterion-Robust (EBIC_R) described in Chapter 5, and it takes into account both the block structure and the high-dimensional scenario. The analytical steps for deriving the generalized version of the EBIC_R for this setting are provided. Simulation results show that the proposed method performs considerably better than the existing state-of-the-art methods. It is also an empirically consistent criterion.

Block-sparsity naturally occurs in a variety of situations, such as in multi-band signals [81–83] or in measurements of the gene expression levels [84]. The multiple measurement vector (MMV) issue, which involves the measurement of a group of vectors that share a common sparsity pattern, is another intriguing specific example of the block-sparse model [85–89]. Furthermore, it was shown in [90] and [85] that the block-sparsity model can be used to treat the problem of sampling signals that lie in a union of subspaces [82, 91–95]. However, the literature on model selection in block-sparse linear regression is very scarce. The latest method for model selection in block-sparse high-dimensional linear regression is the generalized residual ratio thresholding (GRRT) [96]. This method is an extension of RRT [65] developed to handle the block-sparse structure. The authors also present a new approach that

allows GRRT to perform model selection in non-monotonic predictor sequences generated by LASSO. To the best of our knowledge, to date, no known method based on information criterion is available in the open literature to perform model selection in block-sparse linear regression. Therefore, the main goal herein is to develop an information theoretic based model selection method for the general linear regression model assuming a block-sparse structure and high-dimensional setting.

6.1 Problem Statement

Technically there can be four different linear regression models depending on the structure of the parameter matrix/vector. They are: (a) single measurement vector (SMV), (b) block single measurement vector (BSMV), (c) multiple measurement vector (MMV), and (d) block multiple measurement vector (BMMV). For example, as mentioned in [96], SMV models are used in wireless signal detection [97] , MMV models are used in Electroencephalogram (EEG) [98], BSMV models are used in multi-pitch estimation [99] and BMMV models are used in face recognition [100]. In this work, we consider the BMMV model, since it is the general setting and the rest of the models are special cases of BMMV. The BMMV model is given as follows

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}. \tag{6.1}$$

Here, $\mathbf{Y} \in \mathbb{R}^{N\times L}$ is the observed response matrix, $\mathbf{A} \in \mathbb{R}^{N\times p}$ is the design matrix, where $N \ll p$. $\mathbf{X} \in \mathbb{R}^{p\times L}$ is the unknown parameter matrix and $\mathbf{W} \in \mathbb{R}^{N\times L}$ is the noise/error matrix, whose elements are assumed to be i.i.d following $\mathbf{W}[i,j] \sim \mathcal{N}(0,\sigma^2)$ where σ^2 is the true noise variance. The p rows of $\mathbf{X}$ are divided into

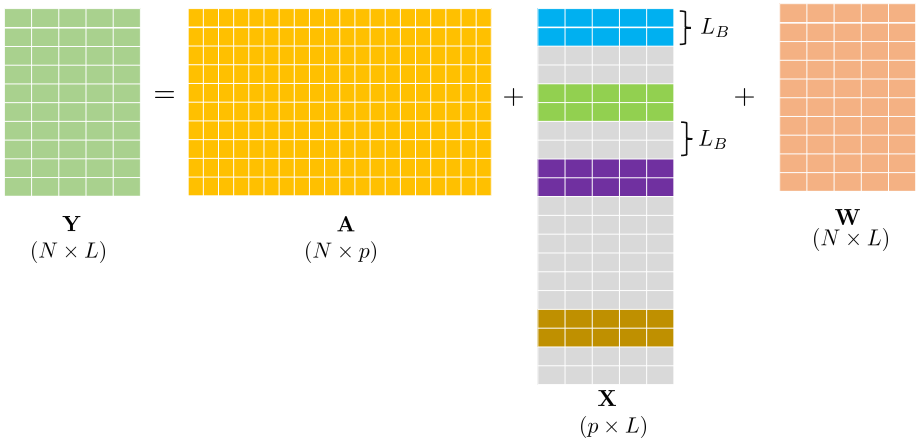


Figure 6.1: BMMV model scenario.

Type	Specifications	$dim(\mathbf{Y})$, $dim(\mathbf{X})$
SMV	$L = 1$, $L_B = 1$, $p_B = p$	$N \times 1$, $p \times 1$
MMV	$L > 1$, $L_B = 1$, $p_B = p$	$N \times L$, $p \times L$
BSMV	$L = 1$, $L_B > 1$, $p_B = p/L_B$	$N \times 1$, $p \times 1$
BMMV	$L > 1$, $L_B > 1$, $p_B = p/L_B$	$N \times L$, $p \times L$

Table 6.1: Type of linear regression models

$p_B = p/L_B$ non-overlapping blocks of equal size L_B. Each of these p_B blocks of size $L_B \times L$ are either completely zero or non-zero. The block size L_B is assumed to be known *a priori*. The jth block consists of the rows of $\mathbf{X}$ indexed by $\mathcal{I}_j = \{(j-1)L_B + 1, (j-1)L_B + 2, \ldots, jL_B\}$. We denote the true block support of $\mathbf{X}$ as $\mathcal{S}_B = \{j \in [p_B] : \mathbf{X}[\mathcal{I}_j, :] \neq \mathbf{0}_{L_B L}\}$. The parameter matrix $\mathbf{X}$ is assumed to be sparse such that $K_B = card(\mathcal{S}_B) \ll p_B$. Fig. 6.1 shows the BMMV scenario. The non-zero blocks in the parameter matrix $\mathbf{X}$ are shown in colour, while the blocks with zero entries are in gray. For the model in Fig. 6.1, it is clear that the true block sparsity is $K_B = 4$. BMMV is the general scenario, and all other scenarios are special cases of BMMV. Table 6.1 shows the configurations for the different types of the linear regression system.

The goal of model selection in the block linear regression model is estimating the true block support $\mathcal{S}_B$ given $\mathbf{Y}$ and $\mathbf{A}$. Here we categorize the model selection process into two major steps: (i) Subset selection, where a competent set of candidate models out of all the $(2^{p_B} - 1)$ possible models is obtained. In our work, we consider the set of competing models as the collection of all plausible combinatorial models up to a maximum cardinality K, under the assumption that $K_B \leq K \ll N$; (ii) estimating the true model among the candidate models using a suitable model selection criterion.

Consider a candidate model with block support $\mathcal{I}_B$ having block cardinality $card(\mathcal{I}_B) = k_B$, where $k_B \in \{1, 2, \ldots, p_B\}$. In this case, the linear model in (6.1) can be reformulated as follows

$$\mathcal{H}_{\mathcal{I}_B} : \mathbf{Y} = \mathbf{A}_{\mathcal{I}_B}\mathbf{X}_{\mathcal{I}_B} + \mathbf{W}_{\mathcal{I}_B}, \tag{6.2}$$

where $\mathcal{H}_{\mathcal{I}_B}$ denotes the hypothesis that the data $\mathbf{Y}$ is truly generated according to (6.2), $\mathbf{A}_{\mathcal{I}_B} \in \mathbb{R}^{N \times (k_B L_B)}$ is the sub-design matrix consisting of columns from the known design matrix $\mathbf{A}$ with block support $\mathcal{I}_B \subseteq \{1, 2, \ldots, p_B\}$, $\mathbf{X}_{\mathcal{I}_B} \in \mathbb{R}^{(k_B L_B) \times L}$ is the corresponding unknown regression coefficient matrix and $\mathbf{W}_{\mathcal{I}_B} \in \mathbb{R}^{N \times L}$ is the associated noise matrix

6.2 Proposed Method

In this section, we provide the necessary steps to derive EBIC_R to perform model selection for block-sparse linear regression models. The further analysis assumes

the following property of the design matrix $\mathbf{A}$ [16, 71, 79]

$$\lim_{N\to\infty}\left\{N^{-1}(\mathbf{A}_{\mathcal{I}_B}^T\mathbf{A}_{\mathcal{I}_B})\right\} = \mathbf{M}_{\mathcal{I}_B} = \mathcal{O}(1), \tag{6.3}$$

where $\mathbf{M}_{\mathcal{I}_B}$ is a $(k_BL_B \times k_BL_B)$ positive definite matrix and bounded as $N \to \infty$. The assumption in (6.3) is true in many applications but not all (see [72] for more details).

To arrive at EBIC_R for the block-sparse linear model, the first step is reformulating the linear model in (6.2) into a vector form as shown below:

$$\text{vec}(\mathbf{Y}) = \mathbf{I}_L \otimes \mathbf{A}_{\mathcal{I}_B}\text{vec}(\mathbf{X}_{\mathcal{I}_B}) + \text{vec}(\mathbf{W}_{\mathcal{I}_B}). \tag{6.4}$$

This allows us to utilize the same derivation steps as shown in Chapter 5 without the necessity to facilitate the analysis from scratch. Also, (6.4) is technically equivalent to (6.2), hence we do not alter the underlying original linear model but just restructure it for our convenience. Now let $\mathbf{y} = \text{vec}(\mathbf{Y}) \in \mathbb{R}^{NL\times 1}$, $\breve{\mathbf{A}}_\mathcal{I} = \mathbf{I}_L \otimes \mathbf{A}_{\mathcal{I}_B} \in \mathbb{R}^{NL\times k_BL_BL}$, $\mathbf{x}_\mathcal{I} = \text{vec}(\mathbf{X}) \in \mathbb{R}^{k_BL_BL\times 1}$ and $\mathbf{e} = \text{vec}(\mathbf{W}) \in \mathbb{R}^{NL\times 1}$. The elements of $\mathbf{e}_\mathcal{I}$ are i.i.d. and $\mathbf{e}_\mathcal{I} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathcal{I}^2\mathbf{I}_{NL})$. Then, we can rewrite (6.4) as

$$\mathcal{H}_\mathcal{I} : \mathbf{y} = \breve{\mathbf{A}}_\mathcal{I}\mathbf{x}_\mathcal{I} + \mathbf{e}_\mathcal{I}, \tag{6.5}$$

where $\mathcal{I} \subseteq \{1, 2, \ldots, pL\}$. Then the pdf of $\mathbf{y}$ under hypothesis $\mathcal{H}_\mathcal{I}$ is

$$p(\mathbf{y}|\boldsymbol{\theta}_\mathcal{I}, \mathcal{H}_\mathcal{I}) = \frac{\exp\{-\|\mathbf{y} - \breve{\mathbf{A}}_\mathcal{I}\mathbf{x}_\mathcal{I}\|_2^2/2\sigma_\mathcal{I}^2\}}{(2\pi\sigma_\mathcal{I}^2)^{NL/2}}, \tag{6.6}$$

where $\boldsymbol{\theta}_\mathcal{I} = [\mathbf{x}_\mathcal{I}^T, \sigma_\mathcal{I}^2]^T$ comprises of all the parameters of the model. Under $\mathcal{H}_\mathcal{I}$, the maximum likelihood estimates (MLEs) of $\hat{\boldsymbol{\theta}}_\mathcal{I} = [\hat{\mathbf{x}}_\mathcal{I}^T, \hat{\sigma}_\mathcal{I}^2]^T$ are obtained as [63]

$$\hat{\mathbf{x}}_\mathcal{I} = \left(\breve{\mathbf{A}}_\mathcal{I}^T\breve{\mathbf{A}}_\mathcal{I}\right)^{-1}\breve{\mathbf{A}}_\mathcal{I}^T\mathbf{y} \quad \& \quad \hat{\sigma}_\mathcal{I}^2 = \frac{\mathbf{y}^T\breve{\boldsymbol{\Pi}}_\mathcal{I}^\perp\mathbf{y}}{NL}, \tag{6.7}$$

where $\breve{\boldsymbol{\Pi}}_\mathcal{I}^\perp$ is the orthogonal projection matrix on the null space of $\breve{\mathbf{A}}_\mathcal{I}^T$. EBIC_R is derived under the Bayesian framework of model selection, which starts with deriving the maximum a-posteriori (MAP) criterion and ending with the final EBIC_R after suitable modifications and reasonable assumptions. We follow similar steps as given in Chapter 5, but incorporate the multiple-measurement and block structure into it. Let us denote the prior pdf of the parameter vector $\boldsymbol{\theta}_\mathcal{I}$ as $p(\boldsymbol{\theta}_\mathcal{I}|\mathcal{H}_\mathcal{I})$, the marginal of $\mathbf{y}$ as $p(\mathbf{y}|\mathcal{H}_\mathcal{I})$ and the prior probability of the model with support $\mathcal{I}$ as $\Pr(\mathcal{H}_\mathcal{I})$. Then the MAP estimate of the true support $\mathcal{S} \subseteq \{1, 2, \ldots pL\}$ is equivalently given by [16, 79, 101]

$$\hat{\mathcal{S}}_\text{MAP} = \arg\max_\mathcal{I}\left\{\ln p(\mathbf{y}|\mathcal{H}_\mathcal{I}) + \ln\Pr\left(\mathcal{H}_\mathcal{I}\right)\right\}. \tag{6.8}$$

Applying a second order Taylor series expansion, an approximation of $\ln p(\mathbf{y}|\mathcal{H}_\mathcal{I})$ is obtained under the presumption that N is large or/and SNR is high (see [16, 79] for details)

$$\ln p(\mathbf{y}|\mathcal{H}_\mathcal{I}) \approx \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\mathcal{I}, \mathcal{H}_\mathcal{I}) + \ln p(\hat{\boldsymbol{\theta}}_\mathcal{I}|\mathcal{H}_\mathcal{I}) + \frac{k_B L_B L + 1}{2} \ln(2\pi) - \frac{1}{2} \ln \left|\hat{\mathbf{F}}_\mathcal{I}\right|. \tag{6.9}$$

Here, $\hat{\mathbf{F}}_\mathcal{I}$ is the sample Fisher information matrix under $\mathcal{H}_\mathcal{I}$ given as [63]

$$\hat{\mathbf{F}}_\mathcal{I} = -\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}_\mathcal{I}, \mathcal{H}_\mathcal{I})}{\partial \boldsymbol{\theta}_\mathcal{I} \partial \boldsymbol{\theta}_\mathcal{I}^T}\bigg|_{\boldsymbol{\theta}_\mathcal{I}=\hat{\boldsymbol{\theta}}_\mathcal{I}}. \tag{6.10}$$

Evaluating (6.10) using (6.6) and (6.7) we get [16]

$$\hat{\mathbf{F}}_\mathcal{I} = \begin{bmatrix} \frac{1}{\hat{\sigma}_\mathcal{I}^2} \breve{\mathbf{A}}_\mathcal{I}^T \breve{\mathbf{A}}_\mathcal{I} & \mathbf{0} \\ \mathbf{0} & \frac{NL}{2\hat{\sigma}_\mathcal{I}^4} \end{bmatrix}. \tag{6.11}$$

From the linear model in (6.5) we have

$$-2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_\mathcal{I}, \mathcal{H}_\mathcal{I}) = NL \ln \hat{\sigma}_\mathcal{I}^2 + \text{const.} \tag{6.12}$$

Therefore, using (6.12), we can rewrite (6.9) as

$$-2 \ln p(\mathbf{y}|\mathcal{H}_\mathcal{I}) \approx NL \ln \hat{\sigma}_\mathcal{I}^2 + \ln \left|\hat{\mathbf{F}}_\mathcal{I}\right| - 2 \ln p(\hat{\boldsymbol{\theta}}_\mathcal{I}|\mathcal{H}_\mathcal{I}) - k_B L_B L \ln 2\pi + \text{const.} \tag{6.13}$$

Furthermore, it is assumed that the prior term in (6.9), i.e., $\ln p(\hat{\boldsymbol{\theta}}_\mathcal{I}|\mathcal{H}_\mathcal{I})$ is flat and uninformative, and hence ignored from the analysis. Thus, dropping the constants and the terms independent of the block model dimension k_B, we can equivalently reformulate the MAP based model estimate as

$$\hat{\mathcal{S}}_{\text{MAP}} = \arg\min_{\mathcal{I}} \Big\{ NL \ln \hat{\sigma}_\mathcal{I}^2 + \ln \left|\hat{\mathbf{F}}_\mathcal{I}\right| - k_B L_B L \ln 2\pi - 2 \ln \Pr(\mathcal{H}_\mathcal{I}) \Big\}. \tag{6.14}$$

EBIC_R is derived from (6.14) with some further modifications and approximations. The two key terms that require further analysis are $\ln|\hat{\mathbf{F}}_\mathcal{I}|$ and the prior term $\Pr(\mathcal{H}_\mathcal{I})$. First, we perform normalization of $\hat{\mathbf{F}}_\mathcal{I}$ under both large-N and high-SNR assumption [16, 102]. For this we factorize the $\ln \left|\hat{\mathbf{F}}_\mathcal{I}\right|$term in a similar manner as performed in Chapter 4 and 5

$$\begin{aligned} \ln \left|\hat{\mathbf{F}}_\mathcal{I}\right| &= \ln \left[|\mathbf{Q}| \left| \mathbf{Q}^{-1/2} \hat{\mathbf{F}}_\mathcal{I} \mathbf{Q}^{-1/2} \right| \right] \\ &= \ln |\mathbf{Q}| + \ln \left| \mathbf{Q}^{-1/2} \hat{\mathbf{F}}_\mathcal{I} \mathbf{Q}^{-1/2} \right|. \end{aligned} \tag{6.15}$$

The goal here is to choose a suitable $\mathbf{Q}$ matrix that normalizes the sample FIM $\hat{\mathbf{F}}_\mathcal{I}$ such that the second term in (6.15) is $\mathcal{O}(1)$, i.e., in this case it should be bounded

as $N \to \infty$ and/or $\sigma^2 \to 0$. To accomplish this objective, and motivated from [101], we choose the following $\mathbf{Q}^{-1/2}$ matrix for this scenario

$$\mathbf{Q}^{-1/2} = \begin{bmatrix} \sqrt{\frac{L_B}{N}}\sqrt{\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_0^2}}\mathbf{I}_{k_B L_B L} & \mathbf{0} \\ \mathbf{0} & \sqrt{\frac{L_B}{N}}\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_0^2} \end{bmatrix}, \tag{6.16}$$

where $\hat{\sigma}_0^2 = \|\mathbf{y}\|_2^2/NL$. Also for the considered generating model (6.5), $\hat{\sigma}_0^2 \to$ const. as $N \to \infty$ and/or $\sigma^2 \to 0$ [71, 79]. Two important points to note here regarding the choice of the $\mathbf{Q}^{-1/2}$ matrix are: (i) The ratio $\left(\frac{\hat{\sigma}_{\mathcal{I}}^2}{\hat{\sigma}_0^2}\right)$ is introduced to normalize the $\hat{\mathbf{F}}_{\mathcal{I}}$ w.r.t. σ^2 where the factor $\hat{\sigma}_0^2$ is especially utilized to counteract the data-scaling problem (as discussed elaborately in Chapter 4 and 5). (ii) The $\frac{1}{N}$ portion of the factor $\frac{L_B}{N}$ is used to normalize the FIM w.r.t. N. However, L_B is also included as part of the normalizing term because for the mean-squared-error of $\hat{\sigma}^2$ to approach the Cramér-Rao bound, we require that the number of measurements is much larger than the number of parameters, i.e., $NL \gg K_B L_B L$ or in other words $N/L_B \gg K_B$. Hence, we use the normalization factor L_B/N instead of just $1/N$ in (6.16). In this way, the penalty will be a function of N/L_B instead of N alone (as will be seen in the subsequent steps). This novel modification helps to counteract the effects of changing L_B on the performance of EBIC_{R}.

Now, using (6.3), (6.11), and (6.16) we can show that

$$\begin{aligned} \left|\mathbf{Q}^{-1/2}\hat{\mathbf{F}}_{\mathcal{I}}\mathbf{Q}^{-1/2}\right| &= \left| \begin{matrix} \frac{L_B}{\hat{\sigma}_0^2}\frac{\check{\mathbf{A}}_{\mathcal{I}}^T\check{\mathbf{A}}_{\mathcal{I}}}{N} & \mathbf{0} \\ \mathbf{0} & \frac{L_B L}{2\hat{\sigma}_0^4} \end{matrix} \right| \\ &= \frac{L_B^{k_B L_B L+1} L}{2\left(\hat{\sigma}_0^2\right)^{k_B L_B L+2}} \left| \mathbf{I}_L \otimes \frac{\mathbf{A}_{\mathcal{I}_B}^T \mathbf{A}_{\mathcal{I}_B}}{N} \right| \\ &= \text{const.} \times \left|\mathbf{I}_L\right|^{k_B \times L_B} \left| \frac{\mathbf{A}_{\mathcal{I}_B}^T \mathbf{A}_{\mathcal{I}_B}}{N} \right|^L \\ &= \mathcal{O}(1) \end{aligned} \tag{6.17}$$

as N grows large and/or $\sigma^2 \to 0$. Hence, this term can be removed without significantly affecting the criterion. Next observe that the $\ln\left|\mathbf{Q}\right|$ term can be expanded as follows

$$\begin{aligned} \ln|\mathbf{Q}| &= \ln \left| \begin{matrix} \frac{N}{L_B}\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right)\mathbf{I}_{k_B L_B L} & \mathbf{0} \\ \mathbf{0} & \frac{N}{L_B}\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right)^2 \end{matrix} \right| \\ &= (k_B L_B L + 1)\ln\left(\frac{N}{L_B}\right) + (k_B L_B L + 2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right). \end{aligned} \tag{6.18}$$

Therefore, using (6.17) and (6.18) we can rewrite (6.15) as

$$\ln\left|\hat{\mathbf{F}}_{\mathcal{I}}\right| = k_B L_B L \ln\left(\frac{N}{L_B}\right) + (k_B L_B L + 2)\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_{\mathcal{I}}^2}\right) + \mathcal{O}(1) + \ln(N/L_B). \tag{6.19}$$

Next, for the model prior probability term $-2\ln\Pr(\mathcal{H}_{\mathcal{I}})$ in (6.14), a similar proposition is taken as in EBIC such that $\Pr(\mathcal{H}_{\mathcal{I}}) \propto \binom{p_B}{k_B}^{-\zeta}$, where $\zeta \geq 0$ is a tuning parameter. If p_B is sufficiently large, the following approximation can be assumed $\ln\binom{p_B}{k_B} \approx k_B \ln p_B$ [18]. This gives

$$-2\ln\Pr(\mathcal{H}_{\mathcal{I}}) = 2\zeta k_B \ln p_B + \text{const.} \tag{6.20}$$

Now, substituting (6.19), (6.20) in (6.14) and dropping the $\mathcal{O}(1)$, the $\ln(N/L_B)$ term (independent of k_B), the constant and the $p(\hat{\boldsymbol{\theta}}_{\mathcal{I}}|\mathcal{H}_{\mathcal{I}})$ term we arrive at the generalized EBIC_R for the BMMV model:

$$\text{EBIC}_\text{R}(\mathcal{I}) = NL\ln\hat{\sigma}^2_{\mathcal{I}} + k_B L_B L \ln\left(\frac{N}{2\pi L_B}\right) + (k_B L_B L + 2)\ln\left(\frac{\hat{\sigma}^2_0}{\hat{\sigma}^2_{\mathcal{I}}}\right) + 2k_B\zeta\ln p_B. \tag{6.21}$$

In practice, we compute the EBIC_R score block-wise, i.e., $\text{EBIC}_\text{R}(\mathcal{I}_B)$ where $\mathcal{I}_B \subseteq \{1,\ldots,p_B\}$. Then the $\hat{\sigma}^2_{\mathcal{I}}$ can be replaced by $\hat{\sigma}^2_{\mathcal{I}_B} = \|\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_B}\mathbf{Y}\|^2_F/NL$, where $\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_B}$ is the orthogonal projection matrix on the null space of $\mathbf{A}^T_{\mathcal{I}_{\mathcal{B}}}$. Finally, the true block support is estimated as

$$\hat{\mathcal{S}}_B = \arg\min_{\mathcal{I}_B}\big\{\text{EBIC}_\text{R}(\mathcal{I}_B)\big\}. \tag{6.22}$$

The EBIC_R form given by 6.21 can be considered as the general form which is applicable for model selection in SVM, BSMV, MMV, and BMMV systems. Clearly, for $L = L_B = 1$ (SMV), 6.21 boils down to the original EBIC_R form already derived and discussed in Chapter 5.

6.3 Predictor Selection Algorithms for Block-Sparse models

In Chapter 3 and 5, we employed OMP for predictor selection in the SMV linear regression scenario and mentioned that LASSO can also be used for this purpose. In the case of OMP, it has been further extended to handle predictor selection in different scenarios. For example, simultaneous OMP (SOMP) [103, 104], block OMP (BOMP) [105] and BMMV-OMP in [106] are extensions of OMP in MMV, BSMV and BMMV scenarios. Similarly, group LASSO and MMV-LASSO are the BSMV and MMV versions of LASSO [107, 108]. The Algorithm in 6.1 presents the generic-OMP (G-OMP) steps for different scenarios, i.e., OMP for SMV, SOMP for MMV, BOMP for BSMV, and BMMV-OMP for BMMV. If we denote K as the iteration when the algorithm stops, then the output of G-OMP is the (block) index sequence $\mathcal{S}^K_{\text{G-OMP}}$.

6.4 Simulation Results

In this section, we provide numerical simulations to highlight the performance of EBIC_R for block-sparse linear regression models. We consider the BMMV model

Algorithm 6.1 Generic-OMP (G-OMP) framework

Inputs: Design matrix $\mathbf{A}$, measurement $\mathbf{Y}$.
Initialization: $\|\mathbf{a}_j\|_2 = 1\ \forall j$, $\mathbf{R}^0 = \mathbf{Y}$, $\mathcal{S}^0_{\text{G-OMP}} = \emptyset$ and counter $i = 0$
repeat
  Step 1: Identify the block that has the highest correlation with $\mathbf{R}^{i-1}$
    (OMP) Next predictor index $d^i = \underset{j=1,\ldots,p}{\arg\max} \left|\mathbf{A}[:, \mathcal{I}_j]^T \mathbf{R}^{i-1}\right|$
    (BOMP) Next block index $d^i = \underset{j=1,\ldots,p_B}{\arg\max} \left\|\mathbf{A}[:, \mathcal{I}_j]^T \mathbf{R}^{i-1}\right\|_2$
    (SOMP) Next predictor index $d^i = \underset{j=1,\ldots,p}{\arg\max} \left\|\mathbf{A}[:, \mathcal{I}_j]^T \mathbf{R}^{i-1}\right\|_2$
    (BMMV-OMP) Next block index $d^i = \underset{j=1,\ldots,p_B}{\arg\max} \left\|\mathbf{A}[:, \mathcal{I}_j]^T \mathbf{R}^{i-1}\right\|_F$
  Step 2: Add current index $\mathcal{S}^i_{\text{G-OMP}} = \mathcal{S}^{i-1}_{\text{G-OMP}} \cup \{d^i\}$
  Step 3: Update residual: $\mathbf{R}^i = \left(\mathbf{I}_N - \mathbf{\Pi}_{\mathcal{S}^i_{\text{G-OMP}}}\right) \mathbf{Y}$
  Step 4: increment counter $i \leftarrow i + 1$
until Stopping rule is satisfied
Output: OMP generated block index sequence $\mathcal{S}^K_{\text{G-OMP}}$ (K is the iteration when the algorithm stopped)

$\mathbf{Y} = \mathbf{AX} + \mathbf{W}$, where the design matrix $\mathbf{A}$ is generated with independent entries following normal distribution $\mathcal{N}(0, 1)$. The cardinality of the true block-support $\mathcal{S}_B$ is chosen to be $card(\mathcal{S}_B) = K_B = 4$. Also, without loss of generality, we assume $\mathcal{S}_B = [1, 2, 3, 4]$. Other parameters are chosen as $L = 5$, $L_B = 10$. The SNR in dB $= 10\log_{10}(\sigma_s^2/\sigma^2)$, where σ_s^2 and σ^2 denote signal and true noise power, respectively. The signal power is computed as $\sigma_s^2 = \|\mathbf{AX}\|_F^2/NL$. Based on σ_s^2 and the chosen SNR (dB), the noise power is set as $\sigma^2 = \sigma_s^2/10^{\text{SNR (dB)}/10}$. Using this σ^2, the elements of the noise matrix $\mathbf{W}$ are generated following $\mathbf{W}[i, j] \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. The probability of correct model selection (PCMS) is estimated over 1000 Monte Carlo trials. To maintain randomness in the data, a new design matrix $\mathbf{A}$ is generated at each Monte Carlo trial. BMMV-OMP (shown in Algorithm 6.1) is used for predictor selection for its simplicity and wider range of applicability. Performing model selection combining BMMV-OMP and EBIC_R is shown in Algorithm 6.2. Since, both EBIC and EFIC in their current forms are not designed for model selection in BMMV system, as such, we exclude them from the comparison. Hence, EBIC_R is compared with GRRT and the oracle, i.e, BMMV-OMP with *a priori* knowledge of K_B. Next, we provide the working principle of GRRT in detail for better clarity. GRRT is a method based on hypothesis testing that uses a test statistic that is a ratio of residuals calculated as follows:

$$RR(k_B) = \frac{\|\mathbf{R}^{k_B}\|_F}{\|\mathbf{R}^{k_B-1}\|_F} \tag{6.23}$$

Algorithm 6.2 Model selection combining EBIC_R with BMMV-OMP

1: Run BMMV-OMP for K iterations to obtain $\mathcal{S}^K_{\text{BMMV-OMP}}$
2: **for** $k_B = 1$ to K **do**
3: $\quad \mathcal{I}_B = \mathcal{S}^{k_B}_{\text{BMMV-OMP}}$
4: $\quad$ Compute $\text{EBIC}_\text{R}(\mathcal{I}_B)$
5: **end for**
6: Estimated true block support: $\hat{\mathcal{S}}_{\text{EBIC}_\text{R}} = \arg\min_{\mathcal{I}_B}\{\text{EBIC}_\text{R}(\mathcal{I}_B)\}$

where $\left\|\mathbf{R}^{k_B}\right\|_F = \left\|\mathbf{\Pi}^{\perp}_{\mathcal{S}^{k_B}_{\text{B-OMP}}}\mathbf{Y}\right\|_F$. The complete steps to perform model selection using GRRT combined with BMMV-OMP are shown in the Algorithm 6.3. Here, the quantity $\Gamma^{\alpha}_{GRRT}(k_B)$ is evaluated as follows

$$\Gamma^{\alpha}_{GRRT}(k_B) = \sqrt{\mathcal{B}^{-1}\left(\rho; \frac{(N - L_B k)L}{2}, \frac{L_B L}{2}\right)}, \tag{6.24}$$

where $\rho = \frac{\alpha}{pos(k)K}$ and $pos(k) = p_B - k_B + 1$. For the simulations, we choose $\alpha = 0.01$ as motivated by the original paper.

Fig. 6.2 shows the PCMS vs SNR (dB) with $N = 150$ and $p = 1000$. Since $L_B = 10$, hence, $p_B = p/L_B = 100$. The first clear observation is that for the considered tuning parameter setting, both EBIC_R and GRRT are empirically consistent in high-SNR, i.e., PCMS $\to 1$ as SNR increases (or inversely $\sigma^2 \to 0$). Second, compared to GRRT, the performance curve of EBIC_R is much closer to the oracle performance.

Fig. 6.3 presents the PCMS versus number of measurements N plot. Here, a fixed value of $p = 2000$ is chosen, hence $p_B = 200$. Also, a low value of SNR = -4 dB is chosen to make the detection more challenging. A similar trend is observed here as well. Both methods achieve empirical consistency (PCMS $\to 1$) as N grows large. However, EBIC_R provides slightly better performance compared to GRRT for smaller N values and is much closer to the oracle performance.

Next, we provide further simulation results with respect to changing variables L and L_B. Fig. 6.4 shows the PCMS versus SNR (dB) but for two different values of L_B, i.e., $L_B = 5$ and $L_B = 20$. It is clearly visible that with the increase in L_B

Algorithm 6.3 GRRT with BMMV-OMP

1: **Inputs:** Design matrix $\mathbf{A}$, observation vector $\mathbf{Y}$.
2: **Step 1** Run K iterations of BMMV-OMP
3: **Step 2** Compute $RR(k_B)$ for $k_B = 1, \ldots, K$
4: **Step 3** Compute $k_{GRRT} = \max\{k_B : RR(k_B) \leq \Gamma^{\alpha}_{GRRT}(k_B)\}$
5: **Outputs:** True support estimate $\hat{\mathcal{S}}_B = \mathcal{S}^{k_{GRRT}}_{\text{OMP}}$.

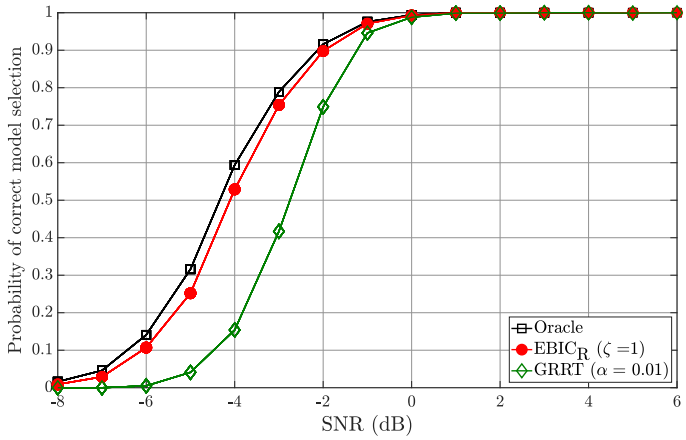


Figure 6.2: PCMS vs SNR (dB) for $N = 150$, $p = 1000$, $L = 5$, $L_B = 10$ and $K_B = 4$.

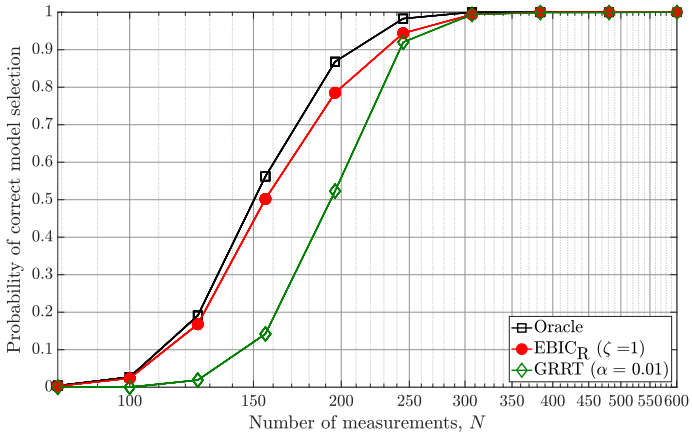


Figure 6.3: PCMS vs N for SNR = -4 dB, $p = 2000$, $L = 5$, $L_B = 10$ and $K_B = 4$.

value the performances of all the methods decline. Thus, an increase in the block length causes a decrease in the performance and vice-versa given that N and L are constants. The reason for this can be attributed to the fact that as L_B increases, the number of true parameters to be estimated also grows large. However, since the total number of available measurements is constant, this produces a higher estimation error. Furthermore, a higher value of L_B lowers the sparse nature of the parameter matrix causing enhanced error in the recovery process. This is reflected in the behaviour of the oracle, which also suffers at higher L_B values and has a slower convergence to PCMS=1.

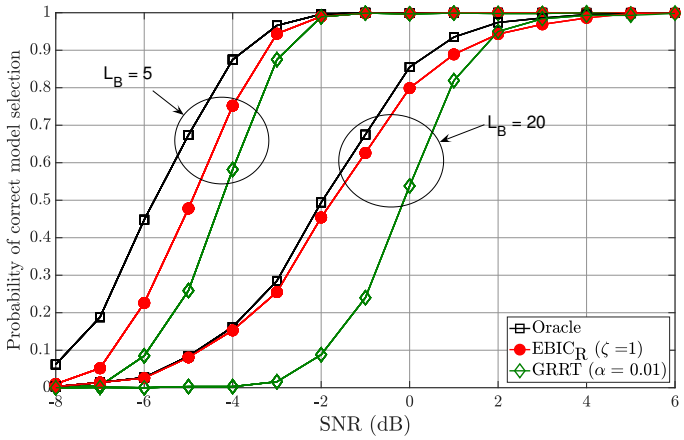


Figure 6.4: PCMS vs SNR for $N = 150$, $p = 1000$, $L = 5$, and $K_B = 4$.

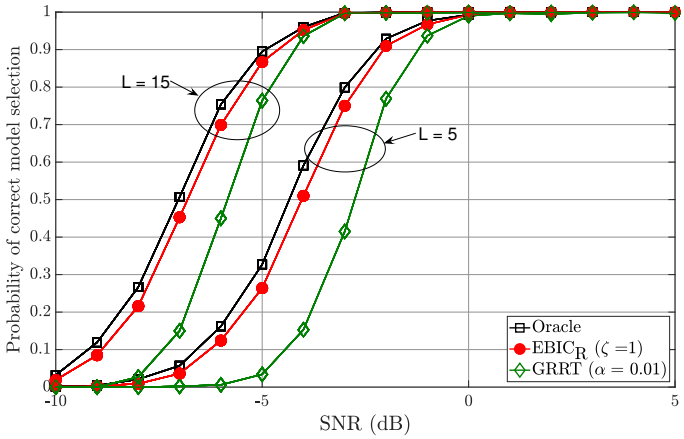


Figure 6.5: PCMS vs SNR for $N = 150$, $p = 1000$, $L_B = 10$, and $K_B = 4$.

Fig. 6.5 shows the PCMS vs SNR for two different values of L, i.e., $L = 5$ and $L = 15$. In this case, the behaviour is contrary to that of changing L_B, and the performances of all the methods improve as L grows large. This is precisely because, with each increase in L, the net measurement sample size grows as $N \times L$.

6.5 Summary

In this chapter, we have extended the EBIC_R to handle model selection in the block-sparse linear regression. A generalized method is developed that is applicable to all

forms of the linear regression structure such as SMV, BSMV, MMV, and BMMV. The steps to arrive at the criterion are shown in detail. Simulation results show that EBIC_{R} is an empirically consistent criterion as $N \to \infty$ and/or $\text{SNR} \to \infty$. Furthermore, we also underline the manner in which the parameters L and the block length L_B affect the model selection performance.

Chapter 7

Conclusion and Future Work

"Nothing in life is to be feared, it is only to be understood."
—Marie Curie (1867–1934)

ROBUST methods are desired in every field of science and technology. In this PhD thesis, we reviewed the classical model selection (MS) problem in linear regression focusing more on the high-dimensional (HD) setting. The primary motivation was to develop robust methods of MS that are consistent in large sample sizes and high-SNR scenarios, and are also invariant to any data-scaling. These properties are desired in any MS method.

We started the journey by developing a MS method based on the hypothesis testing framework called Multi-Beta-Test (MBT). MBT was proposed as a response to deal with the drawback of classical hypothesis testing to handle model selection in a high-dimensional setting employing greedy methods for predictor selection. MBT is specially tailored for this purpose. To perform MS it is combined with a predictor selection algorithm such as OMP that generates a monotonic sequence of the predictor indices in the order of decreasing significance. Using relative least-square cost between successive models as the test statistic, MBT keeps picking models step by step with increasing dimension and stops when the test fails. The order at which the test stops is the estimated true sparsity of the linear regression model. However, MBT has some drawbacks. In its current form, MBT cannot be applied to predictor or subset selection algorithms that generate non-monotonic indices such as LASSO. Hence, it is limited by its choice of predictor selection algorithms. Secondly, MBT is quite sensitive to the tuning parameter β that needs to be chosen beforehand to evaluate the threshold. For $\beta < 1$, MBT will not achieve consistency as $N \to \infty$ or $\sigma^2 \to 0$.

In the next phase of the thesis, the focus is diverted to the information criterion (IC) based approaches for MS. We re-investigate the popular Bayesian IC (BIC), particularly the high-SNR forms of the BIC. These high-SNR forms were proposed to make BIC consistent as $\sigma^2 \to 0$. However, it was discovered that the high-

SNR forms were not invariant to data-scaling and their behavior fluctuates under different scaling conditions. This is not a desirable property of any MS criterion. To eliminate the data-scaling problem, we proposed BIC-Robust or BIC_R. BIC_R completely alleviates the data-scaling problem. Also, it is a consistent estimator of the true model order as $N \to \infty$ or SNR$\to \infty$. This is further verified from the analytical proofs that guarantee the consistency of BIC_R.

A natural extension of BIC_R to deal with the HD situation led to the development of extended BIC_R (EBIC_R). EBIC_R can be seen as a successor to EBIC and EFIC. EBIC is inconsistent in the high-SNR scenarios when N is small. EFIC on the other hand is not invariant to data-scaling. EBIC_R alleviates both these issues which makes it a more robust MS method. We investigate the behaviour of EBIC_R as $N \to \infty$ and $\sigma^2 \to 0$ under the assumption that p grows with N as $p = N^d$ where $d > 1$ is some constant. In both cases, it was shown that EBIC_R can precisely pick the true model with the correct detection probability approaching one.

In the final phase of the thesis, the newly developed EBIC_R was generalized to perform MS in the general linear regression framework that may possess a block structure where the non-zero entries of the parameter matrix occur in blocks or groups. A block-sparse nature of the linear model is assumed under the HD setting. The fundamental derivation procedure of the EBIC_R for the block-sparse model is very similar to the original EBIC_R but it takes the block nature of the linear model into account. The generalized EBIC_R can be employed for MS in all the different forms of the linear regression scenarios, viz., SMV, MMV, BSMV, and BMMV. Thus making it a versatile MS criterion that is also consistent and scale-invariant.

7.1 Future Work

1. In the thesis, the proposed methods BIC_R and EBIC_R relies on the assumption that the design matrix obeys $N^{-1}(\mathbf{A}_\mathcal{I}^T\mathbf{A}_\mathcal{I}) \to \mathcal{O}(1)$ as $N \to \infty$. This is indeed quite generally true in many situations but does not hold valid in all cases. The fundamental issue is how to handle MS in that situation and if it is possible to generalize the proposed method such that it caters to all forms of the design matrix.

2. The entire thesis deals with only linear models which are widely used but also limiting to some extent. Most real-world data nowadays may require non-linear modeling. However, model selection in the non-linear setting is a much more challenging and difficult problem. As such, developing robust MS approaches for non-linear systems is a very interesting direction to look at.

3. Today machine learning methods are at the forefront of all important applications. Can we employ machine learning approaches to perform MS. What are the challenges and drawbacks in this regard? This can be a very intriguing topic of investigation.

References

[1] S. Konishi and G. Kitagawa, "Information criteria and statistical modeling," 2008.

[2] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 6, pp. 16–34, 2018.

[3] A. C. Davison, *Statistical models*, vol. 11. Cambridge university press, 2003.

[4] P. McCullagh, "What is a statistical model?," *The Annals of Statistics*, vol. 30, no. 5, pp. 1225–1310, 2002.

[5] J. Pfanzagl, *Parametric statistical theory*. Walter de Gruyter, 2011.

[6] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[7] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.

[8] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[9] G. W. Corder and D. I. Foreman, *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons, 2014.

[10] O. Z. Maimon and L. Rokach, *Data mining with decision trees: theory and applications*, vol. 81. World scientific, 2014.

[11] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[12] N. Cristianini, J. Shawe-Taylor, *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[13] P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov, *Efficient and adaptive estimation for semiparametric models*, vol. 4. Springer, 1993.

[14] D. Anderson and K. Burnham, "Model selection and multi-model inference," *Second. NY: Springer-Verlag*, vol. 63, no. 2020, p. 10, 2004.

[15] X. Su, X. Yan, and C.-L. Tsai, "Linear regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 3, pp. 275–294, 2012.

[16] P. Stoica and P. Babu, "On the proper forms of BIC for model order selection," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4956–4961, 2012.

[17] J. Chen and Z. Chen, "Extended bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[18] A. Owrang and M. Jansson, "A model selection criterion for high-dimensional linear regression," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3436–3446, 2018.

[19] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012.

[20] S. Weisberg, *Applied linear regression*, vol. 528. John Wiley & Sons, 2005.

[21] J. Marchini, P. Donnelly, and L. R. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nature genetics*, vol. 37, no. 4, pp. 413–417, 2005.

[22] M. Lustig, J. M. Santos, J.-H. Lee, D. L. Donoho, and J. M. Pauly, "Application of compressed sensing for rapid mr imaging," *SPARS,(Rennes, France)*, 2005.

[23] R. Baraniuk and P. Steeghs, "Compressive radar imaging," in *2007 IEEE radar conference*, pp. 128–133, IEEE, 2007.

[24] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[25] S. Chen and D. Donoho, "Basis pursuit," in *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 41–44, IEEE, 1994.

[26] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proceedings of 27th Asilomar conference on signals, systems and computers*, pp. 40–44, IEEE, 1993.

[27] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[29] T. Hastie, R. Tibshirani, and M. Wainwright, "Statistical learning with sparsity," *Monographs on statistics and applied probability*, vol. 143, p. 143, 2015.

[30] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[31] D. L. Donoho, X. Huo, *et al.*, "Uncertainty principles and ideal atomic decomposition," *IEEE transactions on information theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[32] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[33] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on information theory*, vol. 52, no. 1, pp. 6–18, 2005.

[34] Q. Mo, "A sharp restricted isometry constant bound of orthogonal matching pursuit," *arXiv preprint arXiv:1501.01708*, 2015.

[35] J. Wen, Z. Zhou, J. Wang, X. Tang, and Q. Mo, "A sharp condition for exact support recovery with orthogonal matching pursuit," *IEEE Transactions on Signal Processing*, vol. 65, no. 6, pp. 1370–1382, 2016.

[36] C. Liu, Y. Fang, and J. Liu, "Some new results about sufficient conditions for exact support recovery of sparse signals via orthogonal matching pursuit," *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4511–4524, 2017.

[37] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Transactions on Information theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[38] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[39] C.-H. Zhang and J. Huang, "The sparsity and bias of the lasso selection in high-dimensional linear regression," *The Annals of Statistics*, vol. 36, no. 4, pp. 1567–1594, 2008.

[40] P. Stoica, Y. Selén, and J. Li, "On information criteria and the generalized likelihood ratio test of model order selection," *IEEE Signal Processing Letters*, vol. 11, no. 10, pp. 794–797, 2004.

[41] R. Hocking, "The analysis and selection of variables in linear regression," *Biometrica*, vol. 32, pp. 1–49, 1976.

[42] F. E. Harrell *et al.*, *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, vol. 608. Springer, 2001.

[43] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, 2004.

[44] C. Rao, Y. Wu, S. Konishi, and R. Mukerjee, "On model selection," *Lecture Notes-Monograph Series*, pp. 1–64, 2001.

[45] A. Chakrabarti and J. K. Ghosh, "AIC, BIC and recent advances in model selection," *Philosophy of statistics*, pp. 583–605, 2011.

[46] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[47] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[48] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[49] M. H. Hansen and B. Yu, "Model selection and the principle of minimum description length," *Journal of the American Statistical Association*, vol. 96, no. 454, pp. 746–774, 2001.

[50] J. Rissanen, "MDL denoising," *IEEE Transactions on Information Theory*, vol. 46, no. 7, pp. 2537–2543, 2000.

[51] P. Stoica and P. Babu, "Model order estimation via penalizing adaptively the likelihood (PAL)," *Signal Processing*, vol. 93, no. 11, pp. 2865–2871, 2013.

[52] R. Kashyap, "Inconsistency of the AIC rule for estimating the order of autoregressive models," *IEEE Transactions on Automatic Control*, vol. 25, no. 5, pp. 996–998, 1980.

[53] J. Rissanen, "Estimation of structure by minimum description length," *Circuits, Systems and Signal Processing*, vol. 1, no. 3, pp. 395–406, 1982.

[54] H. Bozdogan, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[55] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *Journal of Econometrics*, vol. 187, no. 1, pp. 95–112, 2015.

[56] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation.," *Encyclopedia of database systems*, vol. 5, pp. 532–538, 2009.

[57] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[58] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and akaike's criterion," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977.

[59] L. de Torrenté and T. Hastie, "Does cross-validation work when p≫ n?," 2012.

[60] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[61] M. Chichignoud, J. Lederer, and M. J. Wainwright, "A practical scheme and fast algorithm to tune the lasso with optimality guarantees," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8162–8181, 2016.

[62] S. Foucart and H. Rauhut, "An invitation to compressive sensing," in *A mathematical introduction to compressive sensing*, pp. 1–39, Springer, 2013.

[63] S. Kay, *Fundamental of Statistical Signal Processing: Volume I Estimation Theory*. Prentice Hall, 1998.

[64] C. Walck, "Hand-book on statistical distributions for experimentalists," (No SUF-PFY/96-01), 1996.

[65] S. Kallummil and S. Kalyani, "Signal and noise statistics oblivious orthogonal matching pursuit," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, (Stockholmsmässan, Stockholm Sweden), pp. 2429–2438, PMLR, 10–15 Jul 2018.

[66] V. Koivunen and E. Ollila, "Model order selection," in *Academic Press Library in Signal Processing*, vol. 3, pp. 9–25, Elsevier, 2014.

[67] Q. Ding and S. Kay, "Inconsistency of the MDL: On the performance of model order selection criteria with increasing signal-to-noise ratio," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 1959–1969, 2011.

[68] S. Kay, "Exponentially embedded families-new approaches to model order estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 333–345, 2005.

[69] K. W. Broman and T. P. Speed, "A model selection approach for the identification of quantitative trait loci in experimental crosses," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 641–656, 2002.

[70] P. B. Gohain and M. Jansson, "Relative cost based model selection for sparse high-dimensional linear regression models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5515–5519, IEEE, 2020.

[71] D. F. Schmidt and E. Makalic, "The consistency of MDL for linear regression models with increasing signal-to-noise ratio," *IEEE transactions on signal processing*, vol. 60, no. 3, pp. 1508–1510, 2011.

[72] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2726–2735, 1998.

[73] S. Kallummil and S. Kalyani, "High SNR consistent linear model order selection and subset selection," *IEEE Transactions on Signal Processing*, vol. 64, no. 16, pp. 4307–4322, 2016.

[74] A. M. Mathai and S. B. Provost, *Quadratic forms in random variables: theory and applications*. Dekker, 1992.

[75] H. Yanai, K. Takeuchi, and Y. Takane, "Projection matrices," in *Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition*, pp. 25–54, Springer, 2011.

[76] N. Ravishanker and D. K. Dey, *A first course in linear model theory*. CRC Press, 2020.

[77] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The annals of statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[78] R. R. Picard and R. D. Cook, "Cross-validation of regression models," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 1984.

[79] P. B. Gohain and M. Jansson, "Scale-invariant and consistent Bayesian information criterion for order selection in linear regression models," *Signal Processing*, p. 108499, 2022.

[80] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, pp. 1302–1338, 2000.

[81] H. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Mathematica*, vol. 117, pp. 37–52, 1967.

[82] M. Mishali and Y. C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Transactions on signal processing*, vol. 57, no. 3, pp. 993–1009, 2009.

[83] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," *IEEE Journal of selected topics in signal processing*, vol. 4, no. 2, pp. 375–391, 2010.

[84] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed dna microarrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 275–285, 2008.

[85] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[86] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2009.

[87] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[88] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, 2005.

[89] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[90] Y. C. Eldar and M. Mishali, "Block sparsity and sampling over a union of subspaces," in *2009 16th International Conference on Digital Signal Processing*, pp. 1–8, IEEE, 2009.

[91] P. G. Casazza and G. Kutyniok, "Frames of subspaces," *Contemporary Mathematics*, vol. 345, pp. 87–114, 2004.

[92] Y. M. Lu and M. N. Do, "Sampling signals from a union of subspaces," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 41–47, 2008.

[93] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[94] Y. C. Eldar, "Compressed sensing of analog signals in shift-invariant spaces," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 2986–2997, 2009.

[95] K. Gedalyahu and Y. C. Eldar, "Time-delay estimation from low-rate samples: A union of subspaces approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3017–3031, 2010.

[96] S. Kallummil and S. Kalyani, "Generalized residual ratio thresholding," *Signal Processing*, vol. 197, p. 108531, 2022.

[97] J. W. Choi and B. Shim, "Detection of large-scale wireless systems via sparse error recovery," *IEEE Transactions on Signal Processing*, vol. 65, no. 22, pp. 6038–6052, 2017.

[98] S. Aviyente, "Compressed sensing framework for eeg compression," in *2007 IEEE/SP 14th workshop on statistical signal processing*, pp. 181–184, IEEE, 2007.

[99] T. Kronvall, S. I. Adalbjörnsson, S. Nadig, and A. Jakobsson, "Group-sparse regression using the covariance fitting criterion," *Signal Processing*, vol. 139, pp. 116–130, 2017.

[100] I. Fedorov, R. Giri, B. D. Rao, and T. Q. Nguyen, "Robust bayesian method for simultaneous block sparse signal recovery with applications to face recognition," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3872–3876, IEEE, 2016.

[101] P. B. Gohain and M. Jansson, "Robust information criterion for model selection in sparse high-dimensional linear regression models," *arXiv preprint arXiv:2206.08731*, 2022.

[102] P. B. Gohain and M. Jansson, "New improved criterion for model selection in sparse high-dimensional linear regression models," in *ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5692–5696, 2022.

[103] J.-F. Determe, J. Louveaux, L. Jacques, and F. Horlin, "On the exact recovery condition of simultaneous orthogonal matching pursuit," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 164–168, 2015.

[104] H. Li, L. Wang, X. Zhan, and D. K. Jain, "On the fundamental limit of orthogonal matching pursuit for multiple measurement vector," *IEEE Access*, vol. 7, pp. 48860–48866, 2019.

[105] J. Wen, H. Chen, and Z. Zhou, "An optimal condition for the block orthogonal matching pursuit algorithm," *IEEE Access*, vol. 6, pp. 38179–38185, 2018.

[106] Y. Shi, L. Wang, and R. Luo, "Sparse recovery with block multiple measurement vectors algorithm," *IEEE Access*, vol. 7, pp. 9470–9475, 2019.

[107] P. Pal and P. Vaidyanathan, "Pushing the limits of sparse support recovery using correlation information," *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 711–726, 2014.

[108] X. Lv, G. Bi, and C. Wan, "The group lasso for stable recovery of block-sparse signal representations," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1371–1382, 2011.