Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo XXXIV

Settore Concorsuale: 13/D1 - STATISTICA
Settore Scientifico Disciplinare: SECS-S/01 - STATISTICA

MODEL SELECTION AND THE VECTORIAL MISSPECIFICATION-RESISTANT INFORMATION CRITERION IN MULTIVARIATE TIME SERIES

Presentata da: GERY ANDRÉS DÍAZ RUBIO

Coordinatore Dottorato
Monica Chiogna

Supervisore
Simone Giannerini

Co-Supervisore
Greta Goracci

Esame finale anno 2022

MODEL SELECTION AND THE VECTORIAL MISSPECIFICATION-RESISTANT INFORMATION CRITERION IN MULTIVARIATE TIME SERIES

RELEVANT COMPETITION SECTOR: 13/D1

ACADEMIC DISCIPLINE: SECS-S/01

PHD PROGRAMME IN STATISTICS

CYCLE XXXIV

CANDIDATE: GERY ANDRÉS DÍAZ RUBIO

COORDINATOR: PROF. MONICA CHIOGNA

SUPERVISOR: PROF. SIMONE GIANNERINI

CO-SUPERVISOR: DR. GRETA GORACCI

DEPARTMENT OF STATISTICAL SCIENCES "PAOLO FORTUNATI"

UNIVERSITY OF BOLOGNA

BOLOGNA

30 SEPTEMBER 2022

Dedicada
a mi amada y siempre presente familia,
a mis queridos amigos.

Dedicated to the loving memory of
Mario Maffei Arzate,
Theoretical Physicist
1985 - 2018

A mi Madre,
A mi Mimi.

ABSTRACT

The thesis deals with the problem of Model Selection (MS) motivated by information and prediction theory. The focus is on parametric time series models. The main contribution of the thesis is the extension to the multivariate case of the Misspecification-Resistant Information Criterion (MRIC), a criterion introduced recently that manages to solve the original research problem posed by Akaike 50 years ago, which led to the definition of the well known AIC. Since modern statistics, MS is a fundamental task, both necessary and challenging in contemporary applications and in algorithmic solutions, e.g. big data, high-dimensionality, nonlinearity, automation, and machine learning. The importance of MS is witnessed by the huge amount of literature devoted to it and published in scientific journals of many different disciplines. Despite such a widespread treatment, the contributions that adopt a mathematically rigorous approach are not so numerous and one of the aim of the present project is to review and assess them. Chapter 2 discusses methodological aspects of MS from the perspective of information theory. Common information criteria for the *i.i.d.* setting are surveyed along with their main asymptotic properties. The cases of small samples, misspecification, and further estimators are examined. Chapter 3 surveys criteria for time series. Information and prediction criteria for parametric univariate models (AR, ARMA) in the time and frequency domain are considered. The settings of parametric multivariate (VARMA, VAR), nonparametric nonlinear (NAR), and high-dimensional MS are also covered. As mentioned, the MRIC approach answers to the original question posed by Akaike on efficient criteria, for possibly-misspecified univariate time series models, managing multi-step prediction with high-dimensional data and nonlinear models. Chapter 4 extends the MRIC to possibly-misspecified multivariate time series models for multi-step prediction and introduce the Vectorial MRIC (VMRIC). We show that the VMRIC is an asymptotically efficient MS method. To this aim, we prove the asymptotic decomposition of the Mean-Squared Prediction Error (MSPE) matrix, and the asymptotic consistency of its Method-of-Moments Estimator (MoME), for the Least Squares (LS) multi-step prediction of multivariate time series with an univariate regressor, for possibly-misspecified models. Furthermore, Chapter 5 shows that the VMRIC is valid for the general case of multiple regressors. To this aim, we prove that the MSPE matrix decomposition holds, obtain asymptotic consistency for its MoME, and proofs its asymptotic efficiency. The chapter concludes with a digression on the conditions for possibly-misspecified vector autoregressive models with exogenous variables (VARX).

PUBLICATIONS AND WORKING PAPERS

- Gery Andrés Díaz Rubio, Simone Giannerini, and Greta Goracci. *"On the asymptotic mean-squared prediction error for multivariate time series."* In: Book of Short Papers SIS 2021, (2021), pp. 1599 - 1604, Pearson. 50^{th} scientific meeting of the Italian Statistical Society, Pisa, 21-25 June 2021. See [97].

- Gery Andrés Díaz Rubio, Simone Giannerini, and Greta Goracci. *"A multivariate extension of the Misspecification-Resistant Information Criterion."* (2022) Working paper. See [98]. Poster presented at the 3^{rd} Italian Workshop of Econometrics and Empirical Economics: "High-dimensional and Multivariate Econometrics: Theory and Practice" (IWEEE 2022) of the Italian Econometric Society, Rimini, 20-21 January 2022.

- Gery Andrés Díaz Rubio, Simone Giannerini, and Greta Goracci. *"Model selection via information and prediction criteria: a survey for the i.i.d. case"* (2022) Working paper.

- Gery Andrés Díaz Rubio, Simone Giannerini, and Greta Goracci. *"Model selection via information and prediction criteria: a survey for time series"* (2022) Working paper.

ACKNOWLEDGEMENTS

I would first like to thank Professor *Simone Giannerini*, my doctoral supervisor, from the Department of Statistical Sciences "Paolo Fortunati" at the University of Bologna. Since the beginning of this research work, the insightful comments of Professor Giannerini inspired and helped me in finding this interesting and challenging subject. He undoubtedly allowed this thesis to be my own work, but our correspondence and continuous communication during these months were fundamentals to steer me in the right direction. His availability and kindness, together with his precious technical indications, were fundamental during every step of this period.

I would also like to thank my doctoral co-supervisor Dr. *Greta Goracci* of the Faculty of Economics and Management at the Free University of Bozen-Bolzano. The door of Dr. Goracci was always open whenever I ran into a trouble spot, and helped me continuously since the beginning this research. Her brilliant advices helped in improving profoundly the outcome of this work.

I would also like to acknowledge faculty members and my fellow colleagues of the doctoral programs in Statistical Sciences, Economics and Data Science, at the University of Bologna. The encounter with them during these years enriched valuably me and this research. The doctoral courses were essential to this work, so a special acknowledgement goes to my professors. I am gratefully indebted to them for their precious comments and indications. In particular, the discussions with faculty members with whom I had the pleasure to work with, was a key aspect as their availability to combine didactic work with my research. The quality of this thesis was significantly improved thanks to the useful suggestions of the two reviewers, Professor Pietro Coretto at the Free University of Bozen-Bolzano and Professor Davide Ferrari at the Università degli Studi di Salerno.

Finally, I must express my profound gratitude to my mother, my grandmother, my brother, both my aunts in Peru and Japan, and to my friends, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Some of them include alphabetically: Alberto Ida, Armando Celico, Gino Díaz, Grace Rubio, Paulo Levano, Riccardo Cimarosti, Virna Rubio, Wilma Berrospi, Yvette Rubio. Thank you very much.

This work would not have been possible without the financial support of the Department of Statistical Sciences "Paolo Fortunati" and the Italian *Ministero dell'Università e della Ricerca* through the PhD Scholarship at this department.

CONTENTS

LIST OF FIGURES

ACRONYMS

AIC A Information Criterion

AICC Corrected A Information Criterion

AFPE Asymptotic FPE

APE Accumulated Prediction Error

BIC Bayesian Information Criterion

CAIC Consistent AIC

CAICF CAIC with Fisher information

CAT Criterion autoregressive transfer function

Cp Mallows' C_p

CV Cross-Validation

DGP Data Generating Process

DIC Deviance Information Criterion

FD Fixed Dimensionality

FDR False Discovery Rate

FIC Focused Information Criterion

FPE Final Prediction Error

GAIC Generalized AIC

GBIC Generalized BIC

GIC Generalized Information Criterion

GLR Generalized Likelihood Ratio

HDIC High-Dimensional Criterion

HQ Hannan-Quinn Information Criterion

IC Information Criteria

ID Increasing Dimensionality

KC Kashyap's Criterion

LS Least Squares

MDL Minimum Description Length

MoME Method-of-Moments Estimator

MS Model Selection

OLS Ordinary least squares

PC Prediction Criteria

PLS Predictive Least Squares

PSS Prediction Sum of Squares

RIC Risk Inflation Criterion

RSS Residual Sum of Squares

SRIC Shibata's Regularized Information Criterion

MRIC Misspecification-Resistant Information Criterion

MSPE Mean-Squared Prediction Error

TIC Takeuchi Information Criterion

VMRIC Vectorial MRIC

WAIC Widely Applicable Information Criterion

WBIC Widely applicable Bayesian Information Criterion

Part I

INTRODUCTION

1

MODEL SELECTION: INFORMATION AND PREDICTION CRITERIA

1.1 OVERVIEW

This thesis extends the Misspecification-Resistant Information Criterion (MRIC) proposed in [H.-L. Hsu, C.-K. Ing, H. Tong: *On model selection from a finite family of possibly misspecified time series models*. The Annals of Statistics. 47 (2), 1061–1087 (2019)] [153] to possibly- misspecified multivariate time series models for h-step ahead prediction. The MRIC tackles Akaike's original research question on efficient Model Selection (MS) for possibly- misspecified models that 50 years ago led to the popular A Information Criterion (AIC).

Contribution 1

The first contribution is a vast survey composed of two chapters. These are two selective but broad surveys of the last fifty years to study the impact of information and prediction theory on statistical MS, and how this *connubium* can be related to successive developments in itself and alternative data-oriented strategies. Chapter 2 introduces the statistical problem of MS via information and prediction criteria for independent and identically distributed *(i.i.d.)* data. Departing from Akaike's seminal paper in 1973 [5], we follow his trail on the initial heuristic derivation of the AIC and study its formal mathematical proof. The asymptotic properties of criteria are taken into account, detailing specific definitions connected with the different senses of optimality. We expand our focus to include information and prediction criteria from the Bayesian perspective, regression analysis, and resampling techniques, given their relevance in successive development of the field and their influence in contemporary solutions. The first survey concludes with the study of practical situations that modify asymptotic results: small sample, model misspecification, and alternative estimators.

In Chapter 3, the second survey targets time series models. Departing from Akaike 's seminal paper in 1969 [1], we briefly follow his considerations to obtain the Final Prediction Error (FPE). Derived solutions for MS in autoregressive (AR) models are considered, both in the time and in the frequency domain, including a short introduction to the MRIC. We move to general autoregressive moving-average (ARMA) models, with special attention to the formal setting required for the extension of criteria to these types of models. This led us to view the solutions proposed from Rissanen's Accumulated Prediction Error (APE) and its stochastic regression extension with applications to time series. Furthermore, the problem of MS for multivariate time series models is

studied, to better ground our contributions in the successive chapters. Issues and solutions for vector ARMA (VARMA) models are examined, together with part of the literature for MS with vector AR (VAR) models. This survey concludes with the theoretical and methodological framework behind the Asymptotic FPE (AFPE), i.e. the nonparametric counterpart of the FPE for nonlinear time series models, and some notes on algorithmic approaches and modern solutions to MS in high-dimensional settings.

Contribution 2

The second contribution presents the first extension of the MRIC to multivariate time series with univariate regressor, in the form of three theorems. Chapter 4 establishes the asymptotic decomposition of the Mean-Squared Prediction Error (MSPE) matrix for h-steps ahead least-squares predictor, the asymptotic consistency of the Method-of-Moments estimator (MoME) for the related quantities, and the asymptotic efficiency of the VMRIC as a MS method. We showcase our theoretical derivation with an example where the misspecification is also considered, including simulation studies to assess criterion's performance.

Contribution 3

The third contribution shows the first full extension of the MRIC to possibly-misspecified multivariate time series models with multiple regressor for h-step ahead forecast. Chapter 5 proves three theorems establishing the asymptotic decomposition of the MSPE matrix for h-steps ahead least-squares predictor with multiple regressor, the asymptotic consistency of VMRIC's MoMEs, and its asymptotic efficiency as MS criterion. A digression on the technical conditions required for dynamic simultaneous equations models, also known as VAR model with exogenous regressor (VARX), to satisfy the assumptions required for the VMRIC approach is advanced.

Part II

CONTRIBUTIONS

Chapter 2 surveys information and prediction criteria for independent and identically distributed data. Chapter 3 reviews common solutions for model selection via information and prediction criteria for time dependent data. Chapter 4 presents the first extension of the Misspecification-Resistant Information Criterion (MRIC) [153], with three theorems and a technical lemma obtaining: the asymptotic decomposition of the mean-squared prediction error for weakly stationary h-step ahead possibly-misspecified multivariate time series models with univariate regressor, the derivation of our VMRIC, the asymptotic consistency of the method-of-moments estimator of the defined VMRIC, the asymptotic efficiency of the criterion as a model selection method, and one example with simulations. Chapter 5 completes the first full vectorial extension of the MRIC for the case of multiple regressors, showing that the asymptotic properties of the univariate regressor case still hold in the multiple setting, and advances a digression for possibly-misspecified vector autoregressive with exogenous regressors (VARX), also known as dynamic simultaneous equations models.

2

MODEL SELECTION VIA INFORMATION AND PREDICTION CRITERIA: A SURVEY FOR THE I.I.D. CASE

ABSTRACT

Developments on entropy and information theory encountered fertile ground when met the likelihood approach, multiple testing, regression and multivariate analysis. A path is traced starting around 1974 with Akaike's AIC's approach based on the Kullback-Leibler discrimination information for independently, identically distributed data under correct specification. Further information criteria are discussed in relation to both their asymptotic properties and common settings, e.g., small samples, model misspecification, alternative estimators. This chapter is intended as an introductory discussion on information criteria for *i.i.d.* data, laying the basis for Chapter 3 discussing derived solutions for time dependent data.

Keywords: model selection, information criteria, parametric models, consistency, efficiency, misspecification, small sample, functional estimators.

2.1 INTRODUCTION

Within statistical analysis, model selection (MS) deals with the problem of selecting the best model according to a specified measure. Different dimensions of the problem arise, since the goal of our analysis will influence our model selection's understanding. First, following Schmueli [279], we can differentiate between explanation, prediction, or description. Second, defining the "best" model will rely upon a particular measure, which is of interest in itself from the mathematical statistics standpoint and given its overall consequences. A third dimension of the problem deals with the actual sample size, since a portion of available methods are valid asymptotically for large samples. A fourth level relates to models' correct specification or misspecification. Initial developments involved the case where the 'true' model is among the set of candidates models, i.e. correct specification. If the 'true' model does not exists, or if it may not be postulated, we should pursue misspecification robust strategies.

Among sundry selection techniques, information and prediction criteria developed during the 1970's still enjoy popularity after almost fifty years. Information Criteria (IC) derive from information theory. One of its pioneers, the Japanese statistician Hirogutu Akaike (1927–

2009), proposed the popular A Information Criterion (AIC). Prediction Criteria (PC) refer to selection methods derived from the one-step ahead prediction error, e.g. Akaike's Final Prediction Error (FPE). This chapter focusses on the independent and identically distributed (*i.i.d.*) case. The time series setting will be the topic discussed in Chapter 3.

The plethora of available information and prediction criteria are used in vast areas of human knowledge. Studies for empirical and theoretical problems includes but are not limited to: high-dimensionality, geostatistics, ecology, medicine, phylogenetics, robust estimation, genetics, econometrics, demography, epidemiology, biology, sociology, reaction-diffusion problems in mathematical biology, copula methods, astrostatistics, astrophysics, regularization parameter selection, model averaging, bootstrap variants for small sample mixed models, dynamical systems. In a context of continuously increasing computational capabilities, criteria are employed extensively and also in combination with other algorithmic or data-oriented procedures. Sometimes the presence of these criteria in *canned software packages*[1] may illude its application is indeed straightforward to the point that no further analysis is needed. That is not the case.

In Section 2.2 we expand on the motivation behind this survey, while Section 2.4 discusses Akaike' criterion, including instructions on its rigorous use and some issues with hypothesis testing that justified its introduction in the 1970's. This discussion is continued in Section 2.3 with notes on the original derivation of the AIC and on its connection with recent research. Definitions of asymptotic consistency and efficiency are considered in Section 2.5, together with that of pointwise and uniform convergence. Several refinements over IC for small sample, model misspecification, and further estimators are given in Sections 2.6 and 2.7 respectively.

2.2 MOTIVATION

Concerns on the use of machine-learning algorithms dedicated to automated decision-making on individuals and data ethics are present nowadays [34, 35, 84, 320, 336]. Real-world attempts to tackle the issue are found, for instance, in the Art. 22 of the United Kingdom General Data Protection Regulation (UK GDPR) on "Automated individual decision-making, including profiling", or Art. 29 from the European Union GDPR (EU GDPR) and related key provisions that reference general profiling and automated decision-making.
Discussions on MS derive from diverse scientific communities, both from theoretical and applied areas. The statistical community shares concerns, as the plenary talk of Professor Candès at the 2020 Bernoulli-IMS symposium [63] showed, motivating the interest with examples from sensitive applications such as facial-recognition scans, decision

1 Expression due to Kilian and Lütkepohl [169, p. 56].

tree algorithms for classification in inmate- and jail-management systems, the use of artificial intelligence in human-resources activities including recruiting, or medicine. Contemporary theoretical and empirical considerations are also driven by this interest, aimed at improving current methods for selection, estimation, and prediction. Given the increasing levels of complexity in our techniques and the difficulties we face to obtain neutral comparison studies, we attempt to 'back-to-basics' while underlining theoretical key-points, in terms of assumptions and general settings for MS via IC.

We propose a broad selective survey of the last fifty years to study: *(a)* the impact of information theory in statistical MS, and *(b)* how this *connubium* can be related to successive developments in itself and alternative data-oriented strategies. This since algorithmic culture [54] experienced an important expansion in the last twenty years, thanks to the more readily availability of vast quantity of data. Also given that information theory and machine-learning algorithms are two faces of the same coin [201]. A secondary objective is to give practitioners an accessible guide of definitions, remarks and algorithms to consider in empirical applications of MS methods.

MS is intertwined with the estimation process [57] and has strong effects on inference [188]. We will focus in the necessary work after obtaining parameters' estimates with specific properties, although recent methods incorporate the estimation step. As in many other fields, we see a trend from stronger to weaker sets of assumptions. This, combined with the exponential growth of our processors contributed to the development of nonparametric and semiparametric literature. Recent developments are scattered along the lines of *(i)* less restrictions and more flexibility (e.g. conformal prediction, feature-matching, nonlinear modelling, stochastic regression, robust estimation, Bayesian approach, mixture models), *(ii)* ensemble of techniques, and *(iii)* machine learning algorithms.

We still have issues in selecting, ordering, or ranking models according to a specific sense. The combination of different methodologies and practices makes it relevant to understand the theoretical setting to improve empirical works and develop further solutions. In that sense, we stress the importance of theoretical and methodological considerations to improve both theory and practice altogether.

We drive now the attention to Akaike's seminal contributions [7, 16], and we present comments on contemporary issues related to these.

2.3 HIROTUGU AKAIKE AND INFORMATION THEORY

The work by Kullback and Leibler [177] generalized the ideas of Shannon and Weaver [262, 263]. Let (X, S, μ_i), $i = \{1, 2\}$ be a probability space such that μ_1 and μ_2 are mutually absolutely continuous. Moreover, define λ to be any linear combination of μ_1 and μ_2. By the Radon-

Nikodym theorem [131, pp. 128-129], there exists a positive and finite probability density functions $f_i(x)$, $i = 1, 2$, unique (a part from a set of measure zero in λ) and λ-measurable, such that:

$$\mu_i(E) = \int_E f_i(x) d\lambda(x), \text{ for all } E \in S.$$

Our interest lies in evaluating the hypothesis H_1 that x comes from a population with probability measure μ_1 against the hypothesis H_2 that x comes from a population with probability measure μ_2. For that matter we define the information in x for discrimination between H_1 and H_2 as the difference $\log \frac{f_1(x)}{f_2(x)}$. At this point, we are ready to define the Kullback-Leibler Information (KLI),[2] $I(1:2)$, that measures the divergence between two probability distributions. Note that this relative entropy measure is a pseudo-distance (not a distance) since the triangle inequality does not hold, i.e. it is not the same quantity from f to g than from g to f.

Definition 1. *The mean information for discriminating between H_1 and H_2 is defined by:*

$$\begin{aligned} I(1:2) = I_{1:2}(X) &= \int_X \log \frac{f_1(x)}{f_2(x)} \, d\mu_1(x) \\ &= \int_X f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x). \end{aligned}$$

Properties 1. *The KLI $I(1:2)$ features, among others, the following basic properties:*

$$\begin{aligned} &I(1:2) \geq 0, \\ &I(1:2) = 0 \Leftrightarrow f_1(x) = f_2(x), \\ &I_{1:2}(E) \geq \log \frac{\mu_1(E)}{\mu_2(E)}, \text{ for } \lambda(E) > 0, \ E \in S, \\ &I_{AB}(1:2) = I_A(1:2) + I_B(1:2), \text{ for events } A \perp\!\!\!\perp B. \end{aligned}$$

Here and in the following, $A \perp\!\!\!\perp B$ reads *"event A is independent of event B"*. For further details, refer to Kullback [175], and for alternative measures of divergence, see Konishi and Kitagawa [174, p. 31]).[3] In Akaike's work [4, 5, 7] the KLI has been employed to address the problem of model identification. Let $g(x)$ be the data generating probability distribution and $\mathcal{F}_\theta = \{f(x|\theta_k), k = 1, \ldots, L\}$ be a probabilistic class

2 Kullback [176] preferred the term Discrimination Information.

3 These include, but are not restricted to: the χ^2-statistic, the Hellinger distance, the Generalized information, the Divergence, the L^1-norm, the L^2-norm, the Jensen-Shannon divergence, Jeffreys divergence, Chernoff's α-divergence, the exponential divergence, Kagan's divergence, the (α, β)-product divergence, the Battacharyya divergence. Further notions can be found for instance in the field of information geometry. On the topic, we indicate the recent volume of Rao, Rao, and Plastino on Information Geometry [232]. This is an active and interesting line of research combining many of the topics here presented in a unified manner.

of models such that $g(x) \in \mathcal{F}_\theta$. The problem is defined in the following terms.

Given the data $x \sim g(x)$, we want to identify the member of $\mathcal{F}_\theta$ that corresponds, or is the closest, to $g(x)$. If $\theta_k \in \mathbb{R}^k$, the identification problem reduces to select the true order of the model, $k \in \{1, \ldots, L\}$. By assuming that $g(x)$ is absolutely continuous with respect to the Lebesgue measure, the KLI between $g(x)$ and any parametric family $f(x|\theta) \in \mathcal{F}_\theta$ is:

$$I(g; f(x|\theta)) = \int g(x) \frac{\log g(x)}{\log f(x|\theta)} dx. \tag{1}$$

Akaike's consideration of the KLI as a separation measure between two probability density functions in the context of the statistical problem of model identification was novel. By doing so, he combined information theory and its own extension of the Maximum Likelihood (ML) principle. The latter is stated by Akaike [5] in the following manner:

Definition 2. *Given a set of estimates $\hat{\theta}$ of the parameters' vector θ of a probability distribution with density function $f(x|\theta)$ we adopt as our final estimate the one which will give the maximum of the expected log-likelihood, i.e.*

$$E\left[\log\left(f(X|\hat{\theta})\right)\right] = \int f(x|\theta) \log\left(f(X|\hat{\theta})\right) dx. \tag{2}$$

Akaike noticed that the ML principle was equivalent to maximize minus $E[I(f(x|\theta); f(x|\hat{\theta}))]$, i.e. minus:

$$E\left[\log\left(\frac{f(X|\hat{\theta})}{f(X|\theta)}\right)\right] = \int f(x|\theta) \log\left(\frac{f(X|\hat{\theta})}{f(X|\theta)}\right) dx. \tag{3}$$

His version of the ML principle, together with his definition of the loss and risk functions (detailed in the following paragraphs), led Akaike to avoid seeing both the estimation and testing theory as separated, but rather as a single problem of statistical decision [16, p. 610]. This was stated by Akaike given the use of both the ML principle and the LR test statistic, connecting the concepts by means of loss and risk functions from statistical decision theory, through the lenses of information theory. Specifically, this single problem of statistical decision was faced as a problem of MS: to select $f(x|\theta_k)$, $k = \{0, 1, 2, \ldots, L\}$, based on the observations of the random variable X.[4] Note that the parameter vector θ_k is *restricted to the space* where the $\{k+1, k+2, \ldots, L\}$ parameters are set equal to zero, $\theta_k \in \Theta_k$, with Θ_k being the set of parameters' space.

To proceed with the problem at hand of identifying model's correct order k, the use of Wald's log-LR test [319] for composite hypothesis

4 Further details in Akaike [5, Section 4], Bozdogan [52, p. 351], and deLeeuw [361, p. 601].

testing with hierarchical (nested) models was critically indicated by Akaike [5] as a possible path for its solution, after some adaptations. Akaike [5, 7] proposed an *heuristic approach* in the following terms.

Let $\{x_1, \ldots, x_N\}$ be a random sample of the random variable X, $x_i \sim X$, and let $\hat{\theta}$ be the maximum likelihood estimator (MLE) of θ, i.e.

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(x|\theta), \tag{4}$$

where $\mathcal{L}(x|\theta)$ is the likelihood function on the random sample x, given the parameter θ. In the context of statistical decision[5] define the loss and risk[6] functions between the 'true' parameter and the estimated parameter as stated below. Given that the loss function is minus two times the KLI between the true parameter and some estimate of it, it will feature its same properties defined previously after the proper adaptations. This is stated in the following:

Definition 3. *Define the loss function:*

$$W(\theta, \hat{\theta}) = (-2) \int f(x|\theta) \log \left(\frac{f(X|\hat{\theta})}{f(X|\theta)} \right) dx, \tag{5}$$

and its respective risk function as its expected value:

$$R(\theta, \hat{\theta}) = E_{\hat{\theta}} \left[W(\theta, \hat{\theta}) \right], \tag{6}$$

where the expectation $E_{\hat{\theta}}[\cdot]$ is computed with respect to the distribution of $\hat{\theta}$.[7]

Akaike advocated to estimate the loss function $W(\theta, \hat{\theta})$ as minus two times the *sample mean of the log-likelihood ratio* since it is the natural estimator [7], in addition to being well known and extendedly used in the literature. This is stated formally in the following definition:

Definition 4. *Let $f(x_i|\hat{\theta}_L)$ be the density function of the probability model of reference, i.e. the one with 'true' set of parameters, and $f(x_i|\hat{\theta}_k) \in \mathcal{F}_\theta = \{f(x|\theta_k), k = 1, \ldots, L\}$. Then the natural estimator of the loss function is:*

$$\omega_{k,L} = -\frac{2}{N} \sum_{i=1}^{N} \log \left(\frac{f(x_i|\hat{\theta}_k)}{f(x_i|\hat{\theta}_L)} \right). \tag{7}$$

If we multiply this estimator of the loss function by N, we obtain the usual Log-Likelihood Ratio (LR) test statistic:

$$\eta_{k,L} = N \times \omega_{k,L}, \tag{8}$$

which asymptotically behaves as a χ^2_{L-k} distribution when $\theta \in \theta_k$.

5 See Cox and Hinkley [86, p. 429].

6 See Grünwald [127, p. 515].

7 See Konishi and Kitagawa [173, p. 877]) for an example of generalization to functional estimators via the empirical distribution function. See also Kitagawa and Konishi [171].

Remark 1 (Akaike [16])**.** *Defining the loss function between both the true and the reduced parameter vector (to the k-th order) to be the infimun of those obtained with different values of k, i.e.*

$$W(\theta, \theta_k) = \inf_{\theta_k} W(\theta, \theta_k), \tag{9}$$

then, it is expected that the sample estimate will converge to its population value (i.e. strongly consistent estimator)

$$\omega_{k,L} \to W(\theta, \theta_k) \; a.s. \tag{10}$$

Akaike considered three situations for the convergence of the LR statistic to the χ^2 distribution (deLeeuw [361, p. 604], Akaike [16, p. 205]). If the true parameter θ belongs to the reduced parameter space Θ_k (or the reduced parameter space θ_k is a fair approximation of the true parameter space θ), i.e. $\theta \in \Theta_k$, then we may see that:

i. If $NW(\theta, {}_k\theta)$ is much larger than L, the approximation of the log-likelihood ratio test ${}_k\eta_L$ fails to converge to the χ^2 distribution since it would be larger than the chi-square approximation.

ii. If $NW(\theta, {}_k\theta)$ is much smaller than L, the approximation is appropriate, and we can proceed with the analysis using the LR statistic as usual.

iii. If $NW(\theta, {}_k\theta)$ is close to L, a more precise analysis of the behaviour of the LR test statistic is needed.

Given that the first case is not viable, and the second case is within the LR theory, interest lays on the last one. For this third case, Akaike considered asymptotic (e.g. Taylor expansion) and nonasymptotic arguments to modify the loss function, by the use of the KLI in the parametric case [177, p. 81]. This is related to study the discrepancy between two density functions, where the first is some density function $f(x, \theta)$ computed at θ, while the second is the same density function in a neighbourhood of its parameter space, $f(x, \theta + \Delta\theta)$. An additional requirement for the modification of the loss function was the consistency of the MLEs $\hat{\theta}$ and $\hat{\theta}_k$ for both the full parameter space (θ) and its reduced version (θ_k) respectively. In general, it is required the asymptotic efficiency of the MLEs [7, p. 718].[8] This will lead us to solve the MS problem by considering the *sample mean log-likelihood* to be a proper measure of fit of the model. Let us follow Akaike's steps.

If θ and θ_k are very near to each other, we are allowed to focus on the second-order variations of $W(\theta, \hat{\theta})$.[9] In this case, the initial loss function, which used the KLI, can be modified into:

$$W_2(\theta, \hat{\theta}_k) = \sum_{l=1}^{L} \sum_{m=1}^{L} (\hat{\theta}_{k,l} - \theta_l)(\hat{\theta}_{k,m} - \theta_m) C_{l,m}(\theta), \tag{11}$$

8 Akaike used MLE's asymptotic consistency, efficiency and normality.

9 Kullback and Leibler [177, p. 81]), citing Doob [103, p. 774]), highlight how this argument depends on suitable assumptions on the density function. See also Sawa [254, p. 1274].

where $C_{l,m}(\theta)$ is the (l, m)-th element of Fisher's information matrix, abbreviated in the following by $C_{l,m}$ for simplicity. A version of this modified loss function is:

$$W_2(\theta, \hat{\theta}_k) = \left\| \hat{\theta}_k - \theta \right\|_C^2, \tag{12}$$

where $\|\theta\|_C$ is the norm in the space of θ defined as

$$\|\theta\|_C = \sum_{l=1}^{L} \sum_{m=1}^{L} \theta_l \theta_m C_{l,m}.$$

An estimate for the risk function of the modified loss function $E[W_2(\theta, \hat{\theta}_k)]$, when N is sufficiently large, and both L and k are relatively large integers, is:

$$r(\hat{\theta}, \hat{\theta}_k) = N^{-1}(\eta_{k,L} + 2k - L). \tag{13}$$

For MS through comparison of different models with different orders k against the true one, note that the L value in the previous estimate of the risk function is the same since it only depends on the dimension of the true parameter space. For this reason it can be dropped from the estimate of the modified risk function without influencing the result.

Before proceeding with the formal definition of the AIC present in Section 2.4, we shall briefly highlight some points from the arguments presented until now.

2.3.1 *Discussion*

This subsection collects brief thought stimulating comments and the relation with current issues derived from Akaike's seminal contributions introduced in this section.

RESTRICTING PARAMETERS' SPACE The strategy of restricting the parameters' space was already present in the literature, e.g. subset regression [147, 203, 204]; discriminant analysis [229], and can be seen as connected with the broad successive developments on MS literature, e.g. subset selection, regularization, shrinkage [194, pp. 504-505], stochastic complexity [245, p. 47], or dimension reduction [162, p. 203].

AUTOMATION OF SELECTION PROCEDURE In terms of the data *vis-à-vis* algorithmic modeling cultures as in Breiman [54], recent developments were influenced by ideas from data-oriented procedures, which allow for some type of automation of the selection procedure. Automation is intended in the sense of the implementation of procedures in the MS phase which eliminate practitioner's intervention. This step is sometimes necessary, i.e. no alternatives given that alternatives procedures fail. The objective of the analysis is pivotal: description-identification, or prediction-selection. In this sense, automation is both a debatable but important goal in contemporary applications.

INFORMATION AND HYPOTHESIS TESTING An additional early example of information-theoretic quantities and hypothesis testing can be found in Blahut [45]. Teräsvirta and Mellin [293] studied the connection between the ordinary F test and IC when all models are linear and we have to select the model from nested finite number of candidate alternatives. Vuong [318] proposed a LR test approach using the KLI to measure the divergence between non-nested, overlapping, and nested, also for the case of one, or neither misspecified model.

WHY 'TIMES TWO'? The use of the "magic number 2" (Stone [287, p. 32]) is a convention initially criticized as arbitrary, e.g. Rissanen [241], deLeeuw [361, p. 602]. Theoretical and historical clarifications are given in Akaike [15], Bozdogan [52, p. 356-357]), Burnham and Anderson [59, p. 64]). Successive developments consider different types of penalizations as bias correction. For instance, see Kitagawa and Konishi [171] for bias and variance reduction techniques with a generalization of the AIC; Yanagihara et al. [345] for bias correction of AIC in logistic regression models; and Davies et al. [90] for linear regression.

CONDITIONS AND CASES In the literature, relaxing these assumptions will be dealt in successive generalizations of the AIC, e.g. small or moderate samples, misspecified models, non-*i.i.d.* observations, alternative estimators (viz. conditional, penalized, composite, quasi likelihood, robust, functional), nonparametric statistics, high-dimensionality.

EVALUATION OF THE PERFORMANCE OF IC The asymptotic performance of IC *depends on the problem upon which it is being applied*. An argument for this statement is Akaike's entropy-maximization principle, which recites: *"All statistical activities are directed to maximizing the expected entropy of the predictive distribution in each particular application"* [15, p. 17]. Changing the problem, asymptotic properties behave differently, e.g. under small or moderate samples. We underline it in the light of broad literature in problem definition, decision theory and loss functions. For example, on the loss function, Shibata [274, p. 417] stated: *"In our problem, the choice of the loss function is crucial for discussing the goodness of a selection procedure"*. Partially, this perspective was also present on Ragnar's Frisch 1970 Nobel prize lecture on econometrics [118]. Given that the goal of statistical analysis often is to obtain reliable and precise predictions for decision-making, this connection should guide us to follow methodologies in a precise manner, possibly in a unified way.

INCREASING- AND FIXED-DIMENSIONALITY SETTINGS Akaike's perspective on the study of asymptotic properties enriched the field, i.e. theoretical considerations of an empirical problem: to assess the properties of a MS criterion as n grows to infinity but keeping the 'true'

order k of the model fixed, i.e. Fixed Dimensionality (FD) setting. This approach complements the case when k grows to infinity while also n goes to infinity, i.e. Increasing Dimensionality (ID) setting. See the literature in Schorfheide [256] for indications on its multi-step (direct) loss function based estimation. This distinction further modifies asymptotic properties of IC. See the discussion in Stone [286, p. 277], Leeb and Pötscher [187], and the distinction of cases for time series in Hsu et al. [153, pp. 1061-1065]. Chapters 4 and 5 contribute to the literature in the FD setting for multivariate time series MS.

2.4 MODEL SELECTION WITH INDEPENDENT AND IDENTICALLY DISTRIBUTED DATA

The practicality of Akaike's idea and its lean computation were two drivers for its vast diffusion in statistical sciences and beyond. We observe an *i.i.d.* sample of dimension n, $\{x_1, x_2, \ldots, x_n\}$, taken from some Data Generating Process (DGP). In general, approximate estimates of information criteria are of the following form:

$$\mathrm{IC}(k) = -2\, l(\hat{\theta}_k) + \mathrm{Penalty}, \tag{14}$$

where $l(\cdot)$ denotes the log-likelihood of the estimated reduced model with k parameters and the penalty will depend upon the particular setting of the problem. Our focus is on mathematically-grounded IC.

A short disclaimer is in order. Supplementary paths for MS include, among others, the minimum description length principle [141, 241, 244, 245]; the Bayesian perspective [83, 167, 234, 258, 321]; mathematical decision theory [46]; sieves and approximation theory [29]; resampling or data augmentation methods such as Cross-Validation (CV) [20, 24, 87, 264, 285], bootstrap [265] or shrinkage [65]; machine-learning procedures which also deliver feature selection such as bagging [210], sparse boosting [58], lasso [357], nonconcave penalized likelihood via smoothly clipped absolute deviation (SCAD) [109], neural networks [23, 179], or random forests [92, 281]. Refer to [100, 164, 187, 192, 207, 236, 295, 296] for reviews and surveys.

We start by devoting the rest of this section to understand the theoretical framework behind the A Information Criteria (AIC).

2.4.1 *A Information Criterion (AIC)*

This subsection includes the assumptions, a proposition and formal derivation of the AIC in Section 2.4.1.1. Previously, in Section 2.3 we have followed partially the heuristic derivation of Akaike from his two seminal papers [7, 16]. Here we focus on the formal proof presented by Bozdogan [52]. This is similar to the steps traced until now, so its

development should be eased. A slight theoretical difference is that Bozdogan uses the Entropy Maximization Principle (EMP) [10], which is equivalent to the minimization of the KLI quantity. As we have seen, the latter is connected to Akaike's ML principle. Alternative strategies for the formal proof can be found in [59, 174]. As one reviewer underlined, the technique developed by Akaike is simple, but it is seminal for many of the following contributions. Section 2.4.1.2 presents 15 points to keep in mind for the correct use of the AIC, and then briefly discusses the connection between the AIC and the LR statistic.

2.4.1.1 *Assumptions, definition, and formal derivation*

The AIC in Eq. (16) is an estimate of a measure fit of the model [7, p. 716] where the mean log-likelihood is taken as the preferred measure of fit. First, we present the three assumptions in detail. Then, we follow [52] in Proposition 1 which defines both the AIC and its natural sample estimator, and its respective sketch of proof.

Assumptions 1. *For Eq. (16) to hold the following assumptions are needed:*

(i) $\{x_1, x_2, \ldots, x_n\}$ *are* n *independent observations of a random variable with density function* $g(x)$.

(ii) The maximum likelihood estimator (MLE) based on n *observations,* $\hat{\boldsymbol{\theta}}_n$*, of the 'true' parameter vector* $\boldsymbol{\theta}$ *is estimated under regularity conditions that deliver asymptotic efficiency and normality. The expected log-likelihood is estimated by its natural estimator, i.e. the log-likelihood function evaluated at its supremum* $\hat{\boldsymbol{\theta}}$*:* $n^{-1}l(\hat{\boldsymbol{\theta}}_n) = n^{-1}\sum_{i=1}^{n}\log f(x_i|\hat{\boldsymbol{\theta}}_n)$.

(iii) $f(x|\boldsymbol{\theta})$ *is a parametric family of density functions of the random variable* X*, depending on the parameter vector* $\boldsymbol{\theta}$ *which includes the true model, i.e.* $f(x|\boldsymbol{\theta}_0) = g(x)$*. In other words, the model is correctly specified:* $g(x;\boldsymbol{\theta}) \in \mathcal{F}_{\boldsymbol{\theta}}$.

Proposition 1 ([52])**.** *Let a set of candidate models* $\{\mathcal{J}_k : k = 1, 2, \ldots, K\}$*, with* k *the index of the competing models. Define the population AIC as twice the expected KLI, or, equivalently, as minus twice the expected log-likelihood:*

$$\mathrm{AIC}\left(\boldsymbol{\theta}_k\right) = 2E\left[I\left(\theta^*, \theta_k\right)\right] = -2E\left[\log f\left(X|\theta_k\right)\right], \tag{15}$$

which is minimized to choose a model $\mathcal{J}_k$ *over the set of candidate models. Then, its natural sample estimator*[10] *is given by:*

$$\mathrm{AIC}\left(\hat{\boldsymbol{\theta}}_k\right) = -2\sum_{i=1}^{N}\log f(x_i|\,\hat{\boldsymbol{\theta}}_k) + 2k, \tag{16}$$

where $\sum_{i=1}^{N}\log f(x_i|\,\hat{\boldsymbol{\theta}}_k)$ *is the log-likelihood evaluated at the ML estimated parameter with reduced dimension* k.

10 This is an unbiased estimator of its population value.

Proof. We follow the geometrical derivation of Bozdogan [52, pp. 346-356] based on [5, 7, 9, 170]. Consider a DGP with $g(x) \equiv f(\cdot|\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^K$, from which all candidate models derive. Denote the 'true' parameter vector by $\boldsymbol{\theta} \equiv \boldsymbol{\theta}^* \in \mathbb{R}^K$, with K the total number of parameters of the 'true' full model.

Our goal is to select the model with probability density function $f(\mathbf{x}|\boldsymbol{\theta}_k)$ based on n observations, with parameter vector $\boldsymbol{\theta}_k \in \mathbb{R}^k$, where k is the number of free parameters, with $k < K$. This restricted parameter vector sets $\theta_{k+1} = \theta_{k+2} = \cdots = \theta_K = 0$. Obtain its MLE, $\hat{\boldsymbol{\theta}}_k$, by the usual maximization of the likelihood function, $L(\boldsymbol{\theta}_k|\mathbf{x}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta}_k)$, with respect to $\boldsymbol{\theta}_k$. Consider the log-likelihood function divided by n, i.e. the *mean log likelihood*, $n^{-1}l(\boldsymbol{\theta}_k) = n^{-1}\sum_{i=1}^n \log f(x_i|\boldsymbol{\theta}_k)$, which is a natural consistent estimator of the *expected log likelihood*, $E[\log f(\mathbf{X}|\boldsymbol{\theta}_k)] = \int \log f(\mathbf{x}|\boldsymbol{\theta}_k) f(\mathbf{x}|\boldsymbol{\theta}_k)\mathrm{d}\mathbf{x}$. Furthermore, take the KLI as the *loss function*:

$$I\left(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}\right) = \int \log\left[\frac{f(\mathbf{x}|\boldsymbol{\theta}^*)}{f(\mathbf{x}|\hat{\boldsymbol{\theta}})}\right] f(\mathbf{x}|\boldsymbol{\theta}^*)\mathrm{d}\mathbf{x},$$

and its expected value as its associated *risk function*:

$$E_{\mathbf{X}}\left[I\left(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}\right)\right] = \int I\left(\boldsymbol{\theta}^*; \hat{\boldsymbol{\theta}}\right) f(\mathbf{x}|\boldsymbol{\theta}^*)\mathrm{d}\mathbf{x}.$$

Now, the second-order expansion of $I(\boldsymbol{\theta}^*; \boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$, delivers:

$$I(\boldsymbol{\theta}^*; \boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) \cong \frac{1}{2}\|\Delta\boldsymbol{\theta}\|_{\boldsymbol{J}}^2, \tag{17}$$

where $\|\Delta\boldsymbol{\theta}\|_{\boldsymbol{J}}^2 = \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{J}}^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \boldsymbol{J}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, with $\boldsymbol{J}$ being the positive definite $(K \times K)$ Fisher information matrix:

$$\boldsymbol{J} = E\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{X}|\boldsymbol{\theta})\right]^\top \left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{X}|\boldsymbol{\theta})\right]\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right\}.$$

In order to restrict $\boldsymbol{\theta}^* \in \Theta_K$ to the k-dimensional restricted parameter space, Θ_k, with $k < K$, write the projection of $\boldsymbol{\theta}^*$ onto Θ_k and denote it by $\boldsymbol{\theta}_k^*$, and its MLE by $\hat{\boldsymbol{\theta}}_k$. In that case, we can write: $2I(\boldsymbol{\theta}^*, \boldsymbol{\theta}_k) \cong 2I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)$. Seeing that (17) is similar to our case, by the Pythagorean theorem:

$$2I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right) \cong \left\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_k\right\|_{\boldsymbol{J}}^2 \cong \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*\|_{\boldsymbol{J}}^2 + \left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_{\boldsymbol{J}}^2.$$

Therefore, for large n, the expectation of the KLI is a measure of the average estimation error:

$$\begin{aligned} 2nE\left[I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)\right] &\cong E\left[n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*\|_{\boldsymbol{J}}^2 + n\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_{\boldsymbol{J}}^2\right] \\ &= n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*\|_{\boldsymbol{J}}^2 + E\left[n\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_{\boldsymbol{J}}^2\right]. \end{aligned} \tag{18}$$

Eq. (18) shows that this measure can be decomposed into the bias plus the variance of $\left(\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right)$, first and second term on the right-hand side respectively. For large n, in the second term we have $n\left\|\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right\|_{\boldsymbol{J}}^2 = \left\|n^{1/2}\left(\boldsymbol{\theta}_k^* - \hat{\boldsymbol{\theta}}_k\right)^2\right\|_{\boldsymbol{J}}$, which distributes asymptotically as a χ_k^2. Denoting the non-random first component in the right-hand side with $\delta \equiv n\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*\right\|_{\boldsymbol{J}}^2$, and computing the expectation of the χ_k^2 random variable, Eq. (18) becomes, for large n:

$$2nE\left[I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)\right] \cong \delta + k, \tag{19}$$

The non-random quantity δ needs to be estimated in finite samples. Since: *(i)* the mean log likelihood is a consistent estimate of $I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)$; and *(ii)* the LR statistic,

$$LR(\mathbf{x}) = \eta_{k,K} = -2\sum_{i=1}^{n} \log \frac{f\left(x_i|\hat{\boldsymbol{\theta}}_k\right)}{f\left(x_i|\hat{\boldsymbol{\theta}}_K\right)}, \tag{20}$$

for the *i.i.d.* case under regularity conditions is asymptotically distributed as a noncentral $\chi_v^2(\delta)$ random variable, with $v = K - k$ degrees of freedom and noncentrality parameter δ; then the LR statistic in Eq. (20) is employed to estimate $I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)$.

Given that $E\left[\chi_v^2(\delta)\right] = \delta + v$ and $\eta_{k,K} \cong E\left[\chi_v^2(\delta)\right]$, we can solve for δ obtaining $\delta \cong \eta_{k,K} - v = \eta_{k,K} - (K - k)$. Thus, Eq. (18) becomes:

$$\begin{aligned} 2nE\left[I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)\right] &\cong \eta_{k,K} - (K - k) + k \\ &= \eta_{k,K} + 2k - K. \end{aligned} \tag{21}$$

From these arguments, if the KLI between the 'true' and the estimated parameter, $I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)$, is defined as the loss function in the MS problem, with its associated risk function $E\left[I\left(\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right)\right]$, from Eq. (21) we can get an estimate of the risk function for large n, and both K, and k relatively large integers, which is:

$$\mathcal{R}\left[\boldsymbol{\theta}^*, \hat{\boldsymbol{\theta}}_k\right] = n^{-1}\left(\eta_{k,K} + 2k - K\right). \tag{22}$$

Given that our goal is to find the $\hat{\boldsymbol{\theta}}_k$ that minimizes Eq. (22), or equivalently, that minimizes Eq. (21), we can focus on minimizing over $k = \{1, 2, \ldots, K\}$ the following quantity:

$$\xi_{k,K} = \eta_{k,K} + 2k = -2\sum_{i=1}^{n} \log \frac{f\left(x_i|\hat{\boldsymbol{\theta}}_k\right)}{f\left(x_i|\hat{\boldsymbol{\theta}}_K\right)} + 2k. \tag{23}$$

For hierchical/nested models, where the constant terms are shared by every model, the usual approximate estimate of the AIC is obtained:

$$AIC(k) = -2\sum_{i=1}^{n} \log f\left(x_i|\hat{\boldsymbol{\theta}}_k\right) + 2k. \tag{24}$$

□

2.4.1.2 *Discussion*

CORRECT USE OF THE AIC Akaike's criterion allows to judge a particular model, or judge ignorance about the structure of the model [52, p. 351]. From Sakamoto, Ishiguro and Kitagawa [253, pp. 83-85] we learn the following four points:

(i) The number of free parameters estimated from data should be less than $2\sqrt{n}$ ($\frac{n}{2}$ upper bound). Otherwise, if the number of free parameters is too large, the asymptotic normality of the MLE might fail.

(ii) The actual values of the AIC do not matter. Instead, the differences between the AIC obtained by one model versus the AIC of another model matter for MS. If the difference in absolute value is larger than 1 or 2, then it is considered to be informative. If this difference is much smaller than 1, then the goodness of fits of the model are almost the same. Even if the AIC of two models are very similar, if the distributions of the models are quite different, then it is reasonable to consider that neither of the models is good.

(iii) If the AIC gradually decreases with increasing order and it may not have a clear minimum, then it usually indicates that the parametrization is not appropriate.

(iv) AIC is not a criterion to estimate the true order of the model. Instead, it estimates the best fit model. The concept of true order is meaningless in the context of estimating the true distribution from a finite n.

From Burnham and Anderson [59, pp. 75, 80–89] we learn the following eleven points:

(i) AIC is a step in the Minimum A Information Criteria Estimate (MAICE) procedure [174, p. 69]. Later research considered also Akaike's weights and its relation to MS, which moves beyond the MAICE approach.

(ii) AIC cannot be used to compare models of different data sets (data must be fixed).

(iii) Order is not important in computing AIC values. This highlights a difference with respect to step-up (forward) and step-down (backward) hypotheses testing.

(iv) A common mistake is to mix response variables since all hypotheses have to be modelled using the same response variables.

(v) The AIC requires that error structures of comparing regression models to be the same.

(vi) In applied literature it is a common mistake to mix null hypothesis testing with information-theoretic criteria. Its is not adviced to use words such as *'significant'* or *'rejected'*. It is good practice to complement with evidence ratios, analysis of residuals, adjusted R^2, and other model diagnostics or descriptive statistics.

(vii) Null hypothesis testing is still important in strict experiments, but in observational studies it is not clear. Besides, often the hypotheses are naive or trivial.

(viii) Information-theoretic criteria are not a "test".

(ix) Remember always to perform exploratory data analysis.

(x) Ambivalence of data and multi-model inference. In relation to this last point, see Burnham and Anderson [60].

(xi) IC can be applied to non-nested models.

For an updated informative review of the AIC, see Cavanaugh and Neath [67].

THE AIC AND THE LR STATISTIC AIC departed from considerations on the Likelihood Ratio (LR) test statistic when the convergence to the χ^2-distribution fails, from the Kullback-Leibler Information (KLI) perspective. With hypothesis testing, a system of composite hypotheses H_0 and H_1 are considered. Under the null hypothesis, we use the asymptotic distribution of a test statistic in order to specify a rejection condition. Commonly used test statistics are Wilk's LR [337], Wald's test [319], and Rao's Lagrange-Multiplier (LM) test [233]. See Buse [61] for a clear description of these methods. Issues have been pointed out in the testing approach [32, 59, 93], which are connected with observational studies, and the difference between experiments and pseudo-experiments. For instance, with nested (hierarchical) hypotheses, their sequential order influences the result. Besides, there are issues with the distributional convergence as the number of hypotheses to test approaches sample's dimension.

The latter becomes relevant when applying hypothesis testing in the context of high-dimensional data. Benjamini and Hochberg [31] pointed three issues with the classical approach to multiple testing, which include that in practice test statistics are not multivariate normal and that are not comparisons of multiple treatments. This problem has been addressed in the literature departing from Bonferroni-types corrections. They proposed to control for the False Discovery Rate (FDR), i.e. the rate of type I error, instead of the Family Wise Error Rate (FWER) in multiple testing. A similar motivation leading to a different conclusion was advanced by Romano and Wolf [250] with a feasible computational method to control generalized FWER.

Discussions and extensions followed on asymptotic properties of IC, their asymptotically equivalent counterparts, and alternative criteria. We will briefly see two of these asymptotic properties, namely consistency and efficiency. These concepts are differently defined with respect to estimators, and play a central role in the study of MS via IC. Chapters 4 and 5 show that our multivariate extension of the MRIC displays asymptotic efficiency.

2.5 ASYMPTOTIC CONSISTENCY AND EFFICIENCY OF CRITERIA

Consider the set of candidate models $\mathcal{J}_l$, with the generic model l, $l = \{1, \ldots, K\}$, where K is the total number of candidate models. When the true model is among this candidates set (*model's correct specification*), if the criterion selects the true model with probability tending to one, then the selection method is *weakly consistent*. If this convergence is of the almost sure type, then it is said to be *strongly consistent*. If we do not assume that the true model is among the candidates (*model's misspecification*), then we can be interested in selecting the model which minimizes a particular measure or divergence. This leads us towards the concept of *efficiency*. In the MS literature, these two asymptotic properties deal with the ability of the criterion to select the 'true model' (consistency), or, the model that minimizes a particular measure (efficiency), as the sample size diverges.

Showing that this property is featured requires analytical proofs and numerical studies. It is worth noticing that importance is given to select a particular model over a set of candidate models. This is separated from the asymptotic properties of parameters' estimators. Confusion may arise, given that ultimately, MS can be deployed for both estimation and identification. The focus in our case is the latter. For an overview of these points, see Arlot and Celisse [24, pp. 46-48].

Besides, the setting of ID or FD also may deliver contrasting results. For instance, the AIC is not consistent but it is efficient in the ID case, while it is neither consistent nor efficient in the FD setting. Additionally, consider that the Bayesian Information Criterion (BIC) (introduced in Section 2.5.3.1) is consistent but it is not efficient in both the ID and FD settings. An additional layer of uncertainty is include when there is no information concerning the inclusion or exclusion of the 'true' model in the set of candidate models. See Hsu et al. [153, Tables 1 and 2] for a comparison of five criteria (AIC [7], BIC [258], Generalized AIC (GAIC) [173], Generalized BIC (GBIC) [200], GBIC_p[200], and their MRIC). The MRIC and our vectorial extension, the VMRIC, were developed in this last situation of possibly-misspecified models, i.e. no prior information on whether the 'true' model is included or not in the candidates set.

2.5.1 *Consistency*

Shao [266] developed an asymptotic theory for linear MS. Let $\mathbf{y}_n$ be an n-dimensional vector of independent responses, $\mathbf{X}_n$ be an $(n \times p_n)$ matrix of a p-dimensional regressor with n observations, and $\boldsymbol{\mu}_n = E[\mathbf{y}|\mathbf{X}_n]$ the mean response. This is estimated by the model α from class $\mathcal{A}_n$, via the LS estimator (LSE) $\hat{\boldsymbol{\mu}}_n(\alpha)$. Define $\hat{\alpha}_n$ as the selected model by an information criterion, where the subscript n indicates its dependence on the sample of n observations. Also, define α_n^L as the model which minimizes the following quadratic error loss function:

$$L_n(\alpha) = n^{-1} \left\| \boldsymbol{\mu_n} - \hat{\boldsymbol{\mu}}_n(\alpha) \right\|^2, \tag{25}$$

where $\|\cdot\|$ is the Euclidean norm. Under *i.i.d.* observations and misspecification, he conveniently defined asymptotic consistency in terms of the "best fitting" model:

Definition 5. *An information criterion is asymptotically consistent if*

$$\lim_{n\to\infty} \mathbb{P}\left\{\hat{\alpha}_n = \alpha_n^L\right\} \to 1. \tag{26}$$

Shao observed that this implied that the probability that both the quadratic error loss functions are equal converges asymptotically to one:

$$\lim_{n\to\infty} \mathbb{P}\left\{L_n(\hat{\alpha}_n) = L_n(\alpha_n^L)\right\} \to 1. \tag{27}$$

These two situations are equivalent if $L_n(\alpha)$ has a unique minimum for all large n.

Criteria's asymptotic consistency, in some sense, depends on the type of problem under consideration. In addition, the procedure for its demonstration will depend on the types of loss and risk functions [274], and on the sense of asymptotic consistency under consideration.

Now, a generalization of the AIC is introduced to state the regularity conditions for the asymptotic consistency of a criterion.

Definition 6. *A generalization of the approximate estimate of AIC [25, 41] is given by:*

$$AIC_\alpha = -2l(\hat{\boldsymbol{\theta}}) + \alpha p, \tag{28}$$

where α refers in this case to the weight applied to the penalty p, the number of estimated parameters in the model.

Shibata [278] argued that the following conditions are required for consistency using the above initial generalization of the AIC. For strong consistency he derived these results from the Law of Iterated Logarithms (LIL). See Hannan and Quinn [137].

Assumptions 2. *Under general regularity conditions requested for computing criteria, necessary and sufficient conditions for strong consistency of IC are found in setting $\alpha = \alpha_n$ in AIC_α, Eq. (28), such that:*

$$\liminf_n \frac{\alpha_n}{2\log\log n} > 1, \qquad \limsup_n \frac{\alpha_n}{n} = 0, \tag{29}$$

while for weak consistency such that:

$$\liminf_n \alpha_n = \infty, \qquad \limsup_n \frac{\alpha_n}{n} = 0. \tag{30}$$

These conditions are often seen in the assumptions preceeding the definition of IC. For instance, [153] requires that the penalty weight satisfies condition from Eq. (30), while the problem of its empirical determination is assessed in [154, Section S5]. In Chapters 4 and 5, our derivation of the vectorial version of the MRIC requests this condition for asymptotic results, viz. Eq. (242).

2.5.2 *Efficiency*

If the true model is not included in the set of candidate models, e.g. the parameter space has infinite dimension, then the concept of asymptotic consistency would be misleading. For this reason, *asymptotic efficiency* is defined as the convergence of the selected model to the one that minimizes the KLI or other specified loss function. Occasionally, this property is also defined as *optimality in the efficiency sense*,[11] e.g. [68, 189, 273–275], i.e. the selected model achieving the minimum value of the loss function (in probability). According to Claeskens and Hjort [83], a MS criterion is efficient if the ratio of the expected loss function computed at the selected model and the expected loss function at its theoretical minimizer converges in probability to one. Let us give further details.

As in Section 2.5.1, define $\hat{\alpha}_n$ as the selected model by a criterion with a sample of size n, and α_n^L the model which minimizes the loss function $L_n(\alpha)$ from Eq. (25). Shao [266], similarly to the definition of Li [189, p. 961] and in comparison to consistency, considered a weaker condition in which the selected model "$\hat{\alpha}_n$ *is asymptotically as efficient as* α_n^L *in terms of the loss* $L_n(\alpha)$". In that sense, it can be seen as the minimum requirement in terms of consistency:

Definition 7. *A criterion is asymptotically loss efficient if the ratio between the losses of models $\hat{\alpha}_n$ and α_n^L converges in probability to 1, i.e.*

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_n^L)} \xrightarrow{p} 1. \tag{31}$$

11 Often appears that the concept of optimality is used locally or considering a very restrictive sense, which may appear confusing if we take into consideration its meaning. It is adviced to check the specific sense of optimality to avoid confusions.

Shibata proposed both definitions of asymptotic mean efficiency [273] and approximate efficiency [275, p. 43]. Focus on the former and consider the $\mathcal{L}^2$ Hilbert space of sequences of real numbers. See Definition 32 for details. Denote the inner product $< \boldsymbol{a}, \boldsymbol{b} >= \boldsymbol{a}^\top \boldsymbol{b}$, with $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{L}^2$. Let $\mathbf{x}^\top = (x_1, x_2, \dots)$, $\mathbf{x} \in \mathcal{L}^2$, be the infinite dimensional regressors' vector composed by control variables; and $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \dots)$, $\boldsymbol{\beta} \in \mathcal{L}^2$, be the parameters' vector. The observational equation is $Y =< \mathbf{x}, \boldsymbol{\beta} > +\varepsilon$, where ε is a Gaussian error with zero mean and spherical error variance $\sigma^2 > 0$. Consider a sample of n independent observations for Y, $\boldsymbol{Y} = (Y^{(1)}, \dots, Y^{(n)})$, at $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$. In this case, we can estimate at most n parameters. Denote model $\boldsymbol{j} = (j_1, \dots, j_{k(j)})$ having regression function $f(\mathbf{x}, \boldsymbol{j}) =< \mathbf{x}, \boldsymbol{\beta}(\boldsymbol{j}) >$, where:

$$\boldsymbol{\beta}(\boldsymbol{j})^\top = \left(0, \dots, \beta_{j_1}, 0, \dots, \beta_{j_2}, 0, \dots, \beta_{k(j)}, 0, \dots\right),$$

is the restricted infinite dimensional parameter vector in the $\mathcal{V}(j)$ subspace, with $j_1 < j_2 < \cdots < j_{k(j)}$, $k(j) \geq 1$. Its LSE is $\hat{\boldsymbol{\beta}}(\boldsymbol{j}) = \left\{\hat{\beta}_{j_1}(\boldsymbol{j}), \dots, \hat{\beta}_{j_{k(j)}}(\boldsymbol{j})\right\}$ defined as the solution to the system:
$\boldsymbol{M}_n(\boldsymbol{j})\hat{\boldsymbol{\beta}}(\boldsymbol{j}) = \mathbf{X}(\boldsymbol{j})^\top \mathbf{y}$, where $\mathbf{y}$ is an n-dimensional column vector of observations,

$$\mathbf{X}(\boldsymbol{j}) = \{x_{i,j_l}, 1 \geq i \geq n, 1 \geq l \geq k(j)\}$$

is an $(n \times k(j))$ design matrix generated by vectors $\mathbf{x}^{(i)\top} = (x_{i,1}, x_{i,2}, \dots)$, with $i = \{1, \dots, n\}$, and $\boldsymbol{M}_n(\boldsymbol{j}) = \mathbf{X}(\boldsymbol{j})^\top \mathbf{X}(\boldsymbol{j})$ is a $(k(j) \times k(j))$ variance-covariance matrix. The corresponding LS predictor of a future observation at $x^{(i)}$ is: $\hat{Y}_i =< \mathbf{x}^{(i)}, \hat{\boldsymbol{\beta}}(\boldsymbol{j}) >$, with $i = \{1, \dots, n\}$.
Now, write the Residual Sum of Squares (RSS) as:

$$n\hat{\sigma}^2(\boldsymbol{j}) = \left\|\mathbf{y} - \mathbf{X}(\boldsymbol{j})\hat{\boldsymbol{\beta}}(\boldsymbol{j})\right\|^2,$$

where $\|\cdot\|$ is the Euclidean norm, and define $E\left[\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2 \left|\mathbf{x}^{(i)}\right.\right]$, with $i = \{1, \dots, n\}$, as the expectation of the sum of squared errors of the prediction conditional on future observations, which is then equal to:

$$E\left[\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 |x^{(i)}\right] = n\sigma^2 + \left\|\hat{\beta}(j) - \beta\right\|^2_{\boldsymbol{M}_n}, \tag{32}$$

where $\boldsymbol{M}_n = (\sum_{i=1}^n x_{i,l}x_{i,m}, 1 \geq l, m < \infty)$ is an infinite dimensional matrix and $\|\boldsymbol{a}\|_{\boldsymbol{M}_n} =< \boldsymbol{M}_n\boldsymbol{a}, \boldsymbol{a} >^{1/2}$ is a seminorm for any $\boldsymbol{a} \in \mathcal{L}^2$.

Definition 8. *A MS criterion is asymptotically optimal in the mean efficiency sense if it attains a lower bound for Eq. (32) when n tends to infinity.*

Recently, a definition of efficiency was proposed by Hsu et al. [153] via sequential data-driven methodology in the context of time series.[12]

12 We slightly deviate from the strong focus on the *i.i.d.* case to present an interesting recent definition on asymptotic efficiency of a model selection criterion. This choice is motivated by the intertwining of statistical advances in both settings, which traces back to early statistical analysis.

Let $\{y_t\}$ and $\{\mathbf{x}_t\}$ be two weakly stationary demeaned stochastic processes of dimensions 1 and m respectively. Let $y_{t+h} = \boldsymbol{\beta}_h^* \mathbf{x}_t^* + \boldsymbol{\epsilon}_{t+h}$ be the D.G.P for h-step ahead prediction, n be the sample size with

$$t = \{1, 2, \ldots, N, N+1, \ldots, N+h = n\}$$

, an h-step ahead possibly-misspecified forecasting model of the type: $y_{n+h} = \boldsymbol{\beta}_h \mathbf{x}_n + \boldsymbol{\varepsilon}_n^{(h)}$, where pseudo-true parameters' vector is given by

$$\boldsymbol{\beta}_h = \underset{\mathbf{C} \in \mathbb{R}^m}{\operatorname{argmin}} \, E\left[\left(y_{t+h} - \boldsymbol{c}^\top \mathbf{x}_t\right)^2\right],$$

and $\boldsymbol{\varepsilon}_n^{(h)}$ be the possibly-misspecified error for h-step ahead forecasting. The dependence of $\mathbf{x}_t$ on h exists, but it is suppressed for notational convenience, so the consideration of a specific regressor also depends on the forecast horizon h. The LS estimator delivers $\hat{y}_{n+h} = \hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n$, where

$$\hat{\boldsymbol{\beta}}_n = \left(\sum_{t=1}^{N} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{N} \mathbf{x}_t y_{t+h}$$

is the LS estimator of $\boldsymbol{\beta}_h$ for h-step ahead regression based on the sample of n observations. As measure of interest take the h-step ahead Mean-Squared Prediction Error (MSPE), which is derived from the difference between the observed and the estimated values, i.e. $\mathrm{MSPE}_h = E\left[\left(y_{t+h} - \hat{y}_{t+h}\right)^2\right]$. Theorem 2.1 in Hsu et al. [153] allows for the decomposition of the MSPE in two parts. The first one being the Misspecification Index (MI), which is linked to the goodness-of-fit of the model and is equal to the variance of the h-step ahead prediction error, i.e. $\mathrm{MI}_h = E\left[\varepsilon_{1,h}^2\right]$. The second component is the Variability Index (VI), which depends upon the variance of the h-step ahead predictor $\hat{y}_{n+h}$, and which is also connected to the estimation error of $\hat{\boldsymbol{\beta}}_n(h)$:

$$\mathrm{VI}_h = L_h = \operatorname{tr}\left\{\mathbf{R}^{-1}\mathbf{C}_{h,0}\right\} + 2\sum_{s=1}^{h-1} \operatorname{tr}\left\{\mathbf{R}^{-1}\mathbf{C}_{h,s}\right\}.$$

Here, $\mathbf{R} = E\left[\mathbf{x}_1 \mathbf{x}_1^\top\right]$ is the (non-singular) variance-covariance matrix of the regressors, whereas $\mathbf{C}_{h,s} = E\left[\mathbf{x}_1 \mathbf{x}_{1+s}^\top \varepsilon_{1,h} \varepsilon_{1+s,h}\right]$ represents the cross-covariance matrix between the regressors and the h-step ahead prediction error. The approach proposed by Hsu et al. [153] selects the model that minimises the h-step ahead MSPE_h. The minimization occurs by selecting the model with the smallest VI_h among those with the smallest MI_h, sequentially. We introduce here the main statement of the property. For further details, see Chapter 3 - Section 3.2.2.3, and Chapters 4 and 5.

Consider the set of candidate models $\mathcal{J}_k$, with generic model k, $k = \{1, \ldots, K\}$ and K the total number of candidate models. Define $\hat{l}$ to be the selected model by a criterion in a data-driven fashion for h-step ahead prediction.

Definition 9. *If the MS criterion selects model $\hat{l}$ such that it is the model with the smallest* VI *among those with the smallest* MI*, i.e.*

$$\lim_{n\to+\infty} P\left(\hat{l} \in M_2\right) = 1, \tag{33}$$

where

$$M_2 = \left\{k:\ k \in M_1,\ VI_h(k) = \min_{l\in M_1} VI_h(l)\right\}, \tag{34}$$

$$M_1 = \left\{k : 1 \leq k \leq K,\ \mathrm{MI}_h(k) = \min_{1\leq l\leq K} \mathrm{MI}_h(l)\right\}, \tag{35}$$

then it is asymptotically efficient.

2.5.3 *BIC, C_p, alternative criteria, and cross-validation*

Akaike's extension of the ML principle in relation to information-theoretic quantities for the solution of the model identification problem brought increasing interest. See Table 8 in Appendix A.1 for a list of proposed IC. It includes some IC and PC not discussed in detail here. These are:

i. Allen's Prediction Sum of Squares (PSS) [19];

ii. Amemiya's criterion [21];

iii. Bozdogan's Consistent AIC (CAIC), CAIC with Fisher information (CAICF), and Kashyap's Criterion (KC) [52] (the latter was originally proposed in Kashyap [168]);

iv. Rao and Wu's generalization $D_n(k)$ [235];

v. The Risk Inflation Criterion (RIC) by Foster and George [116];

vi. The Generalized BIC (GBIC) by Konishi and Kitagawa [173];

vii. The criterion proposed by Yang and Barron [349];

viii. The Generalized AIC (GAIC) and Generalized BIC (GBIC) of Lv and Liu [200];

ix. The Deviance Information Criterion (DIC) of Spiegelhalter et. al. [282, 283];

x. The criterion for elliptically symmetric distributions proposed by Boisbunon et al. [48];

xi. Watanabe's Widely Applicable Information Criterion (WAIC) and Widely applicable Bayesian Information Criterion (WBIC) [324–326].

Other criteria, such as the Focused Information Criterion (FIC) by Claeskens and Hjort [82] from the literature of model averaging, were not included, given that would go beyond the scope of the present chapter. Brief notes are included in Chapter 3, Section 3.2.4, in the context of ARMA models. We will give brief details on Schwarz's BIC, Mallows' C_p (Cp), and the general Cross-Validation (CV) procedure, given their importance in applied literature and their strong connection with IC here presented. We introduce Nishii's generalization of the BIC since it is useful for the discussion in Section 2.5.4 between point-wise and uniform convergence for MS criteria. Also, we introduce Shao's Generalized Information Criterion (GIC), since it will be useful in the discussion related to the CV method.

2.5.3.1 *The Bayesian approach*

The Bayesian approach to MS includes the seminal works of Kashyap [167], Akaike [10, 14], and Schwarz [258]. The latter defined the famous Bayesian Information Criterion (BIC), often named Schwarz's Information Criterion (SIC) in the econometric literature, asymptotically equivalent to Wei's Predictive Least Squares (PLS) [328] criterion.

Definition 10. *Schwarz's BIC(k) [258] approximate estimate is defined as:*

$$BIC(k) = -2l(\hat{\theta}_k) + k\log(n), \tag{36}$$

where k is model's dimension and n is the sample size.

The BIC is asymptotically consistent under correct specification in time series and regression models [214, 235, 328]. Under misspecification, the BIC is not asymptotically efficient [266, 271]. In the context of time series, for useful distinctions see Choi [79, pp. 58-66], or de Gooijer et al. [125, pp. 318-323]. Pericchi [221] proposed an introduction to the basics and reviewed different available methodologies. The rest of Pericchi's issue on the Handbook of Statistics is devoted to the Bayesian perspective. Further generalizations of the BIC include Nishii [214], Konishi and Kitagawa [173], Lv and Liu [200], and Watanabe [325], among others.

2.5.3.2 *Mallows' C_p*

In the context regression analysis with independent observations, Mallows [204] presented the famous Cp. This criterion may be considered a PC, since it is related to control model's errors. Let a sample of n observations on k fixed design variables and a single response variable, denoted by $x_0 = 1$, $\mathbf{x} = (x_0, x_1, \ldots, x_k)$ a $(1 \times k)$ vector, $\mathbf{y} = (y_1, \ldots, y_n)^\top$ a $(n \times 1)$ vector. Write $\mathbf{X} = (x_{u,i})$ the $(n \times (k+1))$ regressor matrix of rank $k+1$. Write the observational equation of the model $y_u = \eta(\mathbf{x}_u) + e_u$, with $u = \{1, 2, \ldots, n\}$, where $\eta(\mathbf{x}_u) = \beta_0 + \sum_{i=1}^{k} \beta_i x_{u,i} = \mathbf{x}_u\boldsymbol{\beta}$, and the residuals $\{e_u\}$ being independent

random variables with zero mean and unknown spherical variance $\sigma^2 > 0$. The goal is to estimate $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_k\right)$ in order to obtain a good estimate of function η evaluated at any point $\mathbf{x}$ in a neighbourhood of our data, $\eta(\mathbf{x})$: $\hat{y}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_i$. Specifically, we aim at a subset LS estimate of $\hat{\boldsymbol{\beta}}$ where some of its components are set to zero, while the rest are estimated via LS. Define the subset P of the full set of indices $K^+ = \{0, 1, 2, \ldots, k\}$, and Q the subset complementary to P. Assume the number of elements in each set P, Q are respetively $|P| = p$, $|Q| = p$, with $p + q = k + 1$. Denote by $\hat{\boldsymbol{\beta}}_P = \mathbf{X}_P^{-}\mathbf{y}$ the LS estimated vector of coefficients in P, while the rest are set to zero, where $\mathbf{X}_P^{-}$ is the Moore-Penrose generalized inverse of $\mathbf{X}_P$, with $\mathbf{X}_P$ is $\mathbf{X}$ setting the columns in Q to zeroes. Its associated residual sum of squares (RSS):

$$RSS_P = \sum_{u=1}^{n} \left(y_u - \mathbf{x}_u \hat{\boldsymbol{\beta}}_P\right)^2. \tag{37}$$

Definition 11. *Let $\hat{\sigma}^2$ be an estimate of σ^2. The approximate estimate of the Cp statistic for the selection of regression variables is defined as:*

$$C_p = \left(\hat{\sigma}^2\right)^{-1} (RSS_P) - n + 2p. \tag{38}$$

In the same article [204, pp. 662-663], Mallows proposed a version for multivariate response data of dimension w. Define $\hat{\boldsymbol{\Sigma}}$ as an estimate of the $(w \times w)$ residual covariance matrix. In this case the scalar RSS becomes the $(w \times w)$ matrix

$$\boldsymbol{RSS}_P = \sum_{u=1}^{n} \left(\boldsymbol{y}_u - \hat{\boldsymbol{\beta}}_P \boldsymbol{x}_u\right) \left(\boldsymbol{y}_u - \hat{\boldsymbol{\beta}}_P \boldsymbol{x}_u\right)^{\top}, \tag{39}$$

where $\boldsymbol{y}_u$ is a $(w \times 1)$ multivariate response vector, $\boldsymbol{x}_u$ is an $(m \times 1)$ regressors' vector, both observed at times $u = \{1, 2, \ldots, n\}$, $\hat{\boldsymbol{\beta}}_P$ the $(w \times m)$ matrix of estimated coefficients in P, and an $(w \times w)$ identity matrix $\boldsymbol{I}$. To obtain the size of $\boldsymbol{C}_p$, the trace operator, its largest eigenvalue, or any other suitable norm, was suggested.

Definition 12. *The approximate estimate of the Cp statistic for the selection of regression variables with multivariate response data is defined as:*

$$\boldsymbol{C}_p = \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{RSS}_P - (n - 2p)\boldsymbol{I}. \tag{40}$$

2.5.3.3 *Nishii's and Shao's Generalized IC*

NIHII'S GIC Nishii [214] studied the asymptotic distribution of various IC and PC. Let $\mathbf{y}$ be an $(n \times 1)$ observed vector, $\mathbf{X}$ a $(n \times K)$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^{\top}$ a vector of unknown parameters, and $\boldsymbol{e}$ a $(n \times 1)$ error vector assumed *i.i.d.* Gaussian with zero-vector mean and spherical error variance $\sigma^2 \boldsymbol{I}_N$, $\mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_N)$. Focus on MS for prediction. We denote a general model $j = \{j_1, \ldots, j_k\}$, with

$(1 \leq j_1 < \cdots < j_k \leq K)$ if and only if $\beta_{j_1} \neq 0, \ldots, \beta_{j_k} \neq 0$ while the remaining elements of $\boldsymbol{\beta}$ are set to zero. In this case, there are $k(j) = k + 1$ unknown parameters. Define $\boldsymbol{D}_j$ as a $(K \times k)$ matrix of zeros and ones such $\mathbf{X}\boldsymbol{D}_j$ only includes columns $j_1, \ldots, j_k$ of $\mathbf{X}$. For model j, the multiple regression model is: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}(j) + \boldsymbol{e}$, where $\boldsymbol{\beta}(j) = \boldsymbol{D}_j \boldsymbol{D}_j^\top \boldsymbol{\beta} = \boldsymbol{D}_j \left(\beta_{j_1}, \ldots, \beta_{j_k}\right)^\top$. Nishii assumed:

1. model's correct specification,
2. invertibility of $\mathbf{X}^\top \mathbf{X}$,
3. and the existence and invertibility of $\boldsymbol{M} = \lim_{N \Rightarrow \infty} N^{-1} \mathbf{X}^\top \mathbf{X}$ of the sample variance-covariance.

In the following definition, if $a_N = 2$, a version of the AIC is recovered, while if $a_N = \log N$, the BIC.

Definition 13. *Let $a_N > 0$ is a sequence such that Eq. (30) is satisfied. Then Nishii's generalization of the approximate estimate of BIC, the Generalized Information Criterion (GIC), is defined as:*

$$GIC = N \log \hat{\sigma}^2 + a_N k. \tag{41}$$

SHAO'S GIC As in Subsection 2.5.1, let $\mathbf{y}_n \equiv (y_1, \ldots, y_n)^\top$ be an n-dimensional vector of independent responses, $\mathbf{X}_n \equiv \left(\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top\right)^\top$ be an $(n \times p_n)$ matrix of a p_n-dimensional regressor with n observations (the i-th row, $\mathbf{x}_i$, is a p_n-dimensional vector of explanatory variables associated with y_i), and $\boldsymbol{\mu}_n = E\left[\mathbf{y} | \mathbf{X}_n\right]$ the mean response. The subscript n indicates dependence on the sample of n observations. Let α indicate a model from class $\mathcal{A}_n$. Define $\boldsymbol{e}_n = \mathbf{y}_n - \boldsymbol{\mu}_n$ as model's residual, with $\boldsymbol{e}_n \equiv \{e_1, \ldots, e_n\}$ assumed *i.i.d.* with conditional variance $\mathrm{V}\left[\boldsymbol{e}_n | \mathbf{X}_n\right] = \sigma^2 \boldsymbol{I}_n$, with $\boldsymbol{I}_n$ the n-dimensional identity matrix. Denote with $\|\cdot\|$ the Euclidean norm. Further technical assumptions are defined in [266, p. 224]. Now, let $S_n(\alpha) = \|\mathbf{y}_n - \boldsymbol{\mu}_n(\alpha)\|^2$, $\hat{\sigma}_n^2$ be an estimator of σ^2, and $\{\lambda_n\}$ a sequence of non-random such that $\{\lambda_n\} \geq 2$ and $\lambda_n / n \to 0$. Shao [266, p. 226] shows the criteria covered by the following generalization:[13]

Definition 14. *The approximate estimate of Shao's Generalized Information Criterion (GIC) is given by:*

$$\boldsymbol{\Gamma}_{n,\lambda_n}(\alpha) = n^{-1} S_n(\alpha) + n^{-1} \hat{\sigma}_n^2 p_n(\alpha). \tag{42}$$

2.5.3.4 *MS and Cross-validation*

CV is an established resampling (or data augmentation) method with applications to model selection, popular among the algorithmic and statistical community. Given the centrality of CV in MS, we present

13 See the concluding paragraph of Subsubsection 2.5.3.5.

its definition and how it relates to IC. For details see [24, 143], and the references therein. First, we follow [24], given that its framework allows to *"include most statistical frameworks"* [24, p. 43]. Then, [143, Ch. 7] is followed to understand the connection between IC and CV on the use of in-sample errors.

Denote the *i.i.d.* random variables $\xi_1, \ldots, \xi_n \in \Xi$ with shared distribution P. Let $s \in \mathbb{S}$ be a target feature to be estimated of the unknown distribution P, and denote with $t \in \mathbb{S}$ its approximation. The loss function $\mathcal{L}(t) : \mathbb{S} \mapsto \mathbb{R}$ is minimal for $t = s$. As we have seen, many of these loss functions are defined in the form of $\mathcal{L}_P(t) = E_{\xi \sim P}\left[\gamma\left(t; \xi\right)\right]$, where the subscript in the expectation operator denotes over the domain of the random variable ξ, and $\gamma : \mathbb{S} \times \Xi \mapsto [0, \infty]$ is a contrast function. We can interpret $E_{\xi \sim P}\left[\gamma\left(t; \xi\right)\right]$ as an average measure of discrepancy between t and a new observation ξ with distribution P. Given a loss function $\mathcal{L}_P(t)$, denote with $l(s,t) \equiv \mathcal{L}_P(t) - \mathcal{L}_P(s) \geq 0$, and with $E_{\xi_1,\ldots,\xi_n \sim P}\left[l(s, \hat{s}\left(\xi_1, \ldots, \xi_n\right))\right]$, respectively the excess loss and the risk of an estimator $\hat{s}\left(\xi_1, \ldots, \xi_n\right)$ of the target s.

Now, define a statistical algorithm $\mathcal{A}$ as any measurable mapping $\mathcal{A} : \cup_{n \in \mathbb{N}} \Xi^n \mapsto \mathbb{S}$. If we denote with $D_n = \left(\xi_i\right)_{1 \leq i \leq} \in \Xi^n$ a sample of size n, then the output of the statistical algorithm $\mathcal{A}$ is an estimator of s: $\mathcal{A}\left(D_n\right) = \left(\hat{s}^{\mathcal{A}}(D_n)\right) \in \mathbb{S}$. To asses the quality of the statistical algorithm $\mathcal{A}$, we use $\mathcal{L}_P\left(\hat{s}^{\mathcal{A}}(D_n)\right)$ and aim at its minimization. Denote with $\left(\hat{s}_\lambda\right)_{\lambda \in \Lambda}$ a family of candidate statistical algorithms, so that the *algorithm selection problem* can be phrased as choosing algorithm $\hat{\lambda}(D_n) \in \Lambda$ using data D_n. The final estimator of s is denoted by $\hat{s}_{\hat{\lambda}(D_n)}(D_n)$.

Consider the training set $I^{(t)} : \{1, \ldots, n\}$, such that both $I^{(t)}$ and its complement $\left(I^{(t)}\right)^c$ are non-empty. Define $\hat{\mathcal{L}}^{HO}\left(\mathcal{A}; D_n; I^{(t)}\right)$ as the *hold-out estimator* of the risk of $\mathcal{A}\left(D_n^{(t)}\right)$ with training set $I^{(t)}$,

$$\hat{\mathcal{L}}^{HO}\left(\mathcal{A}; D_n; I^{(t)}\right) \equiv n_v^{-1} \sum_{i \in D_n^{(v)}} \gamma\left(\mathcal{A}\left(D_n^{(t)}\right); \xi_i\right), \tag{43}$$

where $D_n^{(t)} \equiv \left(\xi_i\right)_{i \in I^{(t)}}$ is the *training* sample of size $n_t = Card(I^{(t)})$, $D_n^{(v)} \equiv \left(\xi_i\right)_{i \in I^{(v)}}$ is the *validation* sample of size $n_v = n - n_t$, and $I^{(v)}$ the *validation set*.

Geisser [121] described CV as averaging several hold-out estimators of the risk for different data splits. Let $\left\{I_1^{(t)}, \ldots, I_B^{(t)}\right\}$ be a sequence of subsets of the training set $I^{(t)}$, with $B \geq 1$.

Definition 15. *The CV estimator of the risk $\mathcal{A}\left(D_n\right)$, with different training sets $\left(I_j^{(t)}\right)_{1 \leq j \leq B}$ is defined as:*

$$\hat{\mathcal{L}}^{CV}\left(\mathcal{A}; D_n; \left(I_j^{(t)}\right)_{1 \leq j \leq B}\right) \equiv B^{-1} \sum_{j=1}^{B} \hat{\mathcal{L}}^{HO}\left(\mathcal{A}; D_n; I_j^{(t)}\right) \tag{44}$$

All common CV estimators are of the same form as Eq. (44), varying only in the particular definition of the splitting scheme. It is called "CV with averaging", since the estimates of the risk are averaged. An alternative definition is called "CV with voting" [346, 347]. See [143, Ch. 7], or [24, Sections 4-10] for further details.

2.5.3.5 *IC, CV and in-sample errors*

According to Hastie et al. [143, Ch. 7], the study of the performance of learning methods (a set of various statistical techniques) is related to its predictive capabilities. We follow them in the following paragraphs.

Let Y be a target variable , $\mathbf{X}$ be a vector of inputs, and $\hat{f}(\mathbf{X})$ be a forecasting model estimated from a *training set* $\mathcal{T}$, i.e. an independent test sample. Typical loss functions to measure the error between the observed Y and the predicted $\hat{f}(\mathbf{X})$ are, for instance:

$$L\left(Y, \hat{f}(\mathbf{X})\right) = \begin{cases} \left(Y - \hat{f}(\mathbf{X})\right)^2 & \text{(squared error)}, \\ \left|Y - \hat{f}(\mathbf{X})\right| & \text{(absolute error)}. \end{cases}$$

To evaluate model's performance, Hastie et al. underline that two measures are usually used, one conditional and the other unconditional:

(a) *Test Error*:

$$\mathrm{Err}_{\mathcal{T}} = E\left[L\left(Y, \hat{f}(\mathbf{X})\right) \middle| \mathcal{T}\right], \tag{45}$$

where $\mathbf{X}$ and Y are randomly drawn from the joint distribution in the population; $\mathcal{T}$ is fixed; and the Test Error refers to the specific training set $\mathcal{T}$.

(b) *Expected Prediction Error* (or *Expected Test Error*):

$$\mathrm{Err} = E\left[L\left(Y, \hat{f}(\mathbf{X})\right)\right] = E\left[\mathrm{Err}_{\mathcal{T}}\right], \tag{46}$$

where the expectation averages over all the randomness.

They indicate that, ideally, the goal is to estimate the *Test Error*, but that usually the *Expected Prediction Error* have efficient estimators. Its sample analogue, the *Training Error*, is defined as the average loss over the training sample:

$$\overline{\mathrm{err}} = N^{-1} \sum_{i=1}^{N} L\left(y_i, \hat{f}(x_i)\right). \tag{47}$$

The goal is to study the *Expected Test Error* of the estimated model. As model's complexity increases (i.e. the number of parameters), it also increases the use of the training data and allows for a better fit to more complicated structures. In other words, it creates a decrease in bias by increasing the variance. Let us see more in details this point, given that it is pivotal in much of the present work, e.g. Section 3.3.1 in Chapter 3; Theorem 1 in Chapter 4; Theorem 4 in Chapter 5.

Remark 2 ([143]). *In the context of regression, consider a model* $Y = f(X) + \epsilon$*, where* $E[\epsilon] = 0$*, and* $\text{Var}[\epsilon] = \sigma_\epsilon^2$*. If the squared-error loss is employed, the* Expected Prediction Error *of a prediction* $\hat{f}(\mathbf{X})$ *at* $\mathbf{X} = \mathbf{x}_0$ *is:*

$$\text{Err}(x_0) = E\left[\left(Y - \hat{f}(x_0)\right)^2 \middle| X = x_0\right] \tag{48}$$

$$= \sigma_\epsilon^2 + \text{Bias}^2\left(\hat{f}(x_0)\right) + \text{Var}\left(\hat{f}(x_0)\right), \tag{49}$$

where the first is the Irreducible Error, i.e. the variance of the target around its true mean $f(x_0)$*; the second is the squared bias:*

$$\text{Bias}^2\left(\hat{f}(x_0)\right) = \left(E\left[\hat{f}(x_0)\right] - f(x_0)\right)^2; \tag{50}$$

and the third is the variance, i.e. the expected squared deviation of $\hat{f}$ *from the mean:*

$$\text{Var}\left(\hat{f}(x_0)\right) = E\left[\left(\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right]. \tag{51}$$

A correct level of complexity should deliver the minimum *Expected Test Error* (Err). But there is an issue with the *Training Error* $(\overline{\text{err}})$: it decreases with model's complexity. The updated goal becomes to estimated correctly model's *Expected Test Error* (Err). Again, the main goal of the analysis is relevant:

- Model selection: choose the best model by studying the performance of different models;
- Model assessment: after selecting a final model, estimate its *Prediction Error* (*Generalization Error*) on new data.

In an ideal situation, we would have enough data to split the sample into:

(i) a Training set (for model fitting);

(ii) a Validation set (to estimated prediction error for MS);

(iii) a Testing set (to study the *Generalization Error*).

Generally, in a practical situation there is not enough data for this division. It is in these practical cases where analytical methods for Model Validation are deployed, such as information or prediction criteria, or those derived from the Minimum Description Length (MDL) literature. Another strategy is to use efficient sample re-use (as CV and bootstrap).

For the first strategy, consider the following quantities:

(i) *In-sample error*:

$$\text{Err}_{\text{in}} = N^{-1}\sum_{i=1}^{N} E_{Y^0}\left[L\left(Y_i^0, \hat{f}(x_i)\right)|\mathcal{T}\right], \tag{52}$$

and Y^0 indicating that at each training point $\mathbf{x}_i$, $i = \{1, 2, \ldots, N\}$, we observe new response values.

(ii) *Optimism*:

$$\text{op} \equiv \text{Err}_{\text{in}} - \overline{\text{err}}. \tag{53}$$

(iii) *Average optimism*:

$$w \equiv E_y(\text{op}) \tag{54}$$

where the expectation E_y is over the training sets.

To estimate in-sample Prediction Error, the first strategy (information and prediction criteria) estimates the *optimism* and adds the training error $\overline{\text{err}}$, i.e.

$$\widehat{\text{Err}}_{in} = \overline{\text{err}} + \hat{w}, \tag{55}$$

with $\hat{w}$ estimating the *average optimism*. In particular, the AIC uses the log likelihood function as loss function.

If we have a set of models $f_\alpha(\mathbf{x})$ indexed by α, then denote with $\overline{\text{err}}(\alpha)$ and $d(\alpha)$ both the *training error* and the number of parameters for each model α. Let $\hat{\sigma}_\epsilon^2$ be an estimate of the noise variance computed from the mean-squared error of a low-bias model (e.g. LS). Then we find model $\hat{alpha}$ by minimization of $AIC(\alpha)$, and select model $f_{\hat{alpha}}(\mathbf{x})$, where:

Definition 16. *The function $AIC(\alpha)$ estimates the test error curve:*

$$AIC(\alpha) = \overline{\text{err}}(\alpha) + 2\frac{d(\alpha)}{N}\hat{\sigma}_\epsilon^2. \tag{56}$$

CV directly estimates the expected out-of-sample error:

$$\text{Err} = E\left[L(Y, \hat{f}(X))\right], \tag{57}$$

the average *Generalization Error* when method $\hat{f}(X)$ is used on an independent test sample from the joint distribution of $(Y, \mathbf{X})$.

For K-fold CV, consider a data split into K parts. For the k-th part, we estimate the model using the remaining $K-1$ parts of data, and compute its prediction error. This procedure is repeated for $k = 1, 2, \ldots, K$, and then we combine the K estimates of the prediction error. Let $\kappa : \{1, \ldots, N\} \mapsto \{1, \ldots, K\}$ be an indexing function indicating to which partition belongs observation i by the randomization, and let $\hat{f}^{-k}(\mathbf{x})$ be the estimated function computed leaving the k-th part out. Then:

Definition 17. *The K-fold Cross-Validation (CV) estimate of the prediction error is given by:*

$$CV(\hat{f}) = N^{-1}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{-k(i)}\right). \tag{58}$$

If we have a set of models $f(\mathbf{x}), \alpha$ indexed by α, then denote with $\hat{f}^{-k}(\mathbf{x}, \alpha)$ the α-th model fitted removing the k-th part of the data. Then we find model $\hat{alpha}$ by minimization of $CV(\hat{f}, \alpha)$, specified in the following:

Definition 18. *The function $CV(\hat{f}, \alpha)$ estimates the test error curve:*

$$CV(\hat{f}, \alpha) = N^{-1} \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(\mathbf{x}_i, \alpha)\right). \tag{59}$$

Then we select model $f(\mathbf{x}, \hat{alpha})$, which will be then fitted to the full data.

From Arlot and Celisse [24], and as it is clear from the previous paragraphs, resampling-based techniques are a solution for the asymptotic nature of criteria, e.g. AIC, or for the dependence on data's assumptions, e.g. Cp. However, both strategies aim at a correct estimation of model's *Expected Test Error*. The first via in-sample errors, while the second generating artificial out-of-sample errors. The trade-off between these two solutions depends on the specific problem at hand. CV enjoys quasi-universality if data are effectively *i.i.d.*, but can be less accurate with respect to procedures with information or prediction criteria designed to be optimal if the assumptions hold (e.g. AIC, Cp are efficient and satisfy oracle inequalities), and its computational burden is higher. Furthermore, the connection between CV with both IC and PC has been studied [105], as well as their asymptotically equivalence, e.g. [264, 278, 286]. In particular, Stone [286] showed asymptotic equivalence between AIC and leave-one out CV, leading to understand the minimization of the AIC equivalent to minimize CV values. Shao [266, Theorem 4] showed the criteria inside each class share the same asymptotic behaviour:

(i) GIC_2, Cp, AIC, leave-one out CV, and generalized CV: useful for the case when no fixed-dimension correct model exist,

(ii) GIC_{λ_n} with $\lambda_n \to \infty$, and leave-d out CV with $d/n \to 1$: useful for the case when fixed-dimension correct model exist,

(iii) GIC_λ with fixed $\lambda > 2$, and leave-d out CV with $d/n \to \tau \in (0, 1)$: useful compromise between classes (i) and (ii).,

where GIC_2, GIC_{λ_n}, GIC_λ, refer to specific settings of $\mathbf{\Gamma}_{n,\lambda_n}(\alpha)$.

Appendix A.2.3 summarizes some results of CV and IC for nonlinear time series models and nonparametric regression. In time series, it is sometimes called *split-sample* validation [220]. See [24] for a comprehensive survey.

2.5.4 *Discussion*

ENTROPY AND INFORMATION Rissanen [241] citing Watanabe [323] noticed that both KLI and Shannon's theorem are corollaries of Gibbs' theorem, and that the AIC derives from those considerations. The BIC [258], which departed from an asymptotic expansion of the posterior probability, is a special case of Rissanen's MDL criterion, which considered Gibbs' theorem and coding theory. Josiah W. Gibbs (1839–1903) was influenced by Rudolf Clausius (1822–1888) ideas [172, p. 128], fact which positions this line of research as derived from the early thermodynamics and entropy studies. This field is still very active. For recent developments, see the workshop "Recent Advances in Info-Metrics Research" [313] and Chen et al. [73].

ON CONSISTENCY VERSUS EFFICIENCY While it is logical to find arguments in favour of the central role of obtaining asymptotically consistent estimators of model's dimension, it is important to remind the philosophical issue it highlights, i.e. the existence of a 'true' model when 'all models are wrong'. For early discussions (1989-2007), viz. [278, p. 229], [52, p. 357], [59], and [245, p. 1-2]. Consistency remains a desired property in contemporary applications, e.g. Markov models, Bayesian Networks. For instance see respectively the consistency of the BIC in Markov order estimation and partition Markov models [88, 89, 120], and the use of BIC and MDL as scoring functions for structure learning balancing precision and complexity of the model in Bayesian Networks [289]. Efficient approaches are preferred for empirical data in biology, social sciences, and medicine, while consistent approaches are preferred in the physical sciences and engineering [59, 269].

POINT-WISE AND UNIFORM CONVERGENCE Leeb, Pötscher and Ewald [188] discussed on asymptotic issues and implications for inference in relation to the difference between point-wise and uniform convergence in MS techniques by post-model-selection. The following definitions are due to Moise [209, Ch. 9], where further arguments on continuity and integrability of functions are treated. Let $f_1, f_2, \ldots$ and f be functions $f : A \to \mathbb{R}$, where $A \subset \mathbb{R}$.

Definition 19. *If for each $x \in A$ we have:*

$$\lim_{n \to +\infty} f_n(x) = f(x), \tag{60}$$

then the sequence $f_1, f_2, \ldots$ convergences point-wise to f.

Definition 20. *If for every $\epsilon > 0$ there is an $n_\epsilon \in \mathbb{Z}^+$ such that*

$$x \in A, n \geq n_\epsilon \Rightarrow |f_n(x) - f(x)| < \epsilon, \tag{61}$$

then the sequence $f_1, f_2, \ldots$ convergences uniformly to f.

POINT-WISE AND UNIFORM RESULTS IN IC To go beyond point-wise results, study of the asymptotic distribution of IC is needed. Nishi [214] obtained asymptotic distributions of the selected model by various criteria including: AIC, FPE [4], C_P [204], PSS [19], BIC [258], and its own GIC, and their quadratic risk functions, for regression problems. For further examples, cf. [79, 270, 344, 348]. In particular, Section 2.2 in Yang [348] for a discussion of point-wise consistency, efficiency, and the AIC-BIC dilemma. Possible proposed solutions to this quandary are found in adaptive MS, i.e. combining properties from both sides as in Wu [341], Hansen and Yu [140], Van Erven et al. [314], Ding et al. [99], where the former does it under the framework of Rissanen's MDL principle while the latter refers to autoregression in time series.

2.6 SMALL SAMPLES

The AIC and most IC were developed with the attempt to eliminate the asymptotic bias of the maximum likelihood for large samples with related asymptotic arguments.

Sugiura [290] proposed finite sample corrections by considering the exact bias for different situations. Consider a set of random samples indexed by $i = \{1, \ldots, k\}$, with k the total number of samples. The i-th random sample denoted by $\left\{x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}\right\}$ (where $x_1^{(i)}$ is the first observation from the i-th sample, while n_i being its sample size) comes from a Gaussian distribution $\mathcal{N}\left(\mu^{(i)}, \sigma^2\right)$, where $\mu^{(i)}$ refers to the mean of the i-th sample, and $\sigma^2 > 0$. Define parameters' vector as $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_k, \sigma^2)$, and the total sample size n equals the sum of all sample sizes: $n = \sum_{i=1}^{k} n_i$. In problems of data with different means, we are interested in testing multiple null-hypotheses with unknown common variance σ^2. Define c the total number of means we are interested in testing, with $1 \leq c \leq k$, to determine which samples have equal means. Denote by $\{j_1, \ldots, j_c\}$ the set of indexes for the test, e.g. μ_{j_1} refers to the first mean we are interested in testing, $\ldots$, μ_{j_c} to the c-th mean we want to test, with $1 \leq j_1 < \cdots < j_c \leq k$, i.e. we can consider testing up to k means. In this case, we write the general multiple null-hypothesis as:

$$H_0^{(j_1,\ldots,j_c)} : \mu_{j_1} = \cdots = \mu_{j_c},$$

for all possible values of c. Following similar mathematical arguments as in Section 2.3, developed originally by Akaike, Sugiura proposed the following definition, where the subscript c_1 stands for Criterion 1.

Definition 21. *The finite- and multi-sample correction for the approximate estimate of the AIC is:*

$$AIC_{c_1}(k, c) = -2l(\hat{\theta}) + \frac{2n(k - c + 2)}{n - k + c - 3}. \tag{62}$$

Then we compute the $AIC_{c_1}(k,c)$ for different combinations of hypotheses, and select the one with the smallest value.

Sugiura also considered small-sample correction and the exact bias for regression models with an almost identical result. Let $\mathbf{X} \in \mathbb{R}^n$ an n-dimensional Gaussian vector, $\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{A\theta}, \sigma^2 \boldsymbol{I}\right)$, with $\boldsymbol{A}$ an $(n \times k)$ known matrix with rank k, $\boldsymbol{\theta}$ a $(k \times 1)$ parameters' vector, and $\boldsymbol{I}$ an $(n \times n)$ unitary diagonal matrix. The null hypothesis is

$$\boldsymbol{H_0 : B\theta = 0},$$

with $\boldsymbol{B}$ a known $(b \times k)$ matrix of rank b, where b is the number of parameters set to zero (restrictions). The subscript c_{2_a} stands for Criterion 2, case a (univariate).

Definition 22. *The small-sample correction for regression models of the approximate estimate of the AIC is given by:*

$$AIC_{c_{2_a}} = -2l(\hat{\theta}) + \frac{2n(k-b+1)}{n-k+b-2}. \tag{63}$$

This result can be extended to the multivariate case. Consider an $(n \times p)$ matrix $\mathbf{X}$, with each row having a p-variate independent normal distribution , and $\mathbf{X} \sim \boldsymbol{\mathcal{N}(A\theta, \Sigma)}$, where the mean has an $(n \times k)$ matrix $\boldsymbol{A}$ of rank k, $\boldsymbol{\theta}$ is a $(k \times p)$ parameters' matrix, $\boldsymbol{\Sigma}$ is a $(p \times p)$ covariance matrix, and the hypothesis to study is again similarly $\boldsymbol{H_0 : B\theta = 0}$, with $\boldsymbol{B}$ a known $(b \times k)$ matrix of rank b. The subscript c_{2_a} stands for Criterion 2, case b (multivariate).

Definition 23. *The small-sample correction for multivariate regression of the approximate estimate of the AIC is given by:*

$$AIC_{c_{2b}} = -2l(\hat{\theta}) + \frac{2n(k-b+(p+1)/2)p}{n-k+b-p-1}. \tag{64}$$

For the case of normal populations with different variances, consider a sample $\left\{x_1^{(i)}, \ldots, x_{n_i}^{(i)}\right\}$ where the generic random variable $X \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$, with $i = \{1, \ldots, k\}$, and μ_i unknown. The hypothesis in this case is defined by:

$$H_0 : \sigma_{j_1}^2 = \cdots = \sigma_{j_c}^2,$$

where the set of indexes is defined as in the second paragraph of this section. The subscript c_{3_a} stands for Criterion 3, case a (univariate).

Definition 24. *The corrected approximate estimate AIC for normal populations with different variances is:*

$$AIC_{c3a} = -2l(\hat{\theta}) + 2\left[(c+1)\frac{\sum_{i=1}^{c} n_{j_i}}{\sum_{i=1}^{c} n_{j_i} - c - 2} + \left\{2\sum_{i=c+1}^{k} \frac{n_{j_i}}{n_{j_i}-3}\right\}\right]. \tag{65}$$

Its extension to the multivariate case considers the sample taken from a p-variate $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, with $\boldsymbol{\mu}_i$ an $(n \times p)$ matrix, and $\boldsymbol{\Sigma}_i$ a $(p \times p)$ matrix. Here the hypothesis to consider with unknown μ_i:

$$H_0 : \Sigma_{j_1} = \cdots = \Sigma_{j_c},$$

where again the set of indexes is as previously explained. The subscript in the following approximate criterion, c_{3_a}, stands for Criterion 3, case b (multivariate).

Definition 25. *The approximate estimate of the AIC for multivariate response with normal populations with different variances is:*

$$\begin{aligned} AIC_{c_{3}b} = \\ -2l(\hat{\theta}) + 2p \Bigg[\left(c + \frac{p+1}{2}\right) \frac{\sum_{i=1}^{c} n_{j_i}}{\sum_{i=1}^{c}(n_{j_i} - c - p - 1)} \\ + \left\{ \frac{p+3}{2} \sum_{i=c+1}^{k} \frac{n_{j_i}}{n_{j_i} - p - 2} \right\} \Bigg]. \end{aligned} \tag{66}$$

Extending Sugiura's work, Hurvich and Tsai [157] proposed a second-order bias adjustment for regression models, which includes nonlinear regression and autoregressive time series models.[14] It is efficient if the true parameter space is infinite dimensional. In this situation, the concept of approximating family is employed analogously to candidate models. Consider a DGP $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, with $\mathbf{y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\varepsilon}$ of dimension n, where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_0^2)$. The candidate family of models is $\mathbf{y} = h(\boldsymbol{\theta}) + \boldsymbol{u}$, where $\boldsymbol{\theta}$ is an m-dimensional parameter vector, both $\boldsymbol{u}$ and $h(\boldsymbol{\theta})$ are n-dimensional vectors, h is a twice continuously differentiable function in $\boldsymbol{\theta}$, and $u_i \overset{iid}{\sim} N(0, \sigma^2)$. In the following definition, the subscript c stands for "corrected". Further, following [273] it can be shown that AIC_c is asymptotically efficient if approximating models are linear.

Definition 26. *If the approximating family of models includes the DGP, it can be shown that:*

$$AIC_c = n \log \hat{\sigma}^2 + n \frac{1 + m/n}{1 - (m+2)/n} \tag{67}$$

is an approximately unbiased estimator of the expected KLI of the fitted model.

Bedrick and Tsai [30] extended this small sample correction to multiresponse regression models. Consider a collection of p variables with n observations, $\boldsymbol{Y}$, an $(n \times p)$ matrix. Let $\boldsymbol{X}$ be an $(n \times m)$ known matrix of regressors, $\boldsymbol{B}$ an $(m \times p)$ matrix of unknown parameters, and $\boldsymbol{U}$ an $(n \times p)$ matrix with *i.i.d.* errors with $\boldsymbol{U}_i \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma})$. Consider the multivariate regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}$, and the DGP

14 See also Section 3.2.

$\boldsymbol{Y} = \boldsymbol{X}_0\boldsymbol{B}_0 + \boldsymbol{U}_0$, where $\boldsymbol{X}_0$ is $(n \times m_0)$, $\boldsymbol{B}_0$ is $(m_0 \times p)$, and $\boldsymbol{U}_0$ has $\boldsymbol{U}_{0,i} \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}_0)$. To obtain the following definition, the arguments are similar to Section 2.3. Note that this criterion is an exact unbiased estimator of the expected value, under correct specification, of the KLI between the true and fitted models, using the MLE.

Definition 27. *The small sample correction for multivariate regression with correct specification of the approximate estimate of the AIC is equal to:*

$$\mathrm{AIC}_C = n \log \left|\hat{\boldsymbol{\Sigma}}\right| + dp(n+m), \tag{68}$$

where $d = n/\left(n - (m+p+1)\right)$.

Seghouane and Bekara [261] proposed the Corrected Kullback Information Criterion (KICc), a bias corrected version of the Kullback Information Criterion (KIC) by Cavanaugh [66], which is an asymptotically unbiased estimator of Jeffrey's J-divergence. Following arguments similar to Hurvich and Tsai, their focus was on linear regression models under correct specification or overfitting. Let

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{I}_n), \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_k^2 \boldsymbol{I}_n), \end{aligned}$$

be the DGP and the k-th candidate, where $\mathbf{y}$ is an n-dimensional vector of observations, $\mathbf{X}$ is an $(n \times k)$ design matrix of rank k, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_k$ are k-dimensional parameters' vector, and both $\boldsymbol{\varepsilon}$ and $\boldsymbol{\epsilon}$ are n-dimensional noise vectors. In this case, the full parameters' vector is denoted by $\boldsymbol{\theta}_k = \left[\boldsymbol{\beta}_k^\top \ \sigma_k^2\right]^\top$.

Definition 28. *The approximate estimate of the KICc is given by:*

$$KICc = -2l(\hat{\theta}) + 2\frac{(k+1)n}{n-k-2} - n\psi\left(\frac{n-k}{2}\right) + n\ln\frac{n}{2}, \tag{69}$$

where $\psi(\cdot)$ *is the psi (digamma) function.*

Leeb [186] studied the *out-of-sample* predictive performance of different IC and PC (including AIC and AIC_c) when the sample size is small relative to data's complexity. Two situations were considered on the latter:

(i) the number of parameter and sample size are of the same order; and

(ii) the number of candidate models is larger than sample size.

The setting is that of *random design regression with an infinite-dimensional model*, and focusses on prediction of new observations given unobserved regressor. Let y be a scalar response variable, related to a sequence of explanatory variables $\mathbf{x} = (x_j)_{j=1}^{\infty}$ such that:

$$y = \sum_{j=1}^{\infty} x_j \beta_j + u, \tag{70}$$

with the $\boldsymbol{\beta}$ parameters such that $\boldsymbol{\beta} = (\beta_j)_{j=1}^{\infty}$, the error term u with $E[u] = 0$, $E[u^2] = \sigma^2 \geq 0$, the random sequence of explanatory multiple regressor $\mathbf{x}$ with $E[\mathbf{x}] = \mathbf{0}$, and variance-covariance net $\boldsymbol{\Sigma} = [E[x_i x_j]]_{i,j\geq 1}$ such that (70) converges in $\mathcal{L}^2$. Furthermore, it is assumed that $E[x_{j_k} u] = 0$, for each $k \geq 1$, and for integers $j_1 < j_2 < \cdots < j_k$. Consider a sample $(\boldsymbol{Y}, \boldsymbol{X})$ from (70), with $\boldsymbol{Y} = (y^{(1)}, \ldots, y^{(n)})^\top$ an n-dimensional vector, $\boldsymbol{X} = (\mathbf{x}^{(1)\top}, \ldots, \mathbf{x}^{(n)\top})^\top$ an $(n \times \infty)$-dimensional net, and $(y^{(i)}, \mathbf{x}^{(i)})$ *i.i.d.* copies of $(y, \mathbf{x})$. Parameters' vector $\boldsymbol{\beta}$ is estimated via the restricted least-squares (RLS) estimator considering submodels of (70). Define $\tilde{\boldsymbol{\beta}}(m)$ the RLS estimator referred to model m, of dimension $|m|$. The goal is to select the model with 'best' out-of-sample prediction. Consider a new set of observations $(y^{(f)}, \mathbf{x}^{(f)})$ independent of the sample $(\boldsymbol{Y}, \boldsymbol{X})$. Let model m of dimension $|m| < n - 1$, its RLS estimator $\tilde{\boldsymbol{\beta}}(m)$, and the predictor $x^{(f)\top}\tilde{\boldsymbol{\beta}}(m)$. The evaluation of predictor's performance is based on the conditional Mean-Squared Prediction Error (MSPE):

$$\mathrm{MSPE}_{\mathrm{C}}(m) = E\left[(y^{(f)} - x^{(f)\top}\tilde{\boldsymbol{\beta}}(m))^2 \middle| \boldsymbol{Y}, \boldsymbol{X}\right],$$

where the expectation depends on $n, \boldsymbol{\beta}, \sigma$, and $\boldsymbol{\Sigma}$, and the subscript C stands for "conditional". Let $\mathrm{RSS}(m)$ denote the Residual Sum of Squares (RSS) for model m.

Definition 29. *As approximate estimates for the* $\mathrm{MSPE}_{\mathrm{C}}(m)$*, Leeb considered the Generalized CV [87], GMV(m), the* $S_p(m)$ *criterion [312], and an auxiliary criterion* $\hat{\rho}^2(m)$*. Their approximate estimates are:*

$$\mathrm{GCV}(m) = \frac{\mathrm{RSS}(m)}{n - |m|}\frac{n}{n - |m|}, \tag{71}$$

$$\mathrm{S}_p(m) = \frac{\mathrm{RSS}(m)}{n - |m|}\frac{n-1}{n - 1 - |m|}, \tag{72}$$

$$\hat{\rho}^2(m) = \frac{\mathrm{RSS}(m)}{n - |m|}\frac{n+1}{n + 1 - |m|}, \tag{73}$$

Via theoretical and simulation results, Leeb showed that $\mathrm{GCV}(m)$, $\mathrm{S}_p(m)$, and $\hat{\rho}^2(m)$ perform better than the AIC and AIC_c in terms of approximating the $\mathrm{MSPE}_{\mathrm{C}}(m)$ in this setting.

2.6.1 *Discussion*

GOODNESS-OF-FIT IN TERMS OF THE LIKELIHOOD FUNCTION OR THE RESIDUAL SUM-OF-SQUARES IC were developed considering the connection between entropy and the MLE. Further adaptations included approximate/conditional [12], penalized [173, 278, 280], or partial / composite [315] likelihood functions. When the estimation of parameters was related to regression problems, the Ordinary least squares (OLS) method was considered. It is well known in this type of problems, with Gaussian errors, the OLS estimator and MLE coincide under spherical errors. See Berkson [33] for a discussion. When working

directly with the likelihood function, IC are written in terms of the log-likelihood function, $l(\cdot)$. For regression problems, the RSS is used instead.

MLE, ROBUST ESTIMATION, IC AND REGULARIZATION MLEs are a special case of robust estimation [156], and can be thought as as an $\mathcal{L}_1$ or $\mathcal{L}_2$ regularization problem where the penalty parameter, λ, is set to zero. For details, Schmidt [255]. The link between IC and regularization is explored by Dixon and Ward [102], while it is taken as departing point in the machine learning literature, e.g. Xu et al. [342], Giraud [124, Ch. 2].

2.7 MISSPECIFICATION AND FURTHER ESTIMATORS

The AIC was developed under model's correct specification. To obtain asymptotic consistency, efficiency and normality properties of MLEs for misspecified probabilistic models, we need to modify the classical regularity conditions, as in Huber [155] or White [330]. We follow Konishi and Kitagawa [174, pp. 47-50] for these relaxed conditions.

Assumptions 3. *To obtain the asymptotic distribution of parameters under misspecification, assume the following for the probability density function* $f(x|\boldsymbol{\theta})$*:*

(i) $\log f(x|\boldsymbol{\theta})$ *is three times continuously differentiable with respect to the parameters' vector* $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)^\top$*, i.e. these derivatives are continuous functions;*

(ii) There exist integrable functions $F_1(x), F_2(x) \in \mathbb{R}$*, and function* $H(x)$*, usually referred to as envelope functions, such that:*[15]

$$\int_\Omega H(x) f(x;\boldsymbol{\theta}) \mathrm{d}x < \infty, \tag{74}$$

i.e. the expected value of function $H(x)$ *with respect to the postulated law* $f(\cdot)$ *is finite; and*

$$\left| \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_i} \right| < F_1(x),\ i = \{1, \ldots, p\},\ \forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{75}$$

$$\left| \frac{\partial^2 \log f(x;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| < F_2(x),\ i, j = \{1, \ldots, p\},\ \forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{76}$$

$$\left| \frac{\partial^3 \log f(x;\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| < H(x),\ i, j, k = \{1, \ldots, p\},\ \forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta}; \tag{77}$$

(iii) For some $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*,* $i, j = \{1, \ldots, p\}$*:*

$$0 < \int_{-\infty}^{+\infty} \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_j} f(x;\boldsymbol{\theta}) \mathrm{d}x < +\infty, \tag{78}$$

15 This condition may be further relaxed, without the second and higher derivates of the likelihood function [155].

i.e. the expected Fisher information on $\boldsymbol{\theta}$ of a single generic observation is assumed to be positive and finite for a p-dimensional parametric model, $0 < I_1(\boldsymbol{\theta}) < +\infty$, where:

$$I_1(\boldsymbol{\theta}) = \int \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} g(x)\mathrm{d}x \tag{79}$$

$$= E_G\left[\frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right]; \tag{80}$$

(iv) Denote $d_n = \{x_1, \ldots, x_n\}$ an i.i.d. *random sample extracted from an unknown model characterized by a probability density function $g(x;\boldsymbol{\theta})$ which is* not necessarily equal *to the probability density function $f(x;\boldsymbol{\theta})$;*

(v) Denote with $\boldsymbol{\theta}_0$ the solution to the set of equations:

$$\int_\Omega \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \theta_i} g(x)\mathrm{d}x = \mathbf{0},\ i = \{1, \ldots, p\}, \tag{81}$$

where the order of differentiation and integration may be interchanged if we are in a 'regular estimation problem', which would allow us to view the left-hand side as the Expected Value, with respect to the probability density function $g(\cdot)$, of the first partial derivative of the logarithm of the probability density function $f(\cdot)$ with respect to the parameters' vector $\boldsymbol{\theta}$:

$$E_G\left[\frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \frac{\partial}{\partial \boldsymbol{\theta}} E_G\left[\log f(x;\boldsymbol{\theta})\right], \tag{82}$$

i.e. $\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}$ allows to maximize the Expected Log Likelihood, or equivalently, minimize the KLI of the model characterized by $f(x;\boldsymbol{\theta})$, with respect to the model characterized by $g(x;\boldsymbol{\theta})$.

Let $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$ be a p-dimensional parameters' vector, and its MLE computed with a sample of dimension n, respectively, and $\boldsymbol{I}(\boldsymbol{\theta_0})$ and $\boldsymbol{J}(\boldsymbol{\theta_0})$ be $(p \times p)$ matrices, respectively the squared gradient and minus the Hessian, computed at $\boldsymbol{\theta} = \boldsymbol{\theta_0}$, equal to:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \int \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} g(x)dx,$$

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\int \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} g(x)dx.$$

The following proposition states that the the MLE $\hat{\boldsymbol{\theta}}_n$, using a sample of dimension n, is a weakly consistent estimator of $\boldsymbol{\theta}_0$, and that the probability distribution of $\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ converges to a p-dimensional Gaussian distribution with a zero-valued mean vector and $(p \times p)$ asymptotic variance-covariance matrix computed at $\boldsymbol{\theta}_0$.

Proposition 2. *Under assumptions 3, it can be shown that asymptotically:*

$$\hat{\boldsymbol{\theta}}_n \to_p \boldsymbol{\theta}_0, \tag{83}$$

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \to_d \mathcal{N}_p\left(\mathbf{0}, \left[\boldsymbol{J}^{-1}(\boldsymbol{\theta_0})\right] I(\boldsymbol{\theta_0}) \left[\boldsymbol{J}^{-1}(\boldsymbol{\theta_0})\right]^T\right). \tag{84}$$

TYPES OF MISSPECIFICATION There are different types of model misspecification. For instance, local misspecification as in Schorfheide [256]. The properties of predictors in misspecified AR time series models have been studied by Kunitomo and Yamamoto [178], the consequence of misspecified models on parameter estimation in Long [196], a general definition of model misspecification was considered by Hsu et al. [153] for parametric time series, and the consequences of misspecified models in the quasi ML (QML) estimation framework for common time series models by Bardet et al. [28].

CONSEQUENCES OF MISSPECIFICATION FOR IC Model's misspecification modifies IC's evaluation with respect to the correctly specified case. If we are not sure whether the model is correctly nor misspecified (i.e. possibly-misspecified setting), our conclusions on asymptotic properties of IC are further modified. Related considerations from statistical decision theory include the admissibility of the selection procedure [284, 291] and the unbiasedness of the estimator of the loss function (e.g. [48]).

2.7.1 *TIC and RIC*

To the best of our knowledge, the first generalization considering the bias of the estimation of the expected log likelihood in the possibly-misspecified case was the Takeuchi Information Criterion (TIC) [292]. This work was written in Japanese, but Shibata [278] latter proposed an extension of the TIC, Shibata's Regularized Information Criterion (SRIC), where he also discussed the TIC with examples. Both criteria were latter presented in Burnham and Anderson [59], and were further extended .

For model comparison and under four regularity assumptions similar to those above, Shibata [278] writes the expected KLI, between the unknown $g(\cdot)$ and the $f(\cdot)$ at the estimated parameter:

$$E\left[I(g; f(x|\hat{\boldsymbol{\theta}}_k))\right] = \\ \int g(x) \log g(x)\, \mathrm{d}x + E\left[-l(\hat{\boldsymbol{\theta}}_k)\right] + \mathrm{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\} \\ + o(1).$$

Given that the first component does not depend on any model, it is considered as a constant and omitted. This decomposes the expected KLI as minus the expected log likelihood at the estimated parameter plus a bias term. Different estimates of the bias term $\mathrm{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\}$ deliver different criteria. If $g(\cdot)$ is equal to $f(\cdot)$, the AIC is obtained.

Let $\mathbf{y}_n$ be a vector of n independent observations, and write the joint likelihood as:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i\left(y_i; \boldsymbol{\theta}\right), \tag{85}$$

and define $l_i\left(\hat{\boldsymbol{\theta}}\right) \equiv \log f_i\left(y_i; \hat{\boldsymbol{\theta}}\right)$. In the following definition, notice that the estimator of the bias, $\operatorname{tr}\left\{\hat{I}\hat{J}^{-1}\right\}$, coincides with the LM test statistic [151], where

$$\hat{I} = \sum_{i=1}^{n} \frac{\partial l_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \frac{\partial l_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^\top}, \tag{86}$$

$$\hat{J} = -H(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}, \tag{87}$$

are consistent estimators of $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ respectively. Then the approximate estimate of the TIC is given by minus two times the log maximum likelihood, plus twice the LM test statistic:

Definition 30. *The TIC is defined as a generalization of the AIC robust to model misspecification:*

$$TIC = 2E\left[I(g; f(x|\hat{\boldsymbol{\theta}}_k))\right] \tag{88}$$

$$= 2\left(E\left[-l(\hat{\boldsymbol{\theta}}_k)\right] + \operatorname{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\}\right), \tag{89}$$

and its approximate estimate is given by:

$$TIC(k) = -2l(\hat{\boldsymbol{\theta}}_k) + 2\operatorname{tr}\left\{\hat{I}\hat{J}^{-1}\right\}. \tag{90}$$

To obtain SRIC, Shibata considers the Maximum Penalized Likelihood (MPL) function,

$$l_\lambda(y; \boldsymbol{\theta}) = \log f(y; \boldsymbol{\theta}) + \lambda k(\boldsymbol{\theta}),$$

$$l_\lambda(\mathbf{y}_n; \boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{\log f(y_i; \boldsymbol{\theta}) + \lambda k_i(\boldsymbol{\theta})\right\},$$

for a single observation and the whole sample respectively, with $k(\boldsymbol{\theta}) \leq 0$ a twice differentiable penalty function which may depend on n, and $\lambda \geq 0$ a weight controlling the amount of penalty. The MPL estimate (MPLE), $\hat{\boldsymbol{\theta}}(\lambda)$, is obtained by maximizing the MPL with respect to $\boldsymbol{\theta}$. Similarly to Condition v in Assumptions 3, it is assumed here that $\hat{\boldsymbol{\theta}}(\lambda)$ converges to $\boldsymbol{\theta}^*(\lambda)$ defined as the unique solution to the following set of equations:

$$E\left[\frac{\partial}{\partial \boldsymbol{\theta}} l_\lambda(y; \boldsymbol{\theta})\right] = \mathbf{0},$$

i.e. $\boldsymbol{\theta}^*(\lambda)$ is a point in the parameters' space that allows to maximize the expected penalized log-likelihood (equivalent to minimized the KLI).

Let $l(\hat{\boldsymbol{\theta}}_k(\lambda))$ be the MPL estimator for a model of dimension k, and

$$\hat{I}(\lambda) = \sum_{i=1} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} l_\lambda(y_i; \hat{\boldsymbol{\theta}}(\lambda) \frac{\partial}{\partial \boldsymbol{\theta}^\top} l_\lambda(y_i; \hat{\boldsymbol{\theta}}(\lambda))) \right\},$$

$$\hat{J}(\lambda) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} l_\lambda(\mathbf{y}_n; \hat{\boldsymbol{\theta}}(\lambda)),$$

be the estimates of both the squared gradient and minus the Hessian matrix respectively, evaluated at the MPLE in both cases, but avoided in the notation for simplicity. Extending the TIC as a regularization criterion, Shibata [278] obtained:

Definition 31. *The approximate estimate of* SRIC *is given by:*

$$RIC = -2l(\hat{\boldsymbol{\theta}}_k(\lambda)) + 2\mathrm{tr}\left\{ \hat{I}(\lambda) \hat{J}(\lambda)^{-1} \right\}. \tag{91}$$

In this case, with $\lambda = 0$ we obtain again the TIC. The advice is to choose λ as to minimize SRIC for each model, and then compare the minimized value SRIC between the models.

2.7.2 *GIC and GAIC for functional estimators*

The following definitions deliver some preliminary concepts needed to continue the study of IC under model's misspecification for functional estimators.

Definition 32 (Banach space, Hilbert space, and Functional)**.** *A normed vector space V with the property that each Cauchy sequence $\{\boldsymbol{v}_k\}_{k=1}^{\infty}$ in V converges toward some $\boldsymbol{v} \in V$, is called a Banach space. A vector space with an inner product $\langle \cdot, \cdot \rangle$, which is a Banach space with respect to the norm in inner product space, i.e.*

$$\|\boldsymbol{v}\| := \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle}, \boldsymbol{v} \in V,$$

is a Hilbert space. Now, let $\mathcal{H}$ be a Hilbert space. A linear operator $\Phi : \mathcal{H} \to \mathbb{C}$ is called a functional [81, p. 70].

Definition 33 (Functional estimator [174])**.** *Assume that the parameter θ is defined as a real-valued function of the distribution G, i.e. the functional $T(G)$, with $T(G)$ a real-valued function defined on the set of all distributions on the sample space and does not depend on the sample size n. Then, given the observations $\{x_1, \ldots, x_n\}$, the estimator $\hat{\theta}$ of θ is a functional estimator:*

$$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) = T(\hat{G})$$

with $\hat{G}$ is the Empirical Distribution Function, $\hat{G}(x) = \frac{1}{n} \sum_{\alpha=1}^{n} I(x; x_\alpha)$, where $I(x; x_\alpha)$ is the indicator function equal to 1 if $x \geq x_\alpha$, 0 if $x < x_\alpha$.

Definition 34 (Fisher Consistency [156])**.** *Let $(x_1, \ldots, x_n)$ be a sample of n observations, $F_n = n^{-1} \sum \delta_{x_i}$ be the empirical measure with $\delta_x = 1$ at*

x, $T_n(x_1, \ldots, x_n) = T(F_n)$ be some functional T defined on the space of empirical measures,[16] and F the true underlying common distribution of the observations. Then, if the functional T satisfies:

$$T(F) = \lim_{n\to\infty} T(F_n), \tag{92}$$

then it is called Fisher consistent at F.

Cox and Hinkley [86] note for instance that with discrete random variables, Fisher consistency coincides with requiring that, if all sample proportions are equal to the corresponding probabilities, then the estimate is exactly correct.

Let $\mathbf{X}_n$ be a random sample size n from an unknown distribution $G(x)$ with density function $g(x)$. We estimate using a parametric family of density functions, $\{f(x|\theta); \theta \in \Theta)\}$, possibly misspecified, with $\boldsymbol{\theta}$ a p-dimensional vector of unknown parameters. Konishi and Kitagawa [173] considered the bias correction of the log likelihood when $\hat{\boldsymbol{\theta}}$ for functional estimators and misspecified models under the assumption of Fisher consistency. This delivered the Generalized Information Criterion (GIC). In this case, the expected log likelihood is:

$$\eta(\mathbf{X}_n; G) \equiv \int g(z) \log f(z|\hat{\theta}) \, \mathrm{d}z = \int \log f(z|\hat{\theta}) \, \mathrm{d}G(z)$$

and its estimator uses the empirical distribution of G, i.e.

$$\eta(\mathbf{X}_n; \hat{G}) \equiv \frac{1}{n} \sum_{\alpha=1}^{n} \log f(X_\alpha|\hat{\theta}).$$

Theorem 2.1 [173] shows that if the functional $\boldsymbol{T}(\cdot)$ is second-order compact differentiable at the distribution G, i.e. suitably defined p-dimensional regular functional, then the asymptotic bias of the log likelihood in the estimation of the expected log likelihood can be written as:

$$E_G\left\{\eta(\mathbf{X}_n; \hat{G}) - \eta(\mathbf{X}_n; G)\right\} = \frac{1}{n} b_1(G) + o(\frac{1}{n})$$

where

$$b_1(G) = \mathrm{tr}\left\{\int \boldsymbol{T}^{(1)}(z; G) \frac{\partial \log f(z|\theta)}{\partial \theta'}\bigg|_{T(G)} dG(z)\right\},$$

and

$$\boldsymbol{T}^{(1)}(z; G) = \left(T_1^{(1)}(z; G), \ldots, T_p^{(1)}(z; G)\right)^\top$$

16 Or over the full space of all probability measures on the sample space.

is the influence function[17] of a p-dimensional functional $\boldsymbol{T}(G)$ at the distribution G.

In the following definition, the first part is related to the estimate of the expected log likelihood and the second part to the bias estimate $b_1(\hat{G})$ obtained by replacing the unknown distribution G by its empirical counterpart $\hat{G}$. Let

$$\boldsymbol{T}^{(1)}(x_i;\hat{G}) = \left(T_1^{(1)}(x_i;\hat{G}),\dots,T_p^{(1)}(x_i;\hat{G})\right)^\top$$

be the p-dimensional empirical influence function, then:

Definition 35. *For a functional estimator, the approximate estimate of the GIC is defined as:*

$$\begin{aligned} GIC(\mathbf{X}_n;\hat{G}) = &-2\sum_{i=1}^{n}\log f(x_i|\hat{\boldsymbol{\theta}}) \\ &+\frac{2}{n}\sum_{i=1}^{n}\operatorname{tr}\left\{\boldsymbol{T}^{(1)}(x_i;\hat{G})\frac{\partial\log f(x_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}\bigg|_{\hat{\boldsymbol{\theta}}}\right\}. \end{aligned} \tag{93}$$

If a nonfunctional estimator is employed (e.g. the MLE), then the GIC reduces to the TIC as a generalization of the AIC, the Generalized AIC (GAIC). Let $\hat{\boldsymbol{\theta}}_{ML}$ be the MLE, and write $\hat{\boldsymbol{\theta}}_{ML} = \boldsymbol{T}_{ML}(\hat{G})$, with $\boldsymbol{T}_{ML}$ being the p-dimensional functional that solves the "implicit equation":

$$\int \frac{\partial}{\partial\boldsymbol{\theta}}\log f(z|\boldsymbol{\theta})\Big|_{\boldsymbol{T}_{ML}(G)}\mathrm{d}G(z) = \mathbf{0} \tag{94}$$

In the following definition, the bias estimate is obtained as in Eq. 86, Definition 30, Section 2.7.1, for the TIC:

$$\boldsymbol{I}(G) = \int \frac{\partial\log f(z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\log f(z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{T}_{ML}(G)}\mathrm{d}G(z), \tag{95}$$

$$\boldsymbol{J}(G) = -\int \frac{\partial^2\log f(z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\bigg|_{\boldsymbol{T}_{ML}(G)}\mathrm{d}G(z), \tag{96}$$

but via the empirical distribution function $\hat{G}$.

Definition 36. *The approximate estimate of the GAIC is given by :*

$$GAIC\left(\mathbf{X}_n;\hat{G}\right) \equiv -2\sum_{\alpha=1}^{n}\log f\left(X_\alpha\middle|\hat{\boldsymbol{\theta}}_{ML}\right) + 2\operatorname{tr}\left\{\boldsymbol{I}(\hat{G})\boldsymbol{J}(\hat{G})^{-1}\right\}. \tag{97}$$

17 In nonparametric regression problems, the influence function is used to approximate the standard error of a plug-in estimator, and is connected with the Gâteaux derivative (i.e. directionally differentiable at a point in the sense of Gâteaux, equivalent to the definition of weakly directionally differentiable at a point as in Shapiro [268, p. 478]). See Wasserman [322, p. 18], Konishi and Kitagawa [174], and Huber and Ronchetti [156, Section 2.5] for further indications.

Now consider the case of the robust M-estimator as in Huber and Ronchetti [156], where $\hat{\boldsymbol{\theta}}_M$ is defined as the solution for the following implicit equation:

$$\sum_{\alpha=1}^{n} \psi_i \left(X_\alpha, \hat{\boldsymbol{\theta}}_M \right) = 0, \; i = \{1, \ldots, p\},$$

with function $\psi(X, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \rho(X, \boldsymbol{\theta})|_{\boldsymbol{T}_M(G)}$, defined over the domain $\mathcal{X} \times \Theta$, with ρ an arbitrary function (if $\rho(X, \boldsymbol{\theta}) = -\log f(X, \boldsymbol{\theta})$ the ordinary ML estimate is obtained - see [156, p. 50] for further assumptions on $\psi(\cdot)$), $\mathcal{X}$ the sample space, and $\Theta \subset \mathbb{R}^p$, where p is the total number of equations (parameters). The estimate $\hat{\boldsymbol{\theta}}_M = \boldsymbol{T}_M(\hat{G})$ is such that:

$$\int \psi_i[z, \boldsymbol{T}_M(G)] \mathrm{d}G(z) = 0, \; i = \{1, \ldots, p\}.$$

In this case, the influence function becomes:

$$\boldsymbol{T}_M^{(1)}(z; G) = \boldsymbol{M}(\boldsymbol{\psi}, G)^{-1} \boldsymbol{\psi}[z, \boldsymbol{T}_M(G)],$$

with $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^\top$, and $\boldsymbol{M}(\psi, G)$ a nonsingular $(p \times p)$ matrix such that:

$$\boldsymbol{M}(\boldsymbol{\psi}, G)^T = -\int \frac{\partial \boldsymbol{\psi}(z, \boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{T}_M(G)} \mathrm{d}G(z).$$

For the following definition, let $b_M^{(1)}(\hat{G})$ be the bias estimate of:

$$b_M^{(1)}(G) = \operatorname{tr} \left\{ \boldsymbol{M}(\boldsymbol{\psi}, G)^{-1} \int \boldsymbol{\psi}[z, \boldsymbol{T}(G)] \frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \Big|_{\boldsymbol{T}_M(G)} \mathrm{d}G(z) \right\}.$$

Definition 37. *The approximate estimate of the GIC for the robust M-estimator is given by:*

$$GIC_R(\mathbf{X}_n; \hat{G}) = -2 \sum_{\alpha=1}^{n} \log f \left(X_\alpha | \hat{\boldsymbol{\theta}}_M \right) + 2 b_M^{(1)}(\hat{G}). \tag{98}$$

An analogous version of SRIC for functional estimators is defined. Write:

$$\hat{I}_\lambda(\boldsymbol{T}(\hat{G})) = n^{-1} \sum_{\alpha=1}^{n} \boldsymbol{\psi} \left(X_\alpha, \hat{\boldsymbol{\theta}}_\lambda \right) \frac{\partial \log f(X_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \Big|_{\hat{\boldsymbol{\theta}}_\lambda}, \tag{99}$$

$$\hat{J}_\lambda(\boldsymbol{T}(\hat{G})) = n^{-1} \sum_{\alpha=1}^{n} \frac{\partial \boldsymbol{\psi}(X_\alpha|\boldsymbol{\theta})^\top}{\partial \boldsymbol{\theta}} \Big|_{\hat{\boldsymbol{\theta}}_\lambda}. \tag{100}$$

Definition 38. *If the M-estimator is employed under the penalized likelihood procedure, then the approximate estimate of the GIC is given by:*

$$GIC_\lambda(\mathbf{X}_n; \hat{G}) = -2 \sum_{\alpha=1}^{n} \log f(X_\alpha|\hat{\boldsymbol{\theta}}_\lambda) + 2 \operatorname{tr} \left\{ \hat{I}_\lambda(\boldsymbol{T}(\hat{G})) \hat{J}_\lambda^{-1}(\boldsymbol{T}(\hat{G})) \right\}. \tag{101}$$

2.7.3 *TIC and Composite Maximum Likelihood*

Varin and Vidoni [315, p. 523] also generalized the TIC for the Composite Maximum Likelihood Estimator (CMLE). The CMLE is a class of pseudolikelihood estimators which includes the full likelihood as a special case.

Let $\{f(\mathbf{y};\boldsymbol{\theta}),\ \mathbf{y}=(y_1,\ldots,y_n)\in\mathcal{Y}\subset\mathbb{R}^n,\boldsymbol{\theta}\in\boldsymbol{\Theta}\}$ be a parametric statistical model, with $\mathcal{Y}\subseteq\mathbb{R}^n$, $\boldsymbol{\Theta}\subseteq\mathbb{R}^d$, $n\geq 1$, and $d\geq 1$. Consider a set of events $\{\mathcal{A}_i:\mathcal{A}_i\subseteq\mathcal{F},i\in I\}$, where $I\subseteq\mathbb{N}$, and $\mathcal{F}$ is some σ-algebra on $\mathcal{Y}$. Then:

Definition 39. *A composite likelihood is defined as:*

$$L_c(\boldsymbol{\theta};\mathbf{y})=\prod_{i\in I}f(\mathbf{y}\in\mathcal{A}_i;\boldsymbol{\theta})^{w_i},$$

where $f(\mathbf{y}\in\mathcal{A}_i;\boldsymbol{\theta})=f(\{y_j\in\mathcal{Y}:y_j\in\mathcal{A}_i\};\boldsymbol{\theta})$*, with* $\mathbf{y}=(y_1,\ldots,y_n)$*, while* $\{w_i,i\in I\}$ *is a set of suitable weights.*

Define its associated composite log likelihood as:

$$l_c\left(\boldsymbol{\theta};\mathbf{y}\right)=\log L_c\left(\boldsymbol{\theta};\mathbf{y}\right), \tag{102}$$

and its sample counterpart:

$$l_c\left(\hat{\boldsymbol{\theta}}_{CL};\mathbf{y}\right)=\sum_{i\in I}\log f\left(\mathbf{Y}\in\mathcal{A}_i;\hat{\boldsymbol{\theta}}_{CL}\right)w_i, \tag{103}$$

i.e. the maximised composite log likelihood.

Let $\mathbf{Z}=(Z_1,\ldots,Z_n)$ be a random variable with probability density $g(\mathbf{z})$, and denote:

$$L_c\left(g;\mathbf{Z}\right)=\prod_{i\in I}g(\mathbf{Z}\in\mathcal{A}_i)^{w_i}, \tag{104}$$

$$L_c\left(h;\mathbf{Z}\right)=\prod_{i\in I}h(\mathbf{Z}\in\mathcal{A}_i)^{w_i}. \tag{105}$$

the composite likelihood functions of $g(\mathbf{z})$ and $h(\mathbf{z})$. Then we can define the KLI via the CMLE as a linear combination of the KLI in the MLE case:

Definition 40. *The KLI of a probability density* $h(\mathbf{z})$ *with respect to* $g(\mathbf{z})$ *is given by:*

$$\begin{aligned}I_C\left(g,h\right)&=E_{g(z)}\left[\log\left\{\frac{L_c\left(g;\mathbf{Z}\right)}{L_c\left(h;\mathbf{Z}\right)}\right\}\right]\\&=\sum_{i\in I}E_{g(z)}\left[\left\{\log g(\mathbf{Z}\in\mathcal{A}_i)-\log h(\mathbf{Z}\in\mathcal{A}_i)\right\}w_i\right]\end{aligned} \tag{106}$$

For a first-order unbiased selection criterion, the following regularity conditions are required for robustness to misspecification:

Assumptions 4. *The regularity conditions for inference and model selection with composite log likelihood are the following:*

(i) $\boldsymbol{\Theta}$ *is a compact subset of* $\mathbb{R}^d$, $d \geq 1$, *and* $\forall\ \mathbf{y} \in \mathcal{Y}$, $L_C(\boldsymbol{\theta}; \mathbf{y})$ *is twice continuously differentiable with respect to* $\boldsymbol{\theta}$*;*

(ii) The composite likelihood estimator, $\hat{\boldsymbol{\theta}}_{CL}$*, solves the composite likelihood equation and there exists* $\boldsymbol{\theta}^* \in \operatorname{int}(\boldsymbol{\Theta})$*:*

$$E_{g(y)}\left[\bigtriangledown l_C\left(\boldsymbol{\theta}^*; \mathbf{Y}\right)\right] = 0, \tag{107}$$

exactly or asymptotically for diverging n*, where* $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)$ *is the sample of size* n*, and* $g(y)$ *the true distribution that may or may not be included in the the family* $\{f(\mathbf{y}; \boldsymbol{\theta}), \mathbf{y} \in \mathcal{Y}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$*;*

(iii) $\hat{\boldsymbol{\theta}}_{CL}$ *is a consistent estimator of* $\boldsymbol{\theta}^*$ *and asymptotically Gaussian.*

Varin and Vidoni developed a first-order unbiased selection statistic similarly to Takeuchi [292]. Let $\hat{I}_c$ and $\hat{J}_c$ be consistent, first-order unbiased estimators respectively of:

$$\begin{aligned} J(\boldsymbol{\theta}^*) &= \operatorname{Var}_{g(y)}\left[\bigtriangledown l_c(\boldsymbol{\theta}^*; \mathbf{Y})\right], \\ H(\boldsymbol{\theta}^*) &= E_{g(y)}\left[\bigtriangledown^2 l_c(\boldsymbol{\theta}^*; \mathbf{Y})\right], \end{aligned}$$

with $E_{g(y)}$ and $\mathrm{V}_{g(y)}$ respectively the expected value and the variance with respect to $g(y)$. Then:

Definition 41. *The TIC via CML is defined as:*

$$\begin{aligned} &E_{g(y)}\left[l_c\left\{\hat{\boldsymbol{\theta}}_{CL}\left(\mathbf{Y}\right)\right\}\right] = \\ &E_{g(y)}\left[l_c\left\{\boldsymbol{\theta}^*; \mathbf{Y}\right\}\right] - \frac{1}{2}\operatorname{tr}\left\{J(\boldsymbol{\theta}^*)\left(H^*\right)^{-1}\right\} + o(1), \end{aligned} \tag{108}$$

and its approximate estimate given by:

$$TIC_{CL} = -2l_c\left(\hat{\boldsymbol{\theta}}_{CL}; \mathbf{y}\right) + 2\operatorname{tr}\left\{\hat{I}_c \hat{J}_c^{-1}\right\}. \tag{109}$$

3

MODEL SELECTION VIA INFORMATION AND PREDICTION CRITERIA: A SURVEY FOR TIME SERIES MODELS

ABSTRACT

For dependent data, information's perspective met prediction and stochastic processes theory. We discuss common solutions to lag selection for univariate and multivariate parametric time series models. A selective overview of nonparametric techniques for nonlinear time series models follows, focussing on the nonparametric asymptotic final prediction error. We indicate recent developments, theoretical works, and other surveys of methods for high-dimensional settings.

Keywords: model selection, information criteria, time series, parametric, nonparametric, high-dimensional.

3.1 INTRODUCTION

Temporal dependence arises when moving from random variables to stochastic processes and modifies problem's setting by considering a nonzero autocovariance function, which modifies assumptions and strategies for its handling. Another way of seing it is that, if the observations are taken from a process not displaying the *i.i.d.* feature, but instead exhibiting some type of persistence, we position inside time series analysis.

The formal definition of a stochastic process,[1] requires an abstract *probability space* $(\Omega, \mathcal{A}, \mathbb{P})$, a measurable space $(S, \mathcal{B})$, and any nonempty set T that will serve to index the process, i.e. the set of times. A stochastic process $\{X_t\}$, $t \in (-\infty, \infty)$, is a family of random variables $X_t := \{X_t : t \in T\}$. The random variable X_t is said *measurable* if $X_t^{-1}(B) \in \mathcal{A}$, $\forall B \in \mathcal{B}$, and such that $X_t : \omega \to S$, $\forall\, t \in T, \omega \in \Omega$. The state-space of the process, S, is the arrival space of the random variable X_t.[2] For instance, S a Polish space with $\mathcal{B}$ its Borel σ-algebra, or the measurable space $(S, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Note that the random variable X_t is function of the two variables t and ω. For clarity, we write that: $X_t \equiv X(t, \omega)\ \forall\ (t, \omega) \in T \times \Omega$. We observe the sequence of random variables indexed by the subscript of time, i.e. the series $\{x_t\}$, $t = \{1, 2, \ldots, n\}$. Our goal is to select the appropriate dimension of the model. Sometimes it translates into choosing the lag delivering best prediction in some sense, i.e. under some definition of divergence.

1 See [50, 55, 240].

2 In this sense, Ω can be thought of as a 'departure' space.

The work is organized as follows. Section 3.2 will introduce MS via IC and Prediction Criteria (PC) in autoregressive (AR) models. PC refer to selection methods derived from the one-step ahead prediction error, e.g. Akaike's Final Prediction Error (FPE). Subsection 3.2.2 focusses on the FPE for AR models, Subsection 3.2.3 on criteria for analysis in the frequency domain setting, Subsection 3.2.4 on criteria for ARMA models, and Subsection 3.2.5 on criteria for VARMA and VAR models. Section 3.3 proposes a path for nonparametric analysis of nonlinear time series models, focusing on the methodology behind the nonparametric analogue of Akaike's FPE. Brief indications on recent high-dimensional works discussing MS are included in Section 3.4. One conclusive remark is shared in Section 3.5, and short bibliographic notes are included in Appendix A.2.

3.2 TIME SERIES AND MODEL SELECTION

Our goals is to select the appropriate dimension of the model, which in some applications coincides with the order p. Results on MS for time series initially assumed some type of mixing conditions or strict stationarity, followed by results for weak stationarity conditions. See Brockwell et al. [56, Ch. 1] on stationarity, and Doukhan [104] on mixing conditions. Since concepts of these conditions are usually employed on their formal derivations, short notes are presented in Appendix A.2.4 together with further bibliographic notes.

3.2.1 *Information criteria in time series*

Initial solutions for MS came from the hypothesis testing literature applied specifically to time series, e.g. Quenouille [229], Wold [339], Whittle [331], Hannan [132], and from multivariate regression as in Mallows [203, 204]. Akaike [7] underlined how the AIC and the minimum theoretical information criterion estimate (MAICE) procedure were preceded by the FPE, i.e. a PC based on the MSPE advanced in the context of AR models of finite order. Since the beginning, the development on IC have been associated with those in time series analysis and control, viz. Akaike [7, p. 717]; Rissanen [241]; the conditions for asymptotic consistency in Hannan and Quinn [137] from time series; and Bozdogan [52, p. 347]. See Table 9 in Appendix A.2 for examples of IC and PC in time series. These include:

i. the FPE_α [2] of Akaike;

ii. the FPE^β [41] of Bhansali and Downham;

iii. the HQ [137] of Hannan and Quinn;

iv. the AIC_{AR} [217] of Ogata;

v. the $\text{FPE}_\gamma(k)$ [275] of Shibata;

vi. the APE [243] of Rissanen;

vii. the $I(k, C_n)$ [356] of Zhao et al.;

viii. the AIC_c [157] of Hurvich and Tsai;

ix. the GIC_A [227] and GIC_B [227] of Potscher;

x. the FIC [328] of Wei;

xi. the QAIC [183] and QAIC_c [183] of Lebreton et al.;

xii. the ODQ [354] of Zhang and Wang;

xiii. the FIC (k) [180] of Lai and Lee;

xiv. the WIC [341] of Wu and Sepulveda;

xv. the RIC [269] of Shi and Tsai;

xvi. the EIC [43] of Billah et al.;

CONSISTENCY, EFFICIENCY, AND PARSIMONY Besides asymtotic consistency and efficiency for MS, in time series analysis and control, the property of parsimony is relevant. Box et al. [51] citing Tukey [311] defined parsimony as: "*the smallest possible number of parameters for adequate representations*". Bozdogan [53] connects this property with Occam's Razor: "*the desirability of selecting, among the accurate models of reality, those which are most parsimonious*". On the subject, see the counterexamples to parsimony in Findley [114].

Tong [301] pursued the AIC approach to determine the order k of a Markov chain. Consider the sequence of observations $S = \{x_1, \ldots, x_n\}$ from an ergodic and stationary Markov chain. Each observation may assume $\{1, 2, \ldots, t\}$ states. It is necessary to verify for the smallest integer $k > 0$ that the conditional probability:

$$\mathbb{P}\left\{x_n | x_{n-1}, x_{n-2}, \ldots\right\} = \mathbb{P}\left\{x_n | x_{n-1}, x_{n-2}, \ldots x_{n-k}\right\} \tag{110}$$

for all n, given the sequence of observations S. To proceed, let:

$$_k\eta_L = -2\sum_{i=1}^{N} \log\left(\frac{f\left(X_i|\hat{\theta}_k\right)}{f\left(X_i|\hat{\theta}_L\right)}\right), \tag{111}$$

be minus two times the logarithm of the LR statistic, where $\hat{\theta}_L$ is the unrestricted MLE of θ, and $\hat{\theta}_k$ is its restricted version. Denote $\left(\bigtriangledown t^{L+1} - \bigtriangledown t^{k+1}\right)$ as the degrees of freedom of $_k\eta_L$ which is asymptotically a chi-squared random variable, with $\bigtriangledown t^j = t^j - t^{j-1}, j \geq 1$.

Definition 42. *Following arguments as in Section 2.3, Tong showed that to satisfy (110) one can minimize the approximate estimate of the AIC for Markov chains:*

$$R(k) = {}_k\eta_L - 2\left(\bigtriangledown t^{L+1} - \bigtriangledown t^{k+1}\right). \tag{112}$$

AUTOREGRESSIVE MODELS An autoregressive (AR) model of order p, AR(p), is defined as:

$$x_t = \beta_0 + \beta_1 x_{t-1} + \cdots + \beta_p x_{t-p} + \varepsilon_t, \tag{113}$$

where x_t are observations from the X_t process, $\{\epsilon_t\}$ are white noise innovations with $\text{Cov}\,(x_t, \epsilon_l) = 0$, where $l = \{1, \ldots, n\}$.

Results for AR models are numerous. For instance, Shibata [270] studied the problem of the selection of the order of an AR models by AIC, obtained the asymptotic distribution of the selected order, and evaluated the asymptotic risks functions of parameters' estimates for the order selected by AIC. This work showed that the AIC is not consistent for finite dimensional models. Findley [113] extended rigorously results on bias correction also present in Ogata [217] in misspecified Markov models or AR models, stressing the importance of the bias correction. Wong and Li [340] studied the AICc for self-exciting threshold autoregressive (SETAR) models via the conditional LSE in small samples. Ng and Perron [213] studied theoretically and by simulation the sensitivity of the AIC and the BIC to:

(i) the effective numbers of observations;

(ii) the degrees of freedom adjustment of the estimated variance; and

(iii) the penalty for overfitting in relation to sample size.

They conclude that these are relevant issues for valid model comparison.

The next subsection will present Akaike's FPE [1, 2] for AR models in detail, focussing in some hints for its computation. Additional PC derived from the FPE, the Hannan-Quinn Information Criterion (HQ), and the MRIC are overviewed in Sections 3.2.2.2 and 3.2.2.3.

3.2.2 *Akaike's Final Prediction Error and univariate time series models*

The FPE is defined for time series and for general regression problems. Given that AR is a particular type of regression, there is interest in studying its residual, $\hat{\varepsilon}_t = \hat{x}_t - x_t$, and to develop a measure from this quantity. The MSPE , or some transform of it, has been used frequently as such measure:

$$\text{MSPE} = E\left[(\hat{x}_t - x_t)^2\right] \tag{114}$$

The FPE is not an IC, in the sense that it does not derive from information-theoretic considerations. Instead, it is a technique based on the MSPE, a measure depending on the prediction errors. Therefore, a PC.

PENALTIES Transformations of the MSPE include $\mathcal{L}_1$, $\mathcal{L}_2$ penalizations, or their combination. See respectively ridge regression as in Hoerl and Kennard [148], Tibshirani's lasso [299], elastic net of Zou and Hastie [359], or Xu et al. [342] with $L_{1/2}$ penalty. See also different measures of forecast accuracy for univariate time series forecasts as in Hyndman and Koehler [159] or in Peña and Sánchez [220].

The following paragraphs present different penalties and their associated approximate estimates, where distinct settings and sets of assumptions modify the building and development of IC and PC.

3.2.2.1 *The FPE for AR models*

A detailed description of the procedure to compute the FPE for an AR process was later given in Akaike [2], overviewed in the following paragraphs. It relies on the asymptotic theory from [22, 96].

Let a stationary process $\{X(n)\}$, and its predictor $\hat{X}(n)$. Assuming that the dependence between the past history of $X(n)$, used to obtain $\hat{X}(n)$, and recent values of $X(n)$ is vanishing, we consider that prediction is made with process $\{Y(n)\}$, different from $\{X(n)\}$, but sharing the same statistical properties. Define the FPE as the MSPE of the one-step ahead predictor $\hat{X}(n)$:

$$\text{FPE}[\hat{X}(n)] \equiv \text{MSPE}_1[\hat{X}(n)] = E[\Big(X(n) - \hat{X}(n)\Big)^2],$$

with the subscript highlighting that it refers to the $h = 1$-step ahead forecast. When $\{X(n)\}$ is stationary and the predictor $\hat{Y}(n)$ of $\{Y(n)\}$ is linear, write the predictor $\hat{Y}(n)$:

$$\hat{Y}(n) = \sum_{m=1}^{M} \hat{a}_M(m) Y(n-m) + \hat{a}_M(0), \tag{115}$$

where the estimated parameter $\hat{a}_M(m)$ is a function of the observed $\{X(n)\}$. The FPE of the general predictor will be

$$\text{FPE}[\hat{Y}(n)] = \sigma^2(M) + \sum_{l=0}^{M} \sum_{m=0}^{M} E\left[\Delta a_M(l) \Delta a_M(m)\right] V_{M+1}(l,m), \tag{116}$$

where

$$\begin{aligned} \sigma^2(M) &= E\left[\left(Y(n) - \sum_{m=1}^{M} a_M(m) Y(n-m) - a_M(0)\right)^2\right] \\ &= \min_{a(m)} E\left[\left(Y(n) - \sum_{m=1}^{M} a(m) Y(n-m) - a(0)\right)^2\right], \end{aligned} \tag{117}$$

i.e. $a_M(m)$ delivers the best linear predictor in the mean square sense, $\Delta a_M(l) = \hat{a}_M(l) - a_M(l)$ is the difference between the estimated and the true parameter referred to the l-lag, and

$$V_{M+1}(l,m) = E\left[Y(n-l)(n-m)\right], \tag{118}$$

with $l, m = \{1, 2, \ldots, M\}$ is the autocovariance of order $\{l-m\}$ for the univariate process Y. For MS, we will select the model with the smallest FPE.

Assume that $X(n)$ is a stationary process with a DGP of the type:

$$X(n) = \sum_{m=1}^{M} a(m)X(n-m) + a(0) + \varepsilon(n),$$

for $l, m = \{1, 2, \ldots, M\}$, with $\varepsilon(n)$ *i.i.d.*, with zero mean $E\left[\varepsilon(n)\right] = 0$, and positive variance

$$E\left[\left(\varepsilon(n)\right)^2\right] = \sigma^2 > 0. \tag{119}$$

Let $\{X(n) : n = -M+1, -M+2, \ldots, N\}$ be our sample, and $\hat{a}_M(m)$ be the LS estimate of $a(m)$. Then, for the AR(M), the FPE of predictor $\hat{Y}(n)$ using the Ordinary least squares (OLS) estimate is given by:

$$\begin{aligned}\mathrm{FPE}\left[\hat{Y}(n)\right] = \sigma^2 &+ \sum_{l=1}^{M}\sum_{m=1}^{M} E\left[\Delta a_M(m)\Delta a_M(l)\right] R_{XX}(l-m) \\ &+ E\left[\left(\Delta\overline{X}_0 - \sum_{m=1}^{M}\hat{a}_M(m)\Delta\overline{X}_m\right)^2\right],\end{aligned} \tag{120}$$

where $\Delta\overline{X}_l = \overline{X}_l - \mathbb{E}\left[X(n)\right]$ is the difference between the sample mean and its population counterpart, and

$$R_{XX}(l-m) = E\left[X(n-l)X(n-m)\right] - \left(E\left[\left(X(n)\right)\right]\right)^2 \tag{121}$$

is the population autocovariance matrix of lag $l-m$. This is obtained in the case of correct specification, using the independence property and variables in deviations from their unconditional means.

For the asymptotic evaluation of the FPE for predictor $\hat{Y}(n)$, consider the expectation conditioned on X of the squared difference $(Y(n) - \hat{Y}(n))^2$, $E_X[(Y(n) - \hat{Y}(n))^2]$, instead of the unconditional expectation as in Equation (120). The resulting expression is equal to:

$$\begin{aligned}&E_X[(Y(n) - \hat{Y}(n))^2] = \\ &\sigma^2 + \sum_{l=1}^{M}\sum_{m=1}^{M}\Delta a_M(l)\Delta a_M(m)R_{XX}(l-m) \\ &+ \left(\Delta\overline{X}_0 - \sum_{m=1}^{M}\hat{a}_M(m)\Delta\overline{X}_m\right)^2.\end{aligned}$$

Finally, to give the approximate estimate of the FPE, after the application of the results from [22, 96], write a consistent estimate of σ^2 as:

$$S(M) = C_{xx}(0,0) - \sum_{l=1}^{M} \hat{a}_M(l) C_{xx}(0,l), \tag{122}$$

where:

$$C_{xx}(m,l) = N^{-1} \sum_{n=1}^{M} \left(X(n-m) - \overline{X}_m\right)\left(X(n-l) - \overline{X}_l\right), \tag{123}$$

$$C_{xx}(0,0) = N^{-1} \sum_{n=1}^{M} \left(X(n) - \overline{X}_0\right)\left(X(n) - \overline{X}_0\right), \tag{124}$$

with

$$\overline{X}_m = N^{-1} \sum_{n=1}^{N} X(n-m),\ m = \{0,1,2,\ldots,M\}, \tag{125}$$

and $\hat{a}_M(m)$ the LS estimate solving:

$$\sum_{m=1}^{M} C_{xx}(m,l)\hat{a}_M(m) = C_{xx}(0,l),\ l = \{1,2,\ldots,M\}. \tag{126}$$

The following Definition 43 follows from both Theorem 1 and Lemma 1 in Akaike [2], derived using the asymptotic results in [22, 96], and delivers both the FPE for the predictor of an AR and its approximate estimate. Let $(1 - N^{-1}(M+1))^{-1}S(M)$ be an estimate of σ^2, useful to estimate the FPE given the ergodicity of X and the OLS estimate of the parameters, M be the order of the model under consideration, N the sample size, and $S(M)$ a consistent estimate of σ^2, then:

Definition 43. *The definition of* $(\mathrm{FPE})_M$ *as an asymptotic evaluation of the* $\mathrm{FPE}\left[\hat{Y}(n)\right]$ *is:*

$$\mathrm{FPE}_M\left[\hat{Y}(n)\right] = \left(1 + \frac{M+1}{N}\right)\sigma^2, \tag{127}$$

and is estimated by $(\mathrm{FPE})(M)$*:*

$$(\mathrm{FPE})(M) = (1 + \frac{M+1}{N})(1 - N^{-1}(M+1))^{-1}S(M). \tag{128}$$

To summarize, the algorithm's pseudo-code to compute the FPE [1] for an AR process is presented in Listing 1. Define L to be the upper limit large enough to avoid excluding the efficient model. Notice that L can be smaller than the maximum lag usually considered to estimate the power spectrum in the frequency domain. Define the sample autocovariances for lags $l = \{0,1,\ldots,L\}$, and denote these by "S_ACF(l)". Define AR(L) the AR model of order L and estimate it. Recall that to fit by OLS an AR model of order M, with $M = \{1,2,\ldots,L\}$, requires to:

Listing 1: FPE algorithm's pseudo-code

```
        Center variables to their sample means
        Set L
        Consider an AR(L) to fit model to data
        Compute S_ACF(l)
        Fit AR(M), M= {1,2,...,L}, by OLS
        Estimate the FPE of order M
        Compute its relative value wrt FPE(0)
        Select order M that minimizes FPE
```

(i) minimize the Mean Square of Residuals with respect to parameters of the restricted regression;

(ii) solve the normal equations; and

(iii) obtain the estimates for each order M.

Then, denote with "FPE(M)" the estimate of FPE of order M, with FPE(0) the FPE of order 0.

3.2.2.2 *Further PC from the FPE*

Given that the original FPE is not a consistent estimator of the order, Akaike in the same article proposed the $(FPE)^{\alpha}$ which allows to consistently estimate K of a finite AR process. The required quantities have been already defined for Definition 43.

Definition 44. *The approximate estimate of the consistent estimator of K for a finire AR process is given by:*

$$(FPE)^{\alpha}(M) = (1 + N^{-\alpha}(M+1))(1 - N^{-1}(M+1))^{-1} S(M), \quad (129)$$

with $\alpha \in (0,1)$ a penalty weight.

Since $(FPE)^{\alpha}(M)$ adds a tendency to underestimate M_0 (too small values), Bhansali and Downham [41] further generalized the FPE. Define β as a positive and fixed constant, the sample variance as

$$\hat{\sigma}_k^2 = \frac{1}{T} \sum_{t=1}^{T} \left(X_t + \hat{a}_{k,1} X_{t-1} + \ldots + \hat{a}_{k,k} X_{t-k} \right)^2, \quad (130)$$

with $\hat{a}_{k,u} (u = 1, \ldots, k)$ the OLS estimates of the estimated AR(k) model, k the lag number under examination, and T the total number of observations. Then the following definition follows:

Definition 45. *The approximate estimate of the FPE to avoid underestimation is:*

$$\mathrm{FPE}_{\beta}(k) = \hat{\sigma}_k^2 (1 + \beta k / T). \quad (131)$$

Shibata [273] showed that AIC, FPE, and Mallow's C_p, are asymptotically mean efficient point-wise. Shibata [275] also proposed a further generalization of the previous criterion for regression problems, while suggesting a procedure for the choice of the fixed constant that weights the penalty term. Consider a regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}(k) + \boldsymbol{\epsilon}$, where $\mathbf{y}^\top = (y_1, \ldots, y_n)$ is an n-dimensional column vector observations, $\mathbf{X}$ is an $(n \times K)$ design matrix, $\boldsymbol{\beta}^\top(k) = (\beta_1, \ldots, \beta_k, 0, \ldots, 0)$ is the parameters' vector, and $\boldsymbol{\epsilon}^\top = (\epsilon_1, \ldots, \epsilon_n)$ is an *i.i.d.* vector of zero-mean Gaussian random variables with spherical variance $E[\epsilon_i] = \sigma^2 > 0$. Define $\hat{\boldsymbol{\beta}}^\top(k) = \left(\hat{\beta}_1, \ldots, \hat{\beta}_k\right)$ as the solution to the system: $\mathbf{X}^\top(k)\mathbf{X}(k)\hat{\boldsymbol{\beta}}(k) = \mathbf{X}^\top(k)\mathbf{y}$, with $\mathbf{X}(k)$ is a $(n \times k)$ submatrix with the first k column vectors of $\mathbf{X}$. Let k be the specific dimension being evaluated, K the maximum dimension considered, the Residual Sum of Squares (RSS) for model of dimension k:

$$n\hat{\sigma}^2(k) = \left\| y - X\hat{\beta}(k) \right\|^2, \tag{132}$$

and denote with

$$\tilde{\sigma}^2(K) = [n/(n-K)]\,\hat{\sigma}^2(K) \tag{133}$$

an unbiased estimate of σ^2.

Definition 46. *The approximate estimate of the FPE given by [275] is:*

$$\mathrm{FPE}_\gamma(k) = n\hat{\sigma}^2(k) + \gamma k\tilde{\sigma}^2(K), \tag{134}$$

with γ the penalty term.

Notice that if we set $\gamma = 2$, we obtain the usual AIC; $\gamma = \log n$ Schwarz's BIC; or $\gamma = c\log\log n$ the HQ criterion of Hannan and Quinn, introduced below.

3.2.2.3 *The HQ and the MRIC*

Hannan and Quinn [137] proposed a strongly consistent MS criterion, included in many statistical software, the popular Hannan-Quinn Information Criterion (HQ) criterion. For its development, they used the Law of Iterated Logarithms (LIL),[3] explaining its resemblance to the BIC, for finite ergodic stationary AR models under correct specification and estimated via the consistent Yule-Walker estimator in the identically distributed setting. Consider an AR(k) demeaned model:

$$\sum_{j=0}^{k} \alpha(j)\,\{x(n-j) - \mu\} = \epsilon(n), \tag{135}$$

with $\mathrm{E}[x(n)] = \mu$, $\alpha_0 = 1$, $\{x(n)\}$ an ergodic stationary sequence with mean μ and finite variance. The linear innovations $\epsilon(n)$ are such that

3 For the LIL in stationary ergodic martingale difference sequences, see [288].

$\sum_{j=0}^{k} \alpha(j) z^j \neq 0$, $|z| \leq 1$, $E\left[\epsilon(m)\epsilon(n)\right] = \delta_{m,n}\sigma^2$, and to apply the limit theory is only necessary to assume:

$$E\left[\epsilon(n)\middle|\mathcal{F}_{n-1}\right] = 0, \tag{136}$$

$$E\left[\epsilon^2(n)\middle|\mathcal{F}_{n-1}\right] = \sigma^2, \tag{137}$$

$$E\left[\epsilon^4(n)\right] < +\infty, \tag{138}$$

where the σ-algebra generated by $x(m)$ is denoted by $\sigma\left\{x(m)\right\} \equiv \mathcal{F}_{n-1}$, $m \leq n$, (or equivalently, by $\epsilon(m)$, $m \leq n$). Let k be the considered order of the regression, c a positive constant greater than 1, and n the sample size. Then:

Definition 47. *The approximate estimate of the HQ is given by:*

$$\mathrm{HQ}(k) = -2l(\hat{\theta}_k) + 2kc\log\log(n). \tag{139}$$

[4]Let $\{y_t\}$ and $\{\mathbf{x}_t\}$ be two weakly stationary demeaned stochastic processes of dimensions 1 and m respectively. Let $y_{t+h} = \boldsymbol{\beta}_h^* \mathbf{x}_t^* + \boldsymbol{\epsilon}_{t+h}$ be the D.G.P for h-step ahead prediction, and consider a sample of

$$n = \{1, 2, \ldots, N, N+1, \ldots, N+h = n\}$$

observations, an h-step ahead possibly-misspecified forecasting model of the type: $y_{n+h} = \boldsymbol{\beta}_h \mathbf{x}_n + \boldsymbol{\varepsilon}_n^{(h)}$, where pseudo-true parameters' vector is given by

$$\boldsymbol{\beta}_h = \underset{\mathbf{C}\in\mathbb{R}^m}{\operatorname{argmin}}\, E\left[\left(y_{t+h} - \boldsymbol{c}^\top \mathbf{x}_t\right)^2\right],$$

and $\boldsymbol{\varepsilon}_n^{(h)}$ is the possibly-misspecified error for h-step ahead forecasting. The dependence of $\mathbf{x}_t$ on h exists, but it is suppressed for notational convenience, so the consideration of a specific regressor also depends on the forecast horizon h. The LS estimator delivers $\hat{y}_{n+h} = \hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n$, where

$$\hat{\boldsymbol{\beta}}_n = \left(\sum_{t=1}^{N} \mathbf{x}_t \mathbf{x}_t^\top\right)^{-1} \sum_{t=1}^{N} \mathbf{x}_t y_{t+h}$$

is the LS estimator of $\boldsymbol{\beta}_h$ for h-step ahead regression based on the sample of n observations. As measure of interest take the h-step ahead Mean-Squared Prediction Error (MSPE), which is derived from the difference between the observed and the estimated values, i.e. $\mathrm{MSPE}_h = E\left[\left(y_{t+h} - \hat{y}_{t+h}\right)^2\right]$. Theorem 2.1 in Hsu et al. [153] allows for the decomposition of the MSPE in two parts. The first one being the Misspecification Index (MI), which is linked to the goodness-of-fit of the model and is equal to the variance of the h-step ahead prediction error, i.e. $\mathrm{MI}_h = E\left[\varepsilon_{1,h}^2\right]$. The second component is the Variability Index (VI),

4 The following paragraphs are similar to those in Section 2.5.2, included for convenience.

which depends upon the variance of the h-step ahead predictor $\hat{y}_{n+h}$, and which is also connected to the estimation error of $\hat{\boldsymbol{\beta}}_n(h)$:

$$\mathrm{VI}_h = L_h = \mathrm{tr}\left\{\mathbf{R}^{-1}\mathbf{C}_{h,0}\right\} + 2\sum_{s=1}^{h-1}\mathrm{tr}\left\{\mathbf{R}^{-1}\mathbf{C}_{h,s}\right\}.$$

Here, $\mathbf{R} = E\left[\mathbf{x}_1\mathbf{x}_1^\top\right]$ is the (non-singular) variance-covariance matrix of the regressors, whereas $\mathbf{C}_{h,s} = E\left[\mathbf{x}_1\mathbf{x}_{1+s}^\top\varepsilon_{1,h}\varepsilon_{1+s,h}\right]$ represents the cross-covariance matrix between the regressors and the h-step ahead prediction error. The approach proposed by Hsu et al. [153] selects the model that minimises the h-step ahead MSPE_h. The minimization occurs by selecting the model with the smallest VI_h among those with the smallest MI_h, sequentially.

The asymptotic decomposition of the MSPE for the h-steps ahead prediction is derived in the univariate time series case:

$$\mathrm{MSPE}_h = \mathrm{MI}_h + n^{-1}(\mathrm{VI}_h + o(1)), \tag{140}$$

and their method of moments estimator[5] for both MI_h and VI_h are:

$$\hat{\mathrm{MI}}_h = N^{-1}\sum_{t=1}^{N}\left(\hat{\varepsilon}_t^{(h)}\right)^2, \tag{141}$$

$$\hat{\mathrm{VI}}_h = \mathrm{tr}\left\{\hat{R}^{-1}\hat{\mathbf{C}}_{h,0}\right\} + 2\sum_{s=1}^{h-1}\mathrm{tr}\left\{\hat{R}^{-1}\hat{\mathbf{C}}_{h,s}\right\}, \tag{142}$$

where

$$\hat{\mathbf{R}} = N^{-1}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{x}_t^\top, \tag{143}$$

$$\hat{\mathbf{C}}_{h,s} = (N-s)^{-1}\sum_{t=1}^{N-s}\mathbf{x}_t\mathbf{x}_{t+s}^\top\hat{\varepsilon}_t^{(h)}\hat{\varepsilon}_{t+s}^{(h)}, \tag{144}$$

$$\hat{\varepsilon}_t^{(h)} = y_{t+h} - \hat{\boldsymbol{\beta}}_n(h)\mathbf{x}_t, \tag{145}$$

with $\hat{\varepsilon}_t^{(h)}$ defined as the estimated forecast error.

Definition 48. *Based upon such asymptotic decomposition of the* MSPE *it is possible to derive the estimated* MRIC *as follows:*

$$\hat{\mathrm{MRIC}}_h = \hat{\mathrm{MI}}_h + \frac{\alpha_n}{n}\hat{\mathrm{VI}}_h, \tag{146}$$

where

$$\alpha_n/n^{1/2} \to +\infty, \qquad \alpha_n/n \to 0. \tag{147}$$

5 We refer to the method of moments as in [153], sometimes called 'empirical method of moments', 'analog method', or 'empirical method'. Hall [129, p. 5-7] recalled that Pearson proposed the estimation of parameters' vector *"by the value implied by the corresponding sample moments"* [129, p. 6]. And that he *"called this approach the 'Method of Moments'"* [129, p. 7], where to estimate some parameters, it is requested to satisfy the analogous sample moment condition with specified sample size.

The asymptotic efficiency of the MRIC is proved in Hsu et al. [153], Theorem 3.1. The MRIC approach selects the model that minimises the MSPE_h by selecting the model with the smallest VI_h among those with the smallest MI_h, sequentially. The MRIC is asymptotically efficient in the sense of Definition 9, and has interesting applications in combination to other variable selection and dimension reduction techniques in the high-dimensional setting. For indication on the determination of the penalty weight, see Remark 5 in Chapter 4.

3.2.3 *The frequency domain*

Time series can be studied from both the time and frequency domain. The Criterion autoregressive transfer function (CAT) [39, 219, 302], which appeared for the first time in the same 1974 issue of Akaike's [7], is a nonparametric criterion developed in the setting of an infinite dimensional AR process for lag selection via transfer function. See Brockwell and Davis [56, p. 123] or Box et al. [51, p. 8] for time series analysis by transfer functions.

Following Bhansali [39], consider $\{x_t\}$, $t = \{0, \pm 1, \pm 2, \dots\}$, a zero-mean stationary process with covariance function:

$$\gamma(p) = E\left[x_t x_{t-p}\right], \tag{148}$$

and spectral density function:

$$f(\lambda) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \gamma(p) \exp^{-it\lambda}, \tag{149}$$

which is the Fourier transform of the autocovariance function at frequency $\lambda \in (-\pi, \pi)$. Under *absolute summability* of the covariance function and non-vanishing spectral density, the AR(∞) representation of process $\{x_t\}$ exists, $\sum_{j=0}^{\infty} a_j x_{t-j} = \varepsilon_t$, with $a_0 = 1$, and $\{\varepsilon_t\}$ a sequence of random errors with $E[\varepsilon_t] = 0$, $E[\varepsilon_t^2] = \sigma^2 < \infty$, and $\mathrm{Cov}\,(\varepsilon_t, \varepsilon_{t-j}) = 0$, $j \neq t$.
Now, let sample $\{x_1, \dots, x_T\}$ from the AR(∞) process be modelled with an AR(p), where p is the optimal finite order approximation to the infinite process. Obtain the LSEs $\hat{a}_p(j)$ by minimizing

$$(T-p)^{-1} \sum_{t=p+1}^{T} \left(x_t + c_1 x_{t-1} + \cdots + c_p x_{t-p}\right)^2,$$

with minimum estimated variance $\hat{\sigma}_p^2$, where the optimization is with respect to parameters' vector $\boldsymbol{c} = (c_1, \dots, c_p)$, and each $c_j = a(j)$, $j = \{1, \dots, p\}$ are AR(p) parameters.

Write the transfer functions:

$$A(\lambda) = \sum_{j=1}^{\infty} a(j) \exp(-ij\lambda), \tag{150}$$

$$\hat{A}_p(\lambda) = \sum_{j=1}^{p} \hat{a}_p(j) \exp(-ij\lambda), \tag{151}$$

of both the AR parameters and its estimates respectively, where λ denotes the frequency, and $i = \sqrt{-1}$. For lag selection, define the penalty function as

$$J(p) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\| \hat{A}_p(\lambda) - A(\lambda) \right\|^2 \|A(\lambda)\|^{-2} \, \mathrm{d}\lambda.$$

Focus on estimators of $E\left[J(p)\right]$. Define the nonparametric estimate of σ^2 as:

$$\hat{\sigma}_{\infty}^2 = 2\pi \exp\left\{ \left[N^{-1} \sum_{s=1}^{N} \log \boldsymbol{I}^{(T)}(\lambda_s) \right] + e \right\}, \tag{152}$$

with $N = \lfloor \frac{1}{2}(T-1) \rfloor$, $\lfloor \cdot \rfloor$ the largest integer, e the Euler-Mascheroni constant, $\lambda_s = 2\pi \frac{s}{T}$ the s-th frequency, and the periodogram function as:

$$\boldsymbol{I}^{(T)}(\lambda) = (2\pi T)^{-1} \left\| \sum_{t=1}^{T} x_t \exp(-it\lambda) \right\|^2. \tag{153}$$

Parzen [219] estimated order p so that $\hat{A}_p(\lambda)$ is near $A(\lambda)$. For that matter, he introduced the CAT. Write σ_p^2 the MSPE predicting one-step ahead with memory p. Since it is unknown, let

$$\tilde{\sigma}_p^2 = T(T-p)^{-1} \hat{\sigma}_p^2 \tag{154}$$

be a consistent estimator of σ^2, where $\hat{\sigma}_p^2$ is the sample variance for the AR(p) model, i.e.

$$\hat{\sigma}_p^2 = \sum_{j=0}^{p} \hat{a}_p(j) \mathbf{R}_T(j), \tag{155}$$

with $\mathbf{R}_T(j) = T^{-1} \sum_{t=1}^{T-j} x_t x_{t+j}$ the sample covariance function; or alternative, its unbiased estimator,

$$\hat{\tilde{\sigma}}_p^2 = \frac{T}{T-p} \hat{\sigma}_p^2. \tag{156}$$

Definition 49. *The CAT was proposed as an estimator of $E\left[J(p)\right]$. Its approximate estimate is given by:*

$$CAT(p) = 1 - \frac{\hat{\sigma}_{\infty}^2}{\tilde{\sigma}_p^2} + \frac{p}{T}. \tag{157}$$

To cope with lacunae while deriving $CAT(p)$, Bhansali [39] proposed to modify the penalty function in a general manner.

Definition 50. *The approximate estimate of the modification of the $CAT(p)$ is given by:*

$$CAT_\alpha(p) = 1 - \frac{\hat{\sigma}_\infty^2}{\tilde{\sigma}_p^2} + \alpha\frac{p}{T}, \tag{158}$$

where $\alpha = \alpha' + 1$.

If $\alpha = 2$, the asymptotic distribution of the estimate of p is equivalent to that of the FPE and AIC. If α varies, then it is the same of the FPE_β.

3.2.4 *ARMA models*

Direct extensions of IC and PC to general ARMA models were initially related to developments in the likelihood approach to time series models. These included modified likelihood functions or computational approaches. The detailed survey of MS for ARMA models, proposed in 1985 by de Gooijer et al. [125], departed from the theory of statistical hypothesis testing. Then, it was followed by methods from deterministic or stochastic realization theory which do not require prior model fitting, such as those using inverse autocorrelation and partial autocorrelation functions. They also surveyed techniques based on the one-step ahead prediction, e.g FPE, CV, CAT, further IC, and Bayesian methods. These included:

(i) Schwarz's BIC criterion [258];

(ii) Akaike's [10, 11, 13] and Rissanen's [241] BIC criterion; and

(iii) the Bayesian Estimation Criterion of Geweke and Meese [123], or the HQ criterion [137].

In 1992, Choi [79] published a monograph further updating the overview of ARMA model identification. In the following, we will introduce IC and PC for ARMA models developed under different settings.

An autoregressive moving-average of order (*p*,*q*), ARMA(*p*,*q*), is defined as linear combination of an AR(q) with a moving-average MA(q):

$$\begin{aligned} x_t = {} & \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_p x_{t-p} \\ & + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \cdots + \alpha_q \varepsilon_{t-q} + \varepsilon_t, \end{aligned}$$

with $\{\varepsilon_t\}$ a white noise process, i.e. $E[\varepsilon_t] = 0, E[\varepsilon_t^2] = \sigma^2, E[x_t x_j] = 0$, for $t \neq j$, where the zero-mean *i.i.d* innovations and finite variance is a special case. Let $\beta(B)$ and $\alpha(B)$ be the p^{th} and q^{th} degree polynomials:

$$\begin{aligned} \beta(B) &= 1 - \beta_1 z - \cdots - \beta_p z^p, \\ \alpha(B) &= 1 + \alpha_1 z + \cdots + \alpha_q z^q, \end{aligned}$$

where B is defined as the backward shift operator: $B^j x_t = x_{t-j}$, with $j = \{0, \pm 1, \pm 2, \dots\}$.

Definition 51 (Causality and invertibility [56]). *An ARMA(p, q) process, with $t = \{0 \pm 1, \dots\}$,*

$$\beta(B)x_t = \alpha(B)\varepsilon_t$$

is said to be a causal function *of the process $\{\varepsilon_t\}$ if there exists $\{\phi_j\}$ such that $\sum_{j=0}^{\infty} |\phi_j| < \infty$, and:*

$$x_t = \sum_{j=0}^{\infty} \phi_j \varepsilon_{t-j}.$$

The process is said to be invertible *if there exists $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$, and:*

$$\varepsilon_t = \sum_{j=0}^{\infty} \pi_j x_{t-j}.$$

Equivalently, we say that the process $\{x_t\}$ is causal if it is derived from the application of a causal linear filter to $\{\epsilon_t\}$. Note that both causality and invertibility involve both processes $\{x_t\}$ and $\{\epsilon_t\}$.

Following Hannan [135], let the stationary, ergodic, zero-mean with finite variance $\sigma^2 > 0$ process $\{x_t\}$ be generated by:

$$\sum_{j=0}^{p} \beta_j x_{t-j} = \sum_{j=0}^{q} \alpha_j \varepsilon_{t-j}, \quad \beta_0 = \alpha_0 = 1, \text{ with} \tag{159}$$

$$g(z) = \sum_{j=0}^{\infty} \beta_j z^j \neq 0, \ |z| \leq 1, \tag{160}$$

$$h(z) = \sum_{j=0}^{\infty} \alpha_j z^j \neq 0, \ |z| \leq 1; \tag{161}$$

where $g(z)$ and $h(z)$ are co-prime. Notice that Eq.s (160) and (161) ensure causality and invertibility of the process. See [56, Theorems 3.1.1 and 3.1.2]. Let $\hat{\sigma}^2_{p,q}$ be the maximum likelihood estimate of σ^2, computed when lags p and q are being considered. Then:

Definition 52. *The approximate estimate of the AIC for an ARMA(p, q) model is given by:*

$$\mathrm{AIC}(p, q) = n \log \hat{\sigma}^2 + 2(p + q). \tag{162}$$

Note that, since the AIC is inconsistent for AR models, this also transfers to ARMA models. Specifically, it tends to overestimate the true orders p_0 and q_0 [135].

In his seminal contribution, Rissanen [241] obtained the BIC for ARMA (p, q) models following the MDL principle. Following Hannan [134] and letting $\hat{\sigma}^2_{p,q}$ be defined as before:

Definition 53. *The approximate estimate of the BIC for ARMA(p,q) models is given by:*

$$\mathrm{BIC}(p,q) = \log \hat{\sigma}^2_{p,q} + (p+q)\frac{\log n}{n}. \tag{163}$$

Fan and Yao [110], building upon the bias correction from Hurvich and Tsay [157], seen in Section 2.6, considered the Corrected A Information Criterion (AICC) for ARMA models with Gaussian likelihood under correct specification. The AIC tends to overestimate orders p and q, while the AICC [157] favours parsimony with a larger penalty for large values. Define $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$, and σ^2 as the parameters of a causal and invertible Gaussian ARMA(p,q) process, and obtain its MLE:

$$\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2\right) = \underset{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\mathcal{B},\sigma^2>0}{\operatorname{argmin}} L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2), \tag{164}$$

where

$$\mathcal{B} = \{(\boldsymbol{\beta}, \boldsymbol{\alpha}) : b(z)a(z) \neq 0 \; \forall |z| \leq 1\}, \tag{165}$$

i.e. causal and invertible. Denote with

$$S(\boldsymbol{\beta}, \boldsymbol{\alpha}) = r_{j-1}^{-1} \sum_{j=1}^{T} (x_j - \hat{x}_j)^2, \tag{166}$$

the sum of squares of the predictive errors $(x_j - \hat{x}_j)$ regularized by $r_j = \frac{v_t}{\sigma^2}$, where $\{v_t\}$ is the variance of the predictive error (which can be computed recursively). Let $\hat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}})}{T}$ be the MLE estimate of σ^2, $S(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ be the $S(\boldsymbol{\beta}, \boldsymbol{\alpha})$ computed at the MLE parameters $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$, and T the sample size. Then, a part from a constant:

Definition 54. *The approximate estimate for the corrected AIC for ARMA(p,q) model is given by:*

$$AICC_{(p,q)} = -2\log\left\{L\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}, S(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})/T\right)\right\} + \frac{2(p+q+1)T}{T-p-q-2}. \tag{167}$$

Hannan [134] considered a stationary process $\{x_t\}$ generated by an ARMA (p, q) such that:

$$\sum_{j=0}^{p} \beta_j x_{t-j} = \sum_{j=0}^{q} \alpha_j \varepsilon_{t-j},$$

with $E[\varepsilon_t] = 0$, $E[\varepsilon_t \varepsilon_s] = \delta_{t,s}\sigma^2$, and $\beta_0 = \alpha_0 = 1$. Stationarity is implied by the required causality of the process $\{x_t\}$, i.e. the AR and MA polynomials are such that $g(z) = \sum_{j=0}^{p} \beta_j z^j \neq 0$, and $h(z) = \sum_{j=0}^{q} \alpha_j z^j \neq 0$, with $|z| \leq 1$, and both $g(\cdot)$, $h(\cdot)$ with no common

zero. See [56, p. 82-83] for technical details. This delivers the MA(∞) representation:

$$x_t = \sum_{j=0}^{\infty} \kappa_j \varepsilon_{t-j}, \tag{168}$$

with its polynomial $k(z) = \sum_{j=0}^{\infty} \kappa_j z^j = g^{-1}(h(z))$, where κ_j geometrically decreases to zero, and ε_t are linear innovations. The goal is to estimate the true order (p_0, q_0). The estimated parameters $(\hat{\beta}_j, \hat{\alpha}_j)$ are obtained via MLE without the normality assumption. Define the σ-algebra $\mathcal{F}_t = \sigma(\varepsilon_j)$, $j \leq t$, and assume the following technical conditions:

(i) $E[\varepsilon_t | \mathcal{F}_{t-1}] = 0$, $E[\varepsilon_t^2 | \mathcal{F}_{t-1}] = \sigma^2$, $E[\varepsilon_t^4] < \infty$;

(ii) $p_0 \leq P, q_0 \leq Q$, with P, Q known a priori.

Hannan considers the estimates of p_0 and q_0 based on the maximization of the Gaussian likelihood for the conditional error variance σ^2, namely $\hat{\sigma}^2_{p,q}$, without maintaining the Gaussian assumptions and only requiring conditions (i) and (ii) above. The following result was obtained again by using the Law of Iterated Logarithms (LIL).

Definition 55. *To estimate (p_0, q_0), Hannan [134] proposed the minimization of the following approximate estimate of the criterion:*

$$\phi(p, q) = \log \hat{\sigma}^2_{p,q} + (p + q) c \frac{1}{N} \log \log N, \; c > 2. \tag{169}$$

If the disturbances $\{\varepsilon_t\}$ are independent, then this criterion is *strongly consistent*, while if the last term in $\phi(p, q)$ is replaced by $(p + q)C_N / N$, $C_N \to \infty$ then it is *weakly consistent*. Hannan and Rissanen [138] latter modified this criterion by the substitution of the MLE of $\hat{\sigma}^2_{p,q}$ by an alternative estimator from a series of autoregressions proposing a recursive estimation procedure of (p_0, q_0) in three steps for efficient computation. Poskitt [223] further proposed a modification of this procedure to avoid the bias created by the BIC used in the second step, by the use of the Model Determination Criterion [224] obtained by Bayesian arguments.

Stressing the issues of identifiability of ARMA models, Zhang and Wang [354] proposed the order determination quantity (ODQ). Define the backwards operator $B^k y_t = y_{t-k}$, and consider the ARMA model: $\Phi(B) y_t = \Psi(B) \varepsilon_t$, with $\Phi(B) = 1 - \sum_{j=1}^{p_0} \phi_j B^j$, $\Psi(B) = \sum_{j=0}^{q_0} \psi_j B^j$, $\{\varepsilon_t\}$ the unobservable random errors sequence, (p_0, q_0) the unknown true order such that both ϕ_{p_0} and ψ_{q_0} are not null, and $\{y_t\}$ the sequence of observations. The following assumptions are required:

(i) $\Phi(B)$ and $\Psi(B)$ have no common factor so that both polynomials are unique;

(ii) ε_t is a martingale difference sequence $\mathcal{F}_t$-measurable with:

$$E\left[\varepsilon_t|\mathcal{F}_{t-1}\right] = 0 \quad a.s., \forall\, t \geq 1,$$

i.e. $\{\varepsilon_t\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{F}_t\}$;

(iii) y_t is $\mathcal{F}_t$-measurable for every $t \geq 0$ a.s.;

(iv) (p^*, q^*) fixed upper limit known a priori;

(v) there is a random constant a_n s.t. almost surely we have:

$$a_n/n \to 0, \qquad a_n/(\log n)^\beta \to \infty,$$

e.g. $a_n = n^\delta, \delta \in (0,1)$, with $\beta = 1$ for AR models and $\beta \geq 1$ for general ARMA models.

Let n be the sample size, $\hat{\sigma}_n^2(p,q)$ be the estimated variance of the error term ε_t computed at both (p,q), and $\hat{\sigma}_n^2(p^*,q^*)$ be the the estimated variance computed at (p^*,q^*). Then:

Definition 56. *To determine the order of an ARMA(p,q), Zhang and Wang [354] defined the approximate estimate of ODQ as:*

$$\mathrm{ODQ}_n = n\hat{\sigma}_n^2(p,q) - n\hat{\sigma}_n^2(p^*,q^*) - a_n. \tag{170}$$

The ODQ is consistent for unstable autoregressive models, i.e. ARMA (p,0) where all the roots of the characteristic polynomial are either on or inside the unit circle.

This area of research proposed further generalizations. For instance, Peña and Sánchez [220] proposed a validation procedure for h-step ahead forecast. By the use of a filtered version of the in-sample prediction errors, they showed that the procedure is equivalent to an efficient MS method. Recently, Diop and Kengne [101] studied inference and model selection in a general class of causal processes with exogenous covariates, including ARMA-GARCH, APARCH, ARMAX, GARCH-X and APARCH-X. By the use of Lipschitz-type conditions, they showed the existence of a stationary solution, studied consistency of the QML estimator (QMLE) of the parameters, established its asymptotic distribution, Wald-type tests for significance (also for non-stationary cases), and proposed a penalized criterion and conditions for weak and strong consistency, showing that the HQ with a specific regularization parameter is strongly consistent in large samples.

The following paragraphs will detail Rissanen's APE and its extensions in stochastic regression, given its applications to ARMA models.

3.2.4.1 *APE in stochastic regression and ARMA models*

Rissanen [243] proposed the APE, a one-step ahead prediction-error-based consistent cumulated measure to estimate the order of ARMA processes for both large and small samples.[6] This criterion is also called Predictive Least Squares (PLS) and has theoretical groundings on Rissanen's Minimum Description Length (MDL) principle [241].

Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be an observed sample. At every time $t = \{0, 1, \ldots, n-1\}$ we have the past sequence $\mathbf{x}^t = \{x_0 x_1 \cdots x_t\}$, setting $x_0 = 0$. Interest lays on forecasting x_{t+1} at each instant t, given sequence $\mathbf{x}^t$. Consider to model x_t as a zero-mean stationary ARMA(p,q),

$$x_t = \sum_{j=1}^{p} a_j x_{t-j} + \sum_{s=1}^{q} b_j \varepsilon_{t-s} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is a zero-mean uncorrelated process, such that the first two moments of process $\{x_t\}$ are defined by the $k(= p + q)$ parameters composing $\boldsymbol{\theta} = (a_1, \ldots, a_p, b_1, \ldots, b_q)$, and $E[\varepsilon_t^2] = \sigma^2$. Consider a linear predictor, with prediction error $\varepsilon_{t+1} = x_{t+1} - \hat{x}_{t+1}$ given by $\varepsilon_{t+1} = e_{t+1,t}$ determined by both the k parameters $\boldsymbol{\theta}(t) = (a_{1,t}, \ldots, a_{p,t}, b_{1,t}, \ldots, b_{q,t})$ and the data in the following manner:

$$x_i = \sum_{j=1}^{p} a_{j,t} x_{i-j} + \sum_{s=1}^{q} b_{s,t} e_{i-s,t} + e_{i,t},$$

where $i = \{0, \ldots, t+1\}$, setting both x_i and $e_{i,t}$ equal to zero for $i \leq 0$. The parameter vector $\boldsymbol{\theta}(t)$ is obtained by minimizing:

$$S^2(t) = t^{-1} \sum_{0}^{t-1} e_{i+1,t}^2.$$

Definition 57. *Rissanen [243] consistently estimated the order of an ARMA (p,q) model by minimizing the following criterion:*

$$APE(k, x) = n^{-1} \sum_{t=0}^{n-1} \left(x_{t+1} - \hat{x}_{t+1}\right)^2, \tag{171}$$

where $\hat{x}_{t+1}$ is a prediction at time t based on the past sequence $\mathbf{x}^t$.

Wei [328] provided an interpretation of the APE in terms of goodness-of-fit plus penalization for model's complexity, as in IC and PC, for *stochastic regression*. The latter includes as special cases: multiple regression, *time series* models, dynamic input-output systems, adaptive stochastic approximation schemes, and stochastic control. Consider model M with design vector $\mathbf{x}_i$, defined by:

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i,$$

6 For its strong consistency for finite AR processes, see Hemerly and Davis [146].

where ε_i are *i.i.d.* with $\varepsilon_i \sim N(0, \sigma^2)$, $\mathbf{x}_i$ such that it is $\sigma(\varepsilon_1, \ldots, \varepsilon_{i-1})$-measurable, and conditional Fisher information matrix for β equal to:

$$\sigma^{-2} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top,$$

so $\det|\sigma^{-2} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top|$ can be interpreted as the amount of information about parameters' vector β. Let $\hat{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, be the estimated sample variance from the sample of dimension n with model M, be $\tilde{\sigma}_n^2$ the estimated sample variance based on the full model. Then:

Definition 58. *In the context of stochastic regression, Wei [328] proposed the Fisher information criterion (FIC), with its approximate estimate given by:*

$$FIC(M) = n\hat{\sigma}_n^2 + \tilde{\sigma}_n^2 \log\det \left| \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right| \tag{172}$$

Eq. (172) is based on the Fisher information matrix and departs from considerations on the APE for stochastic regression, allowing the penalty term to be proportional to the logarithm of the statistical information contained in a model M with design vector $\mathbf{x}_i$. The connection between PLS (APE) and $FIC(M)$ can be seen noting that, by [80], Theorem 2.1 [328, p. 4], and model's correct specification:

$$PLS \sim n\hat{\sigma}_n^2 + \sigma^2 \log\det \left| \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \right|, \tag{173}$$

thus, if we substitute σ^2 with $\tilde{\sigma}_n^2$, we obtain the FIC. Under model's misspecification, the PLS has an extra penalty, at a cost of issues regarding computation, tendency to select models with fewer variables for small sample, and dependency on data's order. Instead, the FIC is permutation invariant, solves some problems present with the PLS, and features strong consistency.

Lai and Lee [180] further extended the APE and FIC to consider general stochastic regression models, a broad class which includes both *ARMA models* and *nonlinear AR models with exogenous regressors*, and showed their strong consistency regularity conditions. Consider observation y_t and model it as $y_t = g_t(\boldsymbol{\theta}) + \varepsilon_t$, where $\boldsymbol{\theta}$ is the unknown parameters' vector, and $g_t(\boldsymbol{\theta})$ is a twice continuously differentiable $\mathcal{F}_{t-1}$-measurable function.

Assumptions 5. *To define the asymptotic estimate of the APE, assume the following:*

(i) $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^\kappa$, *with* $\boldsymbol{\Theta}$ *compact set;*

(ii) $\{\varepsilon_t\}$ *is a martingale difference sequence with respect to an increasing sequence of* σ*-fields* $\mathcal{F}_t$ *such that:*

$$\sup_t E\left[|\varepsilon_t|^r \mid \mathcal{F}_{t-1}\right] < \infty \; a.s., \; r > 2; \tag{174}$$

(iii) the dimension of $\boldsymbol{\theta}$ is unknown and a family

$$\{g_{t,k}(\lambda) : k \leq \kappa, t \geq 1, \lambda \in \boldsymbol{\Theta}_k\}$$

of regression functions such that $g_{t,k}(\lambda)$ is $\mathcal{F}_{t-1}$ measurable $\forall\, \lambda \in \boldsymbol{\Theta}_k$, with $\boldsymbol{\Theta}_k$ not necessarily subvectors of $\boldsymbol{\Theta}_{k+1}$ (not nested / hierarchical);

(iv) $\boldsymbol{\Theta}_k \in \mathbb{R}^{d(k)}$, $1 \leq k \leq \kappa$, with $d(k)$ positive integers;

(v) $\exists\, \kappa : \boldsymbol{\theta} \in \boldsymbol{\Theta}_\kappa^{int}$ such that:

(a) if $\boldsymbol{\theta}$ is a subvector of some $\boldsymbol{\theta}^{(k)} \in \boldsymbol{\Theta}_k$, then $\boldsymbol{\theta}^{(k)} \in \boldsymbol{\Theta}_k^{int}$ and $g_{t,k}(\boldsymbol{\theta}^{(k)}) = g_{t,\kappa}(\boldsymbol{\theta})$;

(b) if $\boldsymbol{\theta}$ is not a subvector of any $\lambda \in \boldsymbol{\Theta}_k$, then $g_{t,k}(\lambda) \neq g_{t,\kappa}(\boldsymbol{\theta})$, $\forall\, \lambda \in \boldsymbol{\Theta}_k$;

(vi) if $d(k) = d(\kappa)$, $k \neq \kappa$, then $\boldsymbol{\theta}$ is not a subvector of any $\lambda \in \boldsymbol{\Theta}_k$;

(vii) K^* is a known number such that $\boldsymbol{\theta}$ is a subvector of $\boldsymbol{\theta}^{(K^*)} \in \boldsymbol{\Theta}_{K^*}$;

(viii) a prior distribution on σ^2 exists, with technical conditions on the density function.

Also, obtain the not necessarily unique LSE, i.e.

$$\hat{\boldsymbol{\theta}}_t^{(k)} = \underset{\lambda \in \boldsymbol{\Theta}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \left(y_i - g_{i,k}(\lambda)\right)^2,\ \boldsymbol{\theta}^{(k)} \in \boldsymbol{\Theta}_k. \tag{175}$$

Now, define function $S_n(\lambda) = \sum_{i=1}^{n} \left(y_i - g_{i,k}(\lambda)\right)^2$, with minimum at $\lambda = \hat{\boldsymbol{\theta}}_n^{(k)}$ equal to $\hat{\sigma}_{n,k}^2$. Its first and second derivative are then equal to

$$\bigtriangledown S_n(\lambda) = -2 \sum_{i=1}^{n} \left(y_i - g_{i,k}(\lambda)\right) \bigtriangledown g_{i,k}\left(\lambda\right),$$

and

$$\begin{aligned} &\bigtriangledown^2 S_n(\lambda)/2 \\ &= \sum_{i=1}^{n} \left(\bigtriangledown g_{i,k}\left(\lambda\right)\right) \left(\bigtriangledown g_{i,k}\left(\lambda\right)\right)^\top - \sum_{i=1}^{n} \left(y_i - g_{i,k}(\lambda)\right) \bigtriangledown^2 g_{i,k}\left(\lambda\right), \end{aligned}$$

where $\bigtriangledown g_{i,k}\left(\lambda\right)$ and $\bigtriangledown^2 g_{i,k}\left(\lambda\right)$ are the gradient vector and Hessian matrix of $g(\cdot)$ at parameter λ.

Lai and Lee [180] considered a natural extension of the APE for general stochastic regression models. Let m be the fixed initial sample size. To estimate the correct order κ, minimize over all candidates models k, i.e. $\hat{\kappa}_n = \operatorname{argmin}_{1 \leq k \leq K}\ APE(k)$, which may not be unique. Further regularity conditions are necessary for its uniqueness and strong consistency. See Theorem 1 and its Corollary in [180].

Definition 59. *If $\boldsymbol{\theta}$ is a subvector of $\boldsymbol{\theta}^{(k)}$, then the APE for model k is:*

$$APE(k) = \sum_{i=m}^{n} \left\{ y_i - g_{i,k}\left(\hat{\boldsymbol{\theta}}_{i-1}^{(k)}\right) \right\}^2, \tag{176}$$

Now, let $\hat{\sigma}^2_{n,k} = n^{-}1 \sum_{i=1}^{n} \left[y_i - g_{i,k}\left(\hat{\boldsymbol{\theta}}_n^{(k)} \right) \right]^2$ be the estimated variance of ε_i for model k, $\hat{\sigma}^2_{n,K^*}$ be the estimated variance of ε_i for model K^*, and notice that the argument of the logarithm is the sum over all observations of the square of gradient vector of $g(\cdot)$ evaluated at the LSE, i.e. $\bigtriangledown g_{i,k}\left(\hat{\boldsymbol{\theta}}_n^{(k)} \right)$:

Definition 60. *Lai and Lee [180] extended the FIC for general stochastic regression models:*

$$FIC(k) = n\hat{\sigma}^2_{n,k} + \hat{\sigma}^2_{n,K^*} \log \left| \sum_{i=1}^{n} \left(\bigtriangledown g_{i,k}\left(\hat{\boldsymbol{\theta}}_n^{(k)} \right) \right) \left(\bigtriangledown g_{i,k}\left(\hat{\boldsymbol{\theta}}_n^{(k)} \right) \right)^T \right|. \quad (177)$$

See Lai and Yuan [181] for a very recent review of stochastic approximation, field that includes many of the developments in stochastic regression.

3.2.5 *Multivariate time series models*

Multivariate time series arise naturally when considering multiple simultaneous univariate time series, often equally separated in time. The works of Whittle [332], Hannan [133], Reinsel [238], and Lütkepohl [197] are solid grounds for its study. Initial tools derived from extensions of univariate time series, multivariate regression problems, and multivariate prediction and probability theory. For instance on the former, Akaike [3] extended the FPE to multidimensional vector of input and output variables, in the context of vector AR fitting for control, which he called the MFPE. For multivariate regression problems, Mallows [204] considered the $\boldsymbol{C}_p$ also for cases with multiple responses dependent variable in terms of ridge regression, taking a suitable norm or trace as measure of the dimension of the matrix. The first works on the latter trace back to Wiener and Masani [206, 334, 335], and Helson and Lowdenslager [145]. We follow Brockwell et al. [56], Reinsel [238], Lütkepohl [197], and Tsay [306] in the successive paragraphs. We introduce these concepts given the example in Chapter 5, Section 5.4, and the current research work in proximity of the main topic.

Let $\{\mathbf{y}_t\} \in \mathbb{R}^w$ be a stochastic column vector process with mean vector $E[\mathbf{y}_t] = \boldsymbol{\mu}_{y_t} \in \mathbb{R}^w$, j-th lag cross-covariance matrix:

$$E\left[\left(\mathbf{y}_t - \boldsymbol{\mu}_{y_t} \right) \left(\mathbf{y}_{t-j} - \boldsymbol{\mu}_{y_{t-j}} \right)^{\top} \right] = \boldsymbol{\Sigma}_y(j) \in \mathbb{R}^{w \times w}, \quad (178)$$

with $j = \{0, 1, \dots\}$. If the mean vector and the cross-covariance at lag j, i.e $\boldsymbol{\mu}_y$ and $\boldsymbol{\Sigma}_y(j)$ respectively, are independent of time t, then the vector process $\mathbf{y}_t$ is said covariance or weakly stationary.

A VECTOR AUTOREGRESSIVE MOVING-AVERAGE MODEL of orders p and q, denoted VARMA (p, q) or MARMA (p, q), is the multivariate version of the ARMA(p, q), where the response variable is multivariate. Its standard representation is:

$$\mathbf{y}_t = \boldsymbol{\phi}_0 + \sum_{j=1}^{p} \boldsymbol{\phi}_j \mathbf{y}_{t-j} + \boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \boldsymbol{\theta}_j \boldsymbol{\varepsilon}_{t-j},$$

with p, q non-negative integers, $\boldsymbol{\phi}_0$ a w-dimensional constant vector, $\boldsymbol{\phi}_j$ and $\boldsymbol{\theta}_j$ constant $(w \times w)$ matrices, $\{\boldsymbol{\varepsilon}_t\}$ a sequence of *i.i.d* zero-mean random w-dimensional column vectors with positive-definite covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. With the backshift operator ($B^j \mathbf{y}_t = \mathbf{y}_{t-j}$, $j = 0, 1, \dots$), its compact form is:

$$\boldsymbol{\phi}(B)\, \mathbf{y}_t = \boldsymbol{\phi}_0 + \boldsymbol{\theta}(B)\, \boldsymbol{\varepsilon}_t,$$

with $\boldsymbol{\phi}(B) = \boldsymbol{I}_w - \sum_{j=1}^{p} \boldsymbol{\phi}_j B^j$, and $\boldsymbol{\theta}(B) = \boldsymbol{I}_w - \sum_{j=1}^{q} \boldsymbol{\theta}_j B^j$, and $\boldsymbol{I}_w$ a w-dimensional unitary matrix.

Assumptions 6. *To obtain the VARMA(p,q) in final equation form, assume that:*

1. $\{\mathbf{y}_t\}$ *has constant zero-mean, i.e.* $\boldsymbol{\phi}_0 = \mathbf{0}$*;*
2. $\boldsymbol{\phi}(B)$ *and* $\boldsymbol{\theta}(B)$ *are left-coprime (without common factors in matrix setting); and*
3. $\boldsymbol{\phi}(B) = \alpha(B)\boldsymbol{I}_w$*, with* $\alpha(B) = 1 - \sum_{j=1}^{p} \alpha_j B^j \in \mathbb{R}$, $\alpha_p \neq 0$.

The *structural form* and the *echelon form* are also considered in the literature. The former is a generalization of the VARMA model in *compact form*, setting $\boldsymbol{\phi}(B) = \sum_{j=0}^{p} \boldsymbol{\phi}_j B^j$, and $\boldsymbol{\theta}(B) = \sum_{j=0}^{q} \boldsymbol{\theta}_j B^j$. The latter is a useful representation of the VARMA in *structural form*, sometimes indicated as VARMA$_E$, where both the vector AR and MA polynomials, $\boldsymbol{\phi}(B)$ and $\boldsymbol{\theta}(B)$ are such that $\boldsymbol{\phi}(B) = [\phi_{j,i}(B)]$, and $\boldsymbol{\theta}(B) = [\theta_{j,i}(B)]$, with $j, i = 1, \dots, W$ are left-coprime and with the following form:

$$\begin{aligned}
\phi_{j,j}(B) &= 1 - \sum_{s=1}^{p_j} \phi_{jj,s} B^s, \quad \text{for } j = 1, \dots, w; \\
\phi_{j,i}(B) &= - \sum_{s=p_j - p_{ji}+1}^{p_j} \phi_{ji,s} B^s, \quad \text{for } j \neq i; \\
\theta_{j,i}(B) &= \sum_{s=0}^{p_j} \theta_{ji,s} B^s, \quad \text{for } j, i = 1, \dots, w,
\end{aligned}$$

with $\boldsymbol{\phi}_0 = \boldsymbol{\theta}_0$, the row degrees $(p_1, \dots, p_w)$ the Kronecker indices, and $\sum_{j=1}^{w} p_j$ the McMillan degree. There is abundant literature also for integrated or cointegrated VARMA models, e.g. Poskitt [225, 226], Kascha and Trenkle [166].

MS IN GENERAL VARMA MODELS requires the solution of non-trivial statistical aspects for inference, e.g. *stability/causality, invertibility, identifiability*. If $\mathbf{y}_t$ is a zero-mean stable (or causal) process, i.e. $\det(|\boldsymbol{\phi}(z)|) \neq 0, \forall\, z \in \mathbb{C} : \|z\| \leq 1$, then its infinite vector moving-average representation follows, VMA(∞), i.e. $\mathbf{y}_t = \sum_{j=0}^{\infty} \boldsymbol{\psi}_j \boldsymbol{u}_{t-j}$, where the new w-dimensional error vector is $\boldsymbol{u}_t = \boldsymbol{\varepsilon}_t - \sum_{j=1}^{q} \boldsymbol{\theta}_j \boldsymbol{\varepsilon}_{t-j}$, and the $(w \times w)$ matrices $\boldsymbol{\psi}_j$ are such that, for $\|z\| \leq 1$, $\boldsymbol{\psi}(z) = [\boldsymbol{\phi}(z)]^{-1} \boldsymbol{\theta}(z)$. If the zero-mean process $\mathbf{y}_t$ is invertible, i.e. $\det(|\boldsymbol{\theta}(z)|) \neq 0, \forall\, z \in \mathbb{C} : \|z\| \leq 1$, then its infinite vector AR representation follows, VAR(∞), i.e. $\boldsymbol{\varepsilon}_t = \sum_{j=0}^{\infty} \boldsymbol{\Lambda}_j \mathbf{y}_{t-j}$, where the $(w \times w)$ matrices $\boldsymbol{\Lambda}_j$ are such that, for $\|z\| \leq 1$, $\boldsymbol{\Lambda}(z) = [\boldsymbol{\theta}(z)]^{-1} \boldsymbol{\phi}(z)$. For details on multivariate time series models, cf. Hannan and Deistler [136], Brockwell and Davis [56, Ch. 11], Reinsel [238], Lütkepohl [197], Tsay [306], and Wei [329].

If $\boldsymbol{\phi}(B)$ and $\boldsymbol{\theta}(B)$ are uniquely identified by the weight matrices in its VMA(∞) representation, then identifiability of the VARMA(p,q) model follows. Sufficient conditions for identifiability (i.e block identifiability) [197] require that:

i. $\boldsymbol{\phi}(B)$ and $\boldsymbol{\theta}(B)$ are left-coprime;

ii. the order q is as small as possible while the p order is as small as possible for that q; and

iii. rank $\left(\left[\boldsymbol{\phi}_p, \boldsymbol{\theta}_q\right]\right) = w$, with $p, q > 0$, where $[\boldsymbol{A}, \boldsymbol{B}]$ refers to the joint matrix composed by matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

For technical details on structural identifiability of the more general VARMA with exogenous regressors, VARMAX(p, q, r), where r is the maximum lag of the exogenous part, see Hannan and Deistler [136], Section 2.7.

ESTIMATION OF VARMA To estimate VARMA(p, q) parameters, quasi, conditional or exact likelihood methods can be employed. The maximization of the log likelihood function with respect to parameters's vector usually involves approximations or *ad-hoc* algorithms which are usually employed given that no closed forms are available (with some exceptions) or are under development, and that identifiability is an issue. For this reason it is still a non-trivial problem. Recent interesting results to solve these issues include exact likelihood estimation of time-dependent models [18] (where the parameters $\boldsymbol{\phi}_t(B)$, and $\boldsymbol{\theta}_t(B)$ are dependent on time t), estimation of causal and invertible VARMA models by constrained estimation [251], semiparametric estimation for VARMA models via R-estimation (therefore expanding further than QMLE) [130], identification and estimation in large-scale settings with VARMA [338], and asymptotic properties of quasi MLE for causal, invertible and identifiable for time-dependent models [208].

We will briefly introduce two perspectives for VARMA model specification and selection, before introducing VAR model selection via IC.

These sections will serve as an overview previous to Section 5.4 in Chapter 5.

3.2.5.1 *Structural specification and model selection in VARMA models*

Structural specification is related to MS particularly for VARMA models. Drawing from the sedimented literature prior to 2005, Lütkepohl [197] indicated in his monograph that both the echelon form and the final form ensure its identifiability (among the class of those representations), relying on the proofs from Hannan and Deistler contributions, condensed in their re-published monograph [136]. Numerous procedures for structural specification are indicated in Lütkepohl's work, not as widespread as the Box-Jenkins approach [51] for univariate ARMA models. He divided it into two groups: those using the final form, and those using the echelon form.

MS IN FINAL FORM VARMA Zellner and Palm [353] proposed a two-steps procedure for VARMA models in final form. In the first, a univariate model is specified for each component $y_{i,t}$, $i = 1, \ldots, w$. The suggestion is to employ Box-Jenkins strategy [51], LR tests, Bayesian posterior odds, or automatic procedures with IC as in Hannan and Rissanen [138] or its extension by Poskitt [223]. In the second stage, a common AR polynomial for the vector response has to be selected as the product of each components' polynomial. Then, select the corresponding MA polynomial for each of the w components. Finally, select the order q which is the maximum order obtained over all the components. If there are common factor, both for the AR and MA polynomials, the polynomial degree may be further reduced.

MS IN ECHELON FORM VARMA This strategy involves the estimation of large numbers of parameters. For this reason, Lütkepohl argued that it is not popular in practice even if it is relatively simple. It becomes a major problem when the dimensionality increases. For these cases, model specification of VARMA in echelon form are more appealing. Tsay's 2013 monograph [306, Ch. 4] coincides with this strategy, indicating two approaches for structural specification of VARMA models. Both were presented in his 1991's article [305]. The first draws from the *Kronecker indexes* and *MacMillan polynomial degrees* literature from control systems in the engineering literature, applied to the model in echelon form. The second is considered as a refinement over the first, involving canonical correlation analysis (viz. Hotelling [152], Akaike [9]), presented in Tsay's 1989's article [297].

The recent contribution by Bhansali [40] combined these approaches. After proposing the first definition of an h-step ahead state-space representation (extending the results in Akaike [17] for $h = 0$, and Cooper and Wood [85] for $h = 1$), to estimate the *Kronecker indexes*, Bhansali

modified the Difference Information Criterion (DIC) originally presented by Akaike [9]. For the latter, let vectors $\boldsymbol{u} \in \mathbb{R}^s$ and $\boldsymbol{v} \in \mathbb{R}^r$, and define the model $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{u} + \boldsymbol{w}$, where $\boldsymbol{A}$ is the regression coefficient matrix of $\boldsymbol{v}$ on $\boldsymbol{u}$, with $\text{rank}(\boldsymbol{A}) = q$, and $\boldsymbol{w}$ has null-correlation with $\boldsymbol{u}$. The number of free parameters, $F(q)$ is the sum of the free parameters within the covariance matrices of $\boldsymbol{u}$ and $\boldsymbol{v}$, and within matrix $\boldsymbol{A}$. Assuming $s \geq r$, these are respectively $s(s+1)/2$, $r(r+1)/2$, and $q(s+r-q)$. Consider N observations from two Gaussian random vectors $\boldsymbol{v} = (v_1, v_2, \ldots, v_s)^\top$ and $\boldsymbol{u} = (u_1, u_2, \ldots, u_s)^\top$, and assume that there are q non-null canonical correlation coefficients.

In the context of canonical correlation analysis, Akaike [9] proposed the DIC as the difference between the model under consideration with $\text{rank}(\boldsymbol{A}) = q$ and the unconstrained model r (without restrictions on matrix $\boldsymbol{A}$).

Definition 61. *The approximate estimate of the $DIC(q)$ is given by:*

$$DIC(q) = AIC(q) - AIC(r), \tag{179}$$

with

$$AIC(q) = N \log \prod_{i=1}^{q} (1 - c_i)^2 + 2F(q), \tag{180}$$

where c_i the i-th largest canonical correlation coefficient. Therefore, it is obtained:

$$DIC(q) = -N \log \prod_{i=q+1}^{r} (1 - c_i)^2 - 2(r-q)(s-q). \tag{181}$$

Bhansali [40] proposed a modification of the DIC in the sense of Bhansali and Downham [41]. Let w be the dimension of the dependent multivariate response time series, $\rho_h(v+1), v \geq 0$ be the next canonical correlation between the vectors of past variables,

$$\boldsymbol{\eta}_t(h, M) = \left[\mathbf{y}^\top(t-h), \ldots, \mathbf{y}^\top(t-M)\right]^\top, \tag{182}$$

and the vector of future variables

$$\boldsymbol{\theta}_t(h, M) = \left[\mathbf{y}^\top(t), \ldots, \mathbf{y}^\top(t+M-h)\right]^\top, \tag{183}$$

truncated up to a large integer M, with $h \in [-H, H]$, where both M, H are carefully subjectively selected, and let $(M - h + 1)$ be the dimension of the $\boldsymbol{\eta}_t(h, M)$, and $v \in \{1, \ldots, M\}$. Then:

Definition 62. *The resulting approximate estimate of the $DIC(q)$ is given by:*

$$\begin{aligned} &DICh_\alpha(v) = \\ &-T \log\left\{1 - [\rho_h(v+1)]^2\right\} - \alpha\left\{(M-h+1)\, w - v\right\}. \end{aligned} \tag{184}$$

IC AND VARMA MODELS In the last twenty years, the literature has tackled different issues for VARMA models related to IC. For instance on cointegration, Kapetanios [165] reviewed the formal grounds for the application of IC to MS for selecting the cointegration rank of cointegrated VARMA models. He underlined that necessary and sufficient conditions for weak consistency of criteria are also valid to determine the cointegration rank. They derived the asymptotic distribution of the estimated cointegration rank when selected via the AIC, and showed that it has an upward bias also for large samples. The advice was for the use of the BIC or the Posterior Information Criterion by Phillips [222] instead. This criterion is Bayesian in its spirit and, among other features, is valid for order selection of cointegrating rank, lag length, and trend degree in VAR models. Boubacar Mainassara [49] studied selection of weak VARMA models by a modified AIC using the QMLE. Weak VARMA models are cases where the error vector is a weak white noise, i.e. stationary sequence of zero-mean and uncorrelated processes with invertible variance matrix, whereas strong VARMA are defined with strong white noise error vector, i.e. *i.i.d.*. More recently, Chan et al. [69] developed a Bayesian approach for inference in VARMAs ensuring identification and parsimony in the context of an efficient Markov chain Monte Carlo algorithm, with application to macroeconomics. Fasen and Kimmig [112] proposed a criterion for continuous time VARMA processes, studying the QMLE in the context of misspecification, and deriving regularity conditions for strong and weak consistency of a general IC, with AIC and BIC as special cases.

3.2.5.2 *VAR models*

Historically, because of the identifiability problem, VARMA models not enjoyed vast popularity in applied fields. This paved the way for models which setted $q = 0$, obtaining a vector autoregressive model of order p, VAR(p), which can be written as $\mathbf{y}_t^\top = \mathbf{z}_t^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t^\top$, where the row vector $\mathbf{z}_t^\top = (1, \mathbf{y}_{t-1}^\top, \ldots, \mathbf{y}_{t-p}^\top) \in \mathbb{R}^{(wp+1)}$ and the $((wp+1) \times w)$ coefficients matrix $\boldsymbol{\beta} = [\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p]^\top$. In this way, and using $T - p$ observations from a sample of size T, we can rewrite it as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{E},$$

with $\mathbf{Y} \in \mathbb{R}^{(T-p)\times w}$ the response matrix observed from time $t = \{p+1, \ldots, T\}$, $\mathbf{Z} \in \mathbb{R}^{(T-p)\times(wp+1)}$ the design matrix, and $\boldsymbol{E} \in \mathbb{R}^{(T-p)\times w}$ the error matrix. The generalized least squares estimator (GLSE) of VAR(p) parameters can be obtained as:

$$\operatorname{vec}(\hat{\boldsymbol{\beta}}) = \operatorname{vec}\left[\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\left(\mathbf{Z}^\top\mathbf{Y}\right)\right], \tag{185}$$

delivering

$$\hat{\boldsymbol{\beta}} = \left[\sum_{t=p+1}^{T} \mathbf{z}_t\mathbf{z}_t^\top\right]^{-1}\left[\sum_{t=p+1}^{T} \mathbf{z}_t\mathbf{y}_t^\top\right], \tag{186}$$

which is identical to the Ordinary least squares (OLS) estimator. If we further assume that the vector error $\boldsymbol{\varepsilon}_t$ is multivariate Gaussian, via the conditional likelihood function it can be shown that it also coincides with the MLE. If we further let $\{\mathbf{x}_t\} \in \mathbb{R}^m$ be a column vector of exogenous variables or leading indicators, with $E[\mathbf{x}_t] = \boldsymbol{\mu}_x \in \mathbb{R}^m$, and $V[\mathbf{x}_t] = \boldsymbol{\Sigma}_x \in \mathbb{R}^{m \times m}$, then a VAR with exogeneous variables of order (p, s), VARX(p, s) is equal to:

$$\mathbf{y}_t = \boldsymbol{\phi}_0 + \sum_{j=1}^{p} \boldsymbol{\phi}_j \mathbf{y}_{t-j} + \sum_{j=0}^{s} \boldsymbol{\alpha}_j \mathbf{x}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{187}$$

where s is a non-negative integer, $\boldsymbol{\alpha}$ a $(w \times m)$ constant matrices.

Quinn [230] extended results from scalar (AR) to multivariate (VAR) models assuming: $E[\boldsymbol{\varepsilon}_t | \mathcal{F}_{t-1}] = 0$, $E\left[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\top | \mathcal{F}_{t-1}\right] = \boldsymbol{\Sigma}_\varepsilon$, $E\left[\varepsilon_{i,t}^4\right] < \infty$, with $i = 1, \ldots, w$, where $\mathcal{F}_{t-1} = \sigma(\boldsymbol{\varepsilon}_{t-1}, \boldsymbol{\varepsilon}_{t-2}, \ldots)$. Let $\det|\hat{\boldsymbol{\Sigma}}_{\varepsilon,k}|$ be the determinant of the MLE of $\boldsymbol{\Sigma}_\varepsilon$ for a VAR(k), i.e.

$$\hat{\boldsymbol{\Sigma}}_{\varepsilon,k} = \frac{1}{T-k} \hat{\boldsymbol{E}}^\top \hat{\boldsymbol{E}} = \frac{1}{T-k} \sum_{t=k+1}^{T} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t^\top, \tag{188}$$

where T is the total number of observations, k is the number of lags, w the dimension of the dependent vector variable, and $\hat{\boldsymbol{\varepsilon}}_t$ is the residual vector. Then:

Definition 63. *Quinn [230] indicated one version of the AIC and showed that the HQ with multivariate response is strongly consistent for VAR models:*

$$AIC(k) = \log\left(\det|\hat{\boldsymbol{\Sigma}}_{\varepsilon,k}|\right) + 2kw^2 \frac{1}{T}, \tag{189}$$

$$\phi(k) = \log\left(\det|\hat{\boldsymbol{\Sigma}}_{\varepsilon,k}|\right) + 2kw^2 \frac{1}{T} \log\log T. \tag{190}$$

Let $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be generated by a w-dimensional zero-mean VAR(p_0) process: $\mathbf{y}_t^\top = \sum_{j=1}^{p_0} \mathbf{y}_{t-j}^\top \boldsymbol{\phi}_j^\top + \boldsymbol{\varepsilon}_t^\top$, with $t = \{1, \ldots, n\}$ and $\mathbf{y}_t^\top = (y_{1,t}, \ldots, y_{w,t})$, $\boldsymbol{\phi}_j^\top$ is a $(w \times w)$ coefficients' matrix, and $\boldsymbol{\varepsilon}_t$ are zero-mean Gaussian i.i.d. with covariance matrix $\boldsymbol{\Sigma}_0$. We write the true AR(p_0) model for the full sample as: $\mathbf{Y} = \mathbf{X}_0 \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ in matrix terms. The following representation from Eqs. (7) and (8) in Hurvich and Tsai [158] highlights the correction obtained by avoiding the first-order Taylor series approximation:

Denote with $b = n^{-1}\{n - (pw + w + 1)\}$ a scale factor for the complexity penalty term, n is the sample size, and

$$\hat{\boldsymbol{\Sigma}} = n^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \tag{191}$$

with

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\left(\mathbf{X}^\top \mathbf{Y}\right), \tag{192}$$

the conditional LSEs of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\beta}$ respectively. Then:

Definition 64. *Hurvich and Tsai [158] proposed the approximate estimates for both the AIC and its corrected version for VAR model selection:*

$$AIC(p) = n\left(\log\left(\det\left|\hat{\mathbf{\Sigma}}\right|\right) + w\right) + 2\left\{pw^2 + w(w+1)/2\right\}, \quad (193)$$

$$AIC_c(p) = n\left(\log\left(\det\left|\hat{\mathbf{\Sigma}}\right|\right) + w\right) + 2b\left\{pw^2 + w(w+1)/2\right\}. \quad (194)$$

Also in this case, interest has been sustained lately. Qu and Perron [228] proposed a modified AIC for Johansen's cointegration tests and showed that, if applied, these have the same distribution as when the order is finite and known. Ren and Zhang [239] proposed a computationally efficient algorithm, the adaptive lasso, for subset selection and estimation in VAR models. For tuning parameter's selection the consistent BIC was adopted. They also showed that their method satisfies the oracle property, i.e. no prior knowledge is required on the sparsity to obtain an optimal asymptotic convergence rate.[7] They also highlighted that the elastic net as in Zou and Zhang [360] might be a better candidate for subset selection. Bingham [44] reviewed multivariate prediction theory and matrix orthogonal polynomials on the unit circle from a probabilistic standpoint. Lütkepohl and Netšunajev [199] proposed a review of structural VAR models with heteroskedasticity or conditional heteroskedasticity. They found that for lag selection the AIC is a common strategy, but indicated how it may be problematic given that no full likelihood optimization is performed. Recenlty, the FIC for locally misspecified VAR models was extended by Lohmeyer et al. [195].

We now shift our attention to MS in the case of nonparametric regression with nonlinear time series models. Short bibliographic notes are included in Appendix A.2.4, including a selected sequence of developments in modelling nonlinear time series.

3.3 NONPARAMETRIC ANALYSIS OF NONLINEAR TIME SERIES MODELS

Let us consider the 1997's review of MS methods for nonparametric time series by Härdle et al. [142]. Let a specific class of mean functions $\mu(\cdot)$, such that, the specified candidate function be obtained from finite and fixed number of parameters. In that case, the parametric method to estimate the conditional mean function of a time series requires the formulation of a parametric model for the mean function $\mu(\cdot)$.

There are successful cases where there are parsimonious models capturing linearities and nonlinearities of the process under analysis, e.g. Tong's Threshold Autoregressive (TAR) [303], Exponential Autoregressive (EXPAR) [128]; Self-Exciting Threshold Autoregressive (SETAR) [70, 126]. See [304] for further details. The nonparametric analysis of time series instead "*leaves data speak for itself*", avoiding the

7 See Fan and Li [109, p. 1353], Theorem 2.

subjectivity of choosing a specific parametric model before observing data. But there were costs in terms of the increased complexity of the mathematical arguments involved, issues with their interpretability, initial difficulties in practical applications, e.g. bandwidth selection (i.e. smoothing parameter selection), the curse of dimensionality, and computational costs. See Breiman [54], Wasserman [322], Hastie et al. [143], and James et al. [162] for discussions and coverage of these types of methods.

3.3.1 *Asymptotic Final Prediction Error*

The ideas introduced in Auestad and Tjøstheim [26] aimed at identifying nonlinear time series with nonparametric estimates of the conditional mean and the conditional variance. They noted that most of nonlinear models satisfy the assumptions necessary to apply the nonparametric asymptotic theory (see also Robinzonov et al. [249]). Through simulations, they adjusted the conditional quantities, while also using asymptotic arguments for an AR(1) process. By way of further reasoning on the estimates of the conditional mean and conditional variance, they dealt with the problem of identification, obtaining a MS criterion heuristically that is able to manage distortion and misspecification.

Continuing along this line, Tjøstheim and Auestad [300] presented a nonparametric procedure for lag selection of general nonlinear stationary time series. The original derivation was based on β-mixing properties with an exponentially decreasing mixing rate[8], smoothness, and bandwidth rate assumptions. The objective was to select lags that give a "good description" of the conditional mean and conditional variance structure, where the goodness of approximation was measured by the MSPE.[9] In that sense, it is a nonparametric analogue version of Akaike's FPE criterion [1, 2]. Vieu [317] followed a similar path with a nonparametric approach to estimate autoregression order without restriction on the parametric class of processes. They proposed a technique based on the minimization of some prediction error, similar to the AIC criterion. Under homoskedasticity, they showed consistency of the CV approach, a similar result which was also obtained by Yao and Tong [350]. More recently, Manzan [205] studied the finite-sample performance of MS criteria for local linear regression by simulation, showing that the AIC and FPE perform very poorly because they tend to over fit (use too many parameters compared to optimal).

We will follow Tschernig and Yang [310] and Tschernig [309], focusing on the use of the nonparametric FPE in MS to obtain their Asymptotic FPE (AFPE). By asymptotic analysis, they showed asymptotic consistency in their nonparametric FPE allowing for heteroskedasticity,

8 See Appendix A.2.4 for details.
9 See Remark 3.2.2.

and showed that overfitting is more likely than underfitting, suggesting a correction of the nonparametric FPE to reduce overfitting in favour of correct fitting. This strategy was further explained in Tschernig [309]. The result is a nonparametric estimation of univariate nonlinear time series models, covering conditionally heteroskedastic errors and seasonal features. These two works will be summarized with a focus on the methodology for obtaining the desired asymptotic properties. The similarity of their approach to those presented before should be evident to the reader. It is worth noticing that in the formal derivation, Tschernig and Yang studied a different and independent vector of observations than the original, $\{\tilde{y}_t\}$ independent of $\{y_t\}$, but with the same stochastic properties, a strategy that is common in the literature. Appendix A.2.5 shares some notes on the two nonparametric estimators employed.

Consider a nonlinear conditionally heteroskedastic autoregressive (NAR) model generating a univariate stochastic process $\{Y_t\}_{t\geq 0}$:

$$Y_t = \mu(X_t) + \sigma(X_t)\xi_t. \tag{195}$$

Assumptions 7. *The following technical conditions are required:*

(i) $X_t = (Y_{t-i_1}, Y_{t-i_2}, \ldots, Y_{t-i_m})^\top$, *are all the correct lagged values, with* $i_1 < \cdots < i_m$;

(ii) $\{\xi_t\}$ *is sequence of* i.i.d. *random variables, with* $E[\xi_t] = 0$, $E[\xi_t^2] = 1$;

(iii) Let $t = \{i_m, i_m + 1, \ldots\}$ *be the set of times;*

(iv) Let $\mu(\cdot)$ *be the conditional mean function and* $\sigma(\cdot)$ *the conditional volatility.*

Assumptions 8. *The AFPE is defined under the following assumptions:*

(I) All the lags indicated by the indexes $\{i_1, \ldots, i_m\}$ *are necessary to model the conditional mean function* $\mu(\cdot)$, *but these are not the same necessary to model the conditional volatility function* $\sigma(\cdot)$.

(II) The process $X_{M,t} = (Y_{t-1}, \ldots, Y_{t-M})^T$ *is strictly stationary and* β*-mixing,*[10] *with* $\beta(n) \leq c_0 n^{-(2+\delta)/\delta}$, $\delta > 0$, *and* $c_0 > 0$. *Where*

(i) $M \geq i_m$, M *integer*

(ii) $\beta(n) = E[sup\{|P(A|\mathcal{F}_M^k) - P(A)| : A \in \mathcal{F}_{n+k}^\infty\}]$

(iii) $\mathcal{F}_t^{t'}$ *is the* σ*-algebra generated by* $X_{M,t}, X_{M,t+1}, \ldots, X_{M,t'}$

(III) The stationary distribution of the process $X_{M,t}$ *has continuous density* $f_M(x_M)$, $x_M \in \mathbb{R}^M$.

(IV) (a) The conditional mean function $\mu(\cdot)$ *is twice continuously differentiable; (b) The conditional volatility function* $\sigma(\cdot)$ *is continuous and positive on the support of* $f(\cdot)$.

10 See Appendix A.2.4.

(V) *The random variables* $\{\xi_t\}_{t \geq i_m}$ *have a finite fourth moment* m_4.

(VI) *For the weight function* $w(\cdot)$, $w : \mathbb{R}^M \longrightarrow \mathbb{R}$, *assume that it:*

(i) *is continuous;*

(ii) *is nonnegative;*

(iii) *weights the density* $f(x_M) > 0$, *with* x_M *in the support of* (w);

(iv) *has compact support with nonempty interior.*

(VII) *The kernel function* $K : \mathbb{R}^1 \longrightarrow \mathbb{R}^1$ *is a symmetric probability density (kernel) and* $h = h_T$ *is a positive number (bandwidth), with*

$$\begin{cases} Th^m & \longrightarrow +\infty, \\ h & \longrightarrow 0, \end{cases} \tag{196}$$

as $T \longrightarrow \infty$.

ON ASSUMPTIONS Tshernig and Yang refer, to Tweedie's results in Doukhan's lectures' notes [104] (among others) for technical conditions guaranteeing (I) and (II). Specially, Theorem 7 and Remark 7 [104, pp. 102–103]. Condition (III) delivers that $f(\cdot)$ denotes both $f_M(\cdot)$ and all of its marginal densities. If the Nadaraya-Watson (NW) estimator is used, then $f_M(\cdot)$ has to be continuously differentiable. In their Monte Carlo study, linear processes following AR(1), AR(2), AR(3), and nonlinear processes NLAR(1), NLAR(2), NLAR(3) satisfy conditions (I-IV). Instead, the nonlinear process NLAR(4) with triangular errors violates smoothness condition (II) ($\nabla f(x)$ does not exist at some points). Condition (V) ensures consistency for lag selection. At the same time, under assumptions (I-V), it is no longer necessary to generate the process $\{\tilde{Y}_t\}$ to compute the FPE. Furthermore, using a single weight function defined for the largest lag vector $X_{M,t}$ allows for the treatment of both bounded and unbounded time series. There are other authors showing consistency results only for bounded time series. An exception is Vieu [317], which is a special case in the setting of Tschernig and Yang [310]. Condition (VI) ensures that the asymptotic distribution does not: *(a)* collapses to a point, or *(b)* have an asympotic bias increasing infinitely large. Please refer to Appendix A.2.5 for details on the employed Nadaraya-Watson and Local Linear estimators.

MS ALGORITHM In Tschernig and Yang [310] and Tschernig [309], the algorithm for MS requires to define *a priori* a set of possible lag vectors S, to choose the maximal lag M, and to define the full lag vector $x_{t,M} = (y_{t-1}, y_{t-2}, \ldots, y_{t-M})^\top$ (denote by X, but depending on time t and on lag M). It is specified in the simple pseudo-code in Listing 2. As optimality criterion for the elimination of redundant lags, the authors used the MSPE, i.e FPE.

The following definition requires to consider process $\{\tilde{Y}_t\}$ with exactly the same distribution as $\{Y_t\}$, but independent of it:

Listing 2: MS algorithm's pseudo-code

```
Define S
Select M
Define X
Eliminate redundant lags from X
```

Definition 65. *For a given bandwidth h, and lag vector $\{i_1^+, \dots, i_{m^+}^+\}$, define the FPE of an estimate $\hat{\mu}$ of μ as the functional:*

$$\begin{aligned} &FPE(h, i_1^+, \dots, i_{m^+}^+) = E\Big[\{\tilde{y}_t - \hat{\mu}(\tilde{x}_t^+, h)\}^2 w(\tilde{x}_t, M)\Big] \\ &= \int \left[\int \{\tilde{y} - \hat{\mu}(\tilde{x}^+, h)\}^2 w(\tilde{x}_M)\ f(\tilde{y}, \tilde{x}_M)\ d\tilde{y} d\tilde{x}_M \right] \\ &\times f(y_1, \dots, y_T)\ dy_1 \cdots dy_T = FPE(\hat{\mu}). \end{aligned} \tag{197}$$

The outer integral averages over all possible realizations of the estimator $\hat{\mu}(\tilde{x}^+, h)$, and it depends on a given $\tilde{x}^+$, bandwidth h, and sample realizations $\{y_1, \dots, y_T\}$.

LINEAR AND NON LINEAR AR PROCESSES The FPE measures the discrepancy between $\hat{\mu}$ and the true functional relation between $\tilde{y}_t$ and $\tilde{x}_t$, and:

(i) If: *(a)* the process $\{\tilde{y}_t\}$ is a stationary linear AR process, and *(b)* $\hat{\mu}$ is a linear regressor; then the *usual (linear) FPE* follows ([1, 2]).

(ii) If the process $\{\tilde{y}_t\}$ is a stationary nonlinear AR process and $\hat{\mu}$ is a nonparametric estimator, then we obtain the *Nonparametric FPE* (Auestad and Tjøstheim [26], Tjøstheim and Auestad [300]).

If the $FPE(h, i_1^+, \dots, i_{m^+}^+)$ would be observable, then we may select the lag vector and corresponding bandwidth which minimizes the FPE across all lag combinations considered. Since usually the $FPE(h, i_1^+, \dots, i_{m^+}^+)$ is not observable, it is necessary to estimate it. Possible solutions are given by the CV method (as in Vieu [316] or Yao and Tong [350]), or the one pursued here: to find asymptotic expressions of the $FPE(\cdot)$ as in Auestad and Tjøstheim [26], Tjøstheim and Auestad [300], and Tschernig and Yang [310]. By Theorem 2.1 [310, p. 461], the definition and decomposition of the nonparametric AFPE is obtained as in the following definition.
First, write the FPE as in Eq. 197, and let $a = \{1, 2\}$, where $a = 1$ refers to the NW estimator, and $a = 2$ refers to the Local Linear Estimator (LLE). Then:

Definition 66. *Under assumptions (I–VI), for $a = \{1, 2\}$, as $T \longrightarrow \infty$,*

$$\begin{aligned} &FPE_a(h, i_1, \dots, i_m) \\ &= AFPE_a(h, i_1, \dots, i_m) + o(h^4 + (T - i_m)^{-1} h^{-m}), \end{aligned} \tag{198}$$

in which the Asymptotic FPEs are given by:

$$AFPE_a(h, i_1, \ldots, i_m) = A + b(h)B + c(h)C_a, \tag{199}$$

To help with the interpretation, write Eq. (199) as:

$$AFPE_a(h, \text{correct lags}) = \text{I.V.} + \text{E.V.E.} + \text{S.B.E.} \tag{200}$$

where the initials stand for 'Integrated Variance', 'Expected Variance of the Estimator', and 'Squared Bias of the Estimator' respectively .

Definition 67. *The Integrated Variance is equal to the* FPE *of the true function* $\mu(\cdot)$*:*

$$A = \int \sigma^2(x)\ w(x_M)\ f(x_M)\ dx_m = E\left[\sigma^2(x_t) w(x_{t,M})\right]. \tag{201}$$

Given that $b(h)B$ and $c(h)C_a$ tend to zero as the sample size diverges, both the FPE and AFPE tend asymptotically to the integrated variance.

The $b(h)$ quantity depends on bandwidth and kernel constants:

$$b(h) = ||K||_2^{2m}(T - i_m)^{-1}h^{-m}, \tag{202}$$

with $b(h)B$ vanishing asymptotically.

Definition 68. *The Expected Variance of Estimation is equal to:*

$$\begin{aligned} B &= \int \sigma^2(x)\ w(x_M)\ f(x_M)\ \frac{1}{f(x)} dx_m \\ &= E\left[\sigma^2(x_t)\frac{w(x_{t,M})}{f(x_t)}\right]. \end{aligned} \tag{203}$$

Definition 69. *The integrated SBE, in the LLE case (a=2) is equal to:*

$$\begin{aligned} C_{a=2} &= \int \left(tr\left\{ \frac{\partial^2 \mu(x)}{\partial x \partial x'} \right\} \right)^2 w(x_M)\ f(x_M)\ dx_M \\ &= E\left[\left(tr\left\{ \frac{\partial^2 \mu(x)}{\partial x \partial x'} \right\} \right)^2 w(x_M) \right] \end{aligned} \tag{204}$$

where $c(h)$ *depends on the bandwidth and on the kernel constant:*

$$c(h) = \sigma_K^4 \frac{h^4}{4} \tag{205}$$

with $c(h)C_{a=2}$ *vanishing asymptotically.*

Two cases may be distinguished. In the first, all correct lags are included, plus some additional ones. In this case, all corresponding variables can can be indexed with a "+" sign and the modified FPE expansion can be obtained, delivering Theorem 3.3 in Tschernig and Yang [310, p. 464]. In the second case, the relevant lag is left out and we are in an underfitting situation. In this case, the $AFPE(\cdot)$ of the underfitted and the correct model differ by a contant (independent of the bandwidth and sample size). Theorem 3.4 in Tschernig and Yang [310, p. 465] refers to this case.

3.3.2 *Alternative methodologies for nonlinear time series models*

Robinzonov et al. [249] presented the use of *boosting techniques*[11] in the context of time series. By considering a broad class of nonlinear time series, the class of *nonlinear additive autoregressive* model (NAAR), they obtained estimates of the lags of the time series as flexible functions to detect non-monotone relationships between current and past observations. A component-wise boosting algorithm is applied for simultaneous model fitting, variable selection, and model choice, delivering lag selection and dealing with nonlinearity. Its forecasting potential is exemplified with for German industrial production data with additional exogenous variables. Their work assesses the issues of high dimensionality in models: using Exogenous NAAR (NAARX) models they noted how their boosting technique can cope with large models with the number of explanatory variables much larger than the number of observations.

In Zhang and Wu [355], for a general class of nonparametric time series regression model with time-dependent regression function, the authors established an asymptotic theory for estimates of the time-varying regression functions. This work also proposed an information criterion and proved its asymptotic consistency. The empirical part is an application to the U.S. treasury interest rate data.

To face the problem of dimensionality, an example of recent development in this sense is given by Chen et al. [72] with the use of model averaging, where semiparametric methods (i.e. a combination of both parametric and nonparametric techniques) for dimensionality reduction of the possible regressors is proposed, delivering good results in terms of forecast of the dependent variable, and for:

(a) cases when the number of variables is much larger than the sample dimension, $k >> n$, and

(b) factor models.

For an updated introduction to nonlinear time series analysis, see the recent monograph of Tsay and Chen on nonlinear time series analysis [308]. It offers several applications with the open-source R software. In Chapter 3, nonparametric modelling of univariate time series include methods and techniques such as: kernel smoothing, local polynomial, B-splines, smoothing splines; wavelets and thresholding; index models; and sliced inverse regression. All of these, and many others, are included to explore nonlinearity in a time series and to introduce NAAR models. Their work showed how to increase flexibility in modelling the nonlinearity embedded in the data.

11 For details, see Hastie et al. [143].

For a recent exploration on the limits of distribution-free conditional predictive inference and open questions, see Foygel et al. [117].

3.4 HIGH-DIMENSIONAL AND ALGORITHMIC APPROACHES

For a recent systematic review of MS in high-dimensional regression (least squares, logistic, and quantile regression models), see Lee et al. [184]. For a history of subset selection, see Chen et al. [76]. Two key moments are worth mentioning. In 1994, Chen and Donoho [74] defined the goals of Adaptive Representation . These are:

i. speed (computational time in order $\mathrm{O}(n)$ or $\mathrm{O}(n \log n)$);

ii. sparsity (similar to parsimony, fewer coefficients);

iii. perfect separation (clear decomposition of the representation);

iv. stability (resistance to small perturbations).

In 1996, Tibshirani [298] proposed the famous least absolute shrinkage and selection operator (*lasso*) for regression and generalized regressions. The idea behind it is to define a shrinking operation to produce coefficients equal to zero, and it was exemplified while competing with subset selection and ridge regression.

To conclude, we indicate 17 works from 2013 to 2021 of algorithmic approaches applied to high-dimensional settings:

(i) Lee and Bjornstad [185] rephrased the testing problem in the large-scale setting as a problem of prediction of latent class indicator variables. Through the data, parameters, and unobservables, they extended the likelihood approach to study the unobservable latent indicator. This method delivered oracle tests with an efficient extended LR test, and efficient FDR-control. They used hierarchical random-effect models to test the null hypothesis. Three examples were based in two-sample cases for the analysis of prostate cancer and leukaemia data.

(ii) Lv and Liu [200] contributed to MS in misspecified models under the Bayesian and the KLI principle. Via asymptotic expansions they obtained the GBIC and GBIC_p, then expanded in [94]. Consider a situation as for the MRIC approach in Section 3.2.2.3. Let n be the sample size, k be the cardinality of the assessed model, and consider the three estimates $\hat{\sigma}_h^{-2}$, $\hat{\mathbf{R}}^{-1}$, $\hat{\mathbf{C}}_{h,0}$, as defined in Eq.s (141), (143), and (144). Then approximate estimates of GBIC and GBIC_p are obtained as:

$$\text{GBIC} = \log \hat{\sigma}_h^2 + \frac{k \log n}{n} - \frac{\log \det(\hat{H}_h)}{n} \tag{206}$$

$$\text{GBIC}_p = \log \hat{\sigma}_h^2 + \frac{k \log n}{n} + \frac{\operatorname{tr}\left\{\hat{H}_h\right\}}{n} - \frac{\log \det(\hat{H}_h)}{n} \tag{207}$$

with $\hat{H}_h = \hat{\sigma}_h^{-2}\hat{\mathbf{R}}^{-1}\hat{\mathbf{C}}_{h,0}$, consistent estimator of $\sigma_h^{-2}\boldsymbol{R}^{-1}\boldsymbol{C}_{h,0}$, for generalized linear models. These two criteria offer advantages for both the correctly and misspecified case. In Hsu et al. [154, Section S5] is presented a comparison for possibly-misspecified time series models between these, including the AIC, BIC, Konishi and Kitagawa's GAIC [173], and the MRIC. The comparison was performed for linear, nonlinear, and high-dimensional models. It showed the superiority of the MRIC in different challenging settings.

(iii) Bogdan et al. [47] proposed the Sorted L-One Penalized Estimation (SLOPE), a method based on considering the problem as a convex program with computational complexity similar to that of procedures such as the lasso [298], dealing also with high-dimensional cases, in a way similar to Benjamini and Hockberg [31].

(iv) Candes et al. [64] showed feature selection in high-dimensional nonlinear models. Considering a general conditional model, for *i.i.d.* observations, acceptable in high dimensional applications in genetics, or client behavioural models, and assuming no knowledge about the conditional distribution. When the distribution of the covariates is known, they obtained powerful procedures to control the FDR in finite samples, by extending the results in Barber and Candès [27].

(v) Owrang and Jansson [218] assessed MS when the number of measurements is much smaller than dimension of the parameter space, and proposed the Extended Fisher Information Criterion (EFIC) for high-dimensional linear regression in the *i.i.d.* context. They also showed its consistency with probability one, and built a computationally affordable algorithm for its implementations. An additional feature is that it also determines implicitly the regularization parameter in the lasso estimator.

(vi) Section IV in Ding et al. [100] devoted to an overview of high-dimensional variable selection techniques.

(vii) Hsu et al. [153] addressed a serious lacuna in realistic applications: MS in the fixed-dimensionality setting (as n diverges but keeping the 'true' order k of the model fixed), with possibly misspecified time series models, for multi-step prediction, and the high-dimensional setting. They proposed the MRIC used after sequential procedure. In the first step, the Orthogonal-Greedy-Algorithm (OGA) of Ing and Lai [161] is employed, which is a stepwise regression method that performs variable selection sequentially for regression models where $k >> n$ through Residual Sum of Squares (RSS) minimization. In the second step, the

High-Dimensional Criterion (HDIC), which allows to trim eventual redundant indices corresponding to parameters that should be zero. They called this procedure "OGA+HDIC$_h$+Trim", and showed selection consistency in high-dimensional misspecified time series.

(viii) Xue and Hu [343] proposed an extension of the TIC to the online updating setting, with normal linear regressions models and the Cumulative Updated Estimating Equation (CUEE). They also showed that the BIC and their consistent RIC have more stable performance than the AIC and standard RIC for a fixed block size.

(ix) In the context of Causal Network Discovery, Runge et al. [252] considered the use of the AIC for hyperparameter choice in the condition selection stage. They also indicated that in the Momentary Conditional Independence test, CV, BIC, or AIC can be employed for selection of the regularization parameter.

(x) Demirkaya et al. [94] focussed on ultra-high dimensional MS, with model misspecification, where the dimesionality of the model can grow nonpolynomially with sample size n. Using generalized linear models, they followed Lv and Liu [200] and investigated the asymptotic expansion of the posterior model probability via QMLE [330].

(xi) Ying et al. [352] proposed an automated MS solution for anomaly detection in time series to ensure quality of online service. Their work proposed an automated selection mechanism for the choice of the best anomaly detection model and its hyper-parameters, showing that it can reduce the time-cost of improving unsatisfied detection.

(xii) Narisetty [212] overviewed Bayesian MS for high-dimensional data.

(xiii) Liu and Chen [193] formally introduced a threshold factor model where the dynamic of the time series is assumed to switch between regimes, depending on the value of the threshold variable. They proposed the estimation of the loading spaces and of the number of factor, via eigen-analysis of the cross moment matrices. They also developed an objective function to identify the threshold value, and showed that even in the case of over-estimation of the number of factors, the estimators kept consistency.

(xiv) Chiou et al. [78] assessed MS in the high-dimensional with heteroskedastic and serially correlated errors, also contemporarily. Via a two-part selection procedure, called Twohit, they proved its consistency in selection of regression and dispersion variables

for errors with short-memory, long-memory, or conditionally heteroskedastic component, and showed its finite sample performance.

(xv) Hastie et al. [144], extending Bertsimas et al. [36], studied the relative merit of $\mathcal{L}_0$, $\mathcal{L}_1$, and forward stepwise selection with different Signal to Noise Ratio (SNR). They found that:

(I) stepwise and best subset perform similarly;

(II) best subset often loses to the LASSO, except if there is high SNR;

(III) the relaxed LASSO performs as the best method in almost every scenario.

(xvi) Bertsimas et al. [37] provided a unified perspective for feature selection, focusing on five methods: the NP-hard cardinality- constrained formulation, its Boolean relaxation, $\mathcal{L}_1$-regularized estimators (lasso and elastic-net) and two non-convex penalties (smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP)). The comparison of the methods was in terms of accuracy and False Discovery Rate (FDR). They highlighted how most of the literature has focussed on the accuracy. They noticed the differences in terms of accuracy between convex and non-convex penalties. According to them, it mimics the distinction between robustness (i.e. good out-of-sample predictive performance even in noisy settings) and sparsity. Convex penalties delivers robust estimators, but non-convex regularization are theoretically more appealing given their less stringent assumptions. The best approaches are those combining convex with non-convex components.

(xvii) Lai and Yuan [181] proposed a review of stochastic approximation showing the evolution its introduction in 1951, and its connection with developments in time series and sequential analysis. In relation to our survey, variable selection with the Pure Greedy Algorithm, the OGA and similar works were discussed.

3.5 CONCLUSIVE REMARK

> *"The better approach will be to isolate for critical discussion the separate aspects entering into the final decision; what is contributed by the data? what are the assumed utilities and what is the basis for their calculation? Most importantly, have all the possible decisions been looked at? The contribution of statistical ideas to major decision making is more likely to be in the clarification of these separate elements than in the provision of a final optimum decision rule."*
> Cox and Hinkley, 1979 [86, p. 416]

The evolution of methods and techniques of innovative MS brought us to a current situation were there is increasing interest in: proposing novel solutions; combining alternative approaches; recovering classical solutions to be implemented in novel algorithms; or in refining current methods. See Gelman and Vehtari [122] for a path on the most important statistical ideas of past 50 years and their connections.

MS is still an open problem, with various potential approaches depending strongly on the objective of the analysis or prediction. We advice researchers and practitioners to keep in mind this vast multiplicity of approaches to select the most suitable for their studies.

4

A FIRST MULTIVARIATE EXTENSION OF THE MRIC: MULTIVARIATE RESPONSE AND SINGLE PREDICTOR

ABSTRACT

The Misspecification-Resistant Information Criterion (MRIC) proposed in [H.-L. Hsu, C.-K. Ing, H. Tong: *On model selection from a finite family of possibly misspecified time series models*. The Annals of Statistics. 47 (2), 1061–1087 (2019)] is a model selection criterion for univariate parametric time series that enjoys both the property of consistency and asymptotic efficiency. In this article we extend the MRIC to the case where the response is a multivariate time series and the predictor is univariate. The extension requires novel derivations based upon matrix theory. We obtain an asymptotic expression for the mean squared prediction error matrix, the vectorial MRIC and prove the consistency of its method-of-moments estimator. Moreover, we prove its asymptotic efficiency. Finally, we show with an example that, in presence of misspecification, the vectorial MRIC identifies the best predictive model whereas traditional information criteria like AIC or BIC fail to achieve the task.[1]

Keywords: multivariate time series, MSPE matrix, information criteria, vectorial MRIC, asymptotic efficiency, model selection.
2020 MSC: Primary 62H12, Secondary 62F12

4.1 INTRODUCTION

The appealing properties of the MRIC make it an ideal tool for omnibus time series model selection but, to date, only the univariate response case has been studied [153]. In this work we extend the MRIC to multivariate time series with a single regressor as to obtain the vectorial MRIC (hereafter VMRIC). As it will be clear, such an extension does not easily derive from the univariate case since it requires dealing with the dependence structure within the components of the vector of forecasting error and hence relies upon matrix theory. Such multivariate extension can be used in all those models where many time series depend upon a single regressor, like for instance, in econometrics, where many interest rates depend upon a single macroeconomic indicator,

1 This chapter is an updated version of Diaz Rubio, Giannerini and Goracci [98], and the results from Theorem 1 have been presented in Diaz Rubio, Giannerini and Goracci [97].

such as inflation. Other possible applications include dimension reduction and hedging, which is intimately connected to the problem of model selection [38].

The rest of the chapter is organized as follows: in Section 4.2 we introduce the notation and in Section 4.2.1 summarize the available results for the univariate case; in Section 4.3 we extend the MRIC approach to multivariate time series with a single regressor. In particular, in Section 4.3.1 we obtain the asymptotic decomposition of the Mean Squared Prediction Error (hereafter MSPE) matrix into two parts: the first one is linked to the goodness of fit of the model and the second one depends upon the prediction variance. In Section 4.3.2 we present the VMRIC and derive a consistent estimator for it, whereas in Section 4.3.3, we prove the asymptotic efficiency of the VMRIC. Section 4.4 presents an example to assess the effect of misspecification in the VMRIC framework. All the proofs are detailed in Section 4.5. Appendix B.1 contains an auxiliary technical lemma valid for the multivariate regressor setting.

4.2 NOTATION AND PRELIMINARIES

For each t, let $\{\mathbf{x}_t\}$ and $\{\mathbf{y}_t\}$, with $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,m})^\top$ and $\mathbf{y}_t = (y_{t,1}, \dots, y_{t,w})^\top$, be two weakly stationary stochastic processes defined over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. When $m = 1$ ($w = 1$, respectively) we write x_t (y_t). Given a vector $\mathbf{v}$ and a matrix $\mathbf{M}$, we use $\|\mathbf{v}\|$ and $\|\mathbf{M}\|$ to refer to the $\mathcal{L}_2$ vectorial norm and the matrix norm induced by the Euclidean norm, respectively. We write $o(1)$ ($o_p(1)$) to indicate a sequence that converges (in probability) to zero and $O(1)$ ($O_p(1)$) to indicate a sequence that is bounded (in probability). Moreover, let $\{c_n\}$ be a sequence of scalar random variables whereas $\{\mathbf{v}_n\}$ and $\{\mathbf{M}_n\}$ are sequences of random vectors and random matrices, respectively. We adopt the following notation: $\mathbf{v}_n = o_p(c_n)$ if $\|\mathbf{v}_n\| / c_n = o_p(1)$; $\mathbf{v}_n = O_p(c_n)$, if $\|\mathbf{v}_n\| / c_n = O_p(1)$, $\mathbf{M}_n = o_p(c_n)$ if $\|\mathbf{M}_n\| / c_n = o_p(1)$; $\mathbf{M}_n = O_p(c_n)$ if $\|\mathbf{M}_n\| / c_n = O_p(1)$. For further details on matrix algebra see [150, 216, 260], for multivariate time series see [198, 237, 307], and for asymptotic tools for vector and matrices, see [163].

Let $\{(\mathbf{x}_t, \mathbf{y}_t), t \in \{1, \dots, n\}\}$ be the observed sample, and divide the interval $\{1, 2, \dots, n\}$ into the *training set* $\{1, 2, \dots, N\}$ and the *test set* $\{N+1, \dots, N+h\}$, with h being the forecasting horizon. Note that $\mathbf{x}_t$ can contain both endogenous and exogenous variables, therefore, Model (208) encompasses many different models including, inter alia, VAR and VARX models. Without loss of generality assume $\mathrm{E}[\mathbf{x}_t] = \mathrm{E}[\mathbf{y}_t] = \mathbf{0}$. In order to forecast $\mathbf{y}_{n+h}$, $h \geq 1$, we adopt the following h-step ahead forecasting Model:

$$\mathbf{y}_{t+h} = \mathbf{B}_h \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(h)}, \tag{208}$$

where $\mathbf{B}_h$ is a $(w \times m)$ matrix parameters' matrix defined as the pseudo-true parameter for the possibly-misspecified model

$$\mathbf{B}_h = \underset{\mathbf{C}\in\mathbb{R}^{(w\times m)}}{\operatorname{argmin}} \; E\left[(\mathbf{y}_{t+h} - \mathbf{C}\mathbf{x}_t)\,(\mathbf{y}_{t+h} - \mathbf{C}\mathbf{x}_t)^\top\right], \tag{209}$$

and $\boldsymbol{\varepsilon}_t^{(h)}$ is the vector containing the w h-step ahead forecast errors; as before, if $w = 1$ we write $\varepsilon_t^{(h)}$.

Remark 3. *This general definition of the model includes:* (i) *the setting of multi-step forecasting with $h < 1$;* (ii) *exogenous and endogenous variables in $\mathbf{x}$; and* (iii) *cases where the prediction error vector $\boldsymbol{\varepsilon}_t^{(h)}$ can be serially correlated, but also correlated with $\mathbf{x}_s$, $s \neq t$. Moreover, the multivariate framework differs from [153] in different key aspects. For instance, (a) the components of the error vector can be cross-correlated, and (b) $\mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)}$ and $\mathbf{x}_k\boldsymbol{\varepsilon}_k^{(h)}$, for $t \neq k$, can also be both serially and cross correlated. Besides, as in the original MRIC, $\mathbf{x}$ may vary also with h, but for notational simplicity, it is avoided.*

Define

$$\hat{\mathbf{R}} = N^{-1}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{x}_t^\top \qquad \text{and} \qquad \mathbf{R} = \mathrm{E}[\mathbf{x}_1\mathbf{x}_1^\top]. \tag{210}$$

Then, the ordinary least squares estimator (hereafter OLS) of $\mathbf{B}_h$ results:

$$\hat{\mathbf{B}}_n(h) = \hat{\mathbf{R}}^{-1}\left(N^{-1}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{y}_{t+h}^\top\right). \tag{211}$$

When $m = 1$, $\mathbf{R}$ and $\mathbf{B}$ become R and β, respectively. The prediction of $\mathbf{y}_{n+h}$, $h \geq 1$, is given by

$$\hat{\mathbf{y}}_{n+h} = \hat{\mathbf{B}}_n(h)\mathbf{x}_n \tag{212}$$

and the corresponding Mean Squared Prediction Error matrix is

$$\mathbf{MSPE}_h = \mathrm{E}\left[(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})^\top\right]. \tag{213}$$

4.2.1 *The MRIC for parametric univariate time series models*

In [153], the authors focused on the case $w = 1$ and $m \geq 1$. Under appropriate conditions, they obtained the following asymptotic decomposition of MSPE:

$$\mathrm{MSPE}_h = \mathrm{E}\left[(y_{n+h} - \hat{y}_{n+h})^2\right] = \mathrm{MI}_h + n^{-1}(\mathrm{VI}_h + o(1)), \tag{214}$$

with

$$\mathrm{MI}_h = \mathrm{E}\left[\left(\varepsilon_n^{(h)}\right)^2\right],$$

$$\mathrm{VI}_h = \mathrm{tr}\left\{\mathbf{R}^{-1}\mathbf{C}_{h,0}\right\} + 2\sum_{s=1}^{h-1}\mathrm{tr}\left\{\mathbf{R}^{-1}\mathbf{C}_{h,s}\right\},$$

where $\mathbf{C}_{h,s} = \mathrm{E}\left[\mathbf{x}_1 \mathbf{x}_{1+s}^\top \varepsilon_1^{(h)} \varepsilon_{1+s}^{(h)}\right]$, $s \geq 0$, is the cross-covariance matrix between the regressors and the h-step ahead prediction error at lag s.

Remark 4. *The first part of Eq. (214) is the Misspecification Index (*MI*), linked to the goodness-of-fit of the model and coincides with the h-step ahead prediction error variance. The second component is the Variability Index (*VI*), which depends upon the variance of the h-step ahead predictor, $\hat{y}_{n+h} = \hat{\boldsymbol{\beta}}_n^\top(h)\mathbf{x}_n$, and is also linked to the bias of the estimator of $\boldsymbol{\beta}_h$.*

Based upon the above decomposition, the MRIC is defined as follows:

$$\mathrm{MRIC}_h = \hat{\mathrm{MI}}_h + \frac{\alpha_n}{n}\hat{\mathrm{VI}}_h, \tag{215}$$

with $\hat{\mathrm{MI}}_h$ and $\hat{\mathrm{VI}}_h$ being the estimators of MI_h and VI_h respectively, i.e.:

$$\hat{\mathrm{MI}}_h = N^{-1}\sum_{t=1}^{N}\left(\hat{\varepsilon}_t^{(h)}\right)^2, \quad \hat{\mathrm{VI}}_h = \mathrm{tr}\left\{\hat{R}^{-1}\hat{\mathbf{C}}_{h,0}\right\} + 2\sum_{s=1}^{h-1}\mathrm{tr}\left\{\hat{R}^{-1}\hat{\mathbf{C}}_{h,s}\right\},$$

where $\hat{\mathbf{C}}_{h,s} = (N-s)^{-1}\sum_{t=1}^{N-s}\mathbf{x}_t\mathbf{x}_{t+s}^\top\hat{\varepsilon}_t^{(h)}\hat{\varepsilon}_{t+s}^{(h)}$ and $\hat{\varepsilon}_t^{(h)} = y_{t+h} - \hat{\boldsymbol{\beta}}_n(h)\mathbf{x}_t$ is the estimated forecast error; α_n is a penalization term sequence such that, as n increases:

$$\frac{\alpha_n}{\sqrt{n}} \to +\infty \qquad \text{and} \qquad \frac{\alpha_n}{n} \to 0. \tag{216}$$

It is shown that $\hat{\mathrm{MI}}_h$ and $\hat{\mathrm{VI}}_h$ are consistent estimators of MI_h and VI_h, moreover the asymptotic efficiency of the MRIC is proved. By minimizing this criterion, the model which minimizes VI among those with minimum MI is selected. Among other features, the MRIC is particularly helpful in situations where competing models present the same goodness-of-fit and the same number of parameters.

Remark 5. *Hsu et al. [154, Section 6] indicated the required steps to determine α in the penalty weight of the type $\alpha_n = n^\alpha$, which satisfies Eq. (216). Let $\{S_t\}$, $1 \leq t \leq n$, be the possibly stationary series of interest; $\lfloor nd \rfloor$ with $d = 0.3$ be the latest sample's portion of $\{S_t\}$ retained for model evaluation; $\hat{S}_{t+h}$ be the predictor of S_{t+h} selected by a criterion and estimated via LS using observations up to time t; and the empirical MSPE, i.e. EMSPE:*

$$\mathrm{EMSPE} = \frac{1}{\lfloor nd \rfloor}\sum_{t=n-2\lfloor nd \rfloor - h + 1}^{n-\lfloor nd \rfloor - h}\left(S_{t+h} - \hat{S}_{t+h}\right)^2. \tag{217}$$

In a real data analysis, α is chosen as the minimizer of the in-sample empirical MSPE:

$$\frac{1}{\lfloor nd \rfloor}\sum_{t=n-2\lfloor nd \rfloor - h + 1}^{n-\lfloor nd \rfloor - h}\left(S_{t+h} - \hat{S}_{t+h}^{(\alpha)}\right)^2, \tag{218}$$

over $\alpha \in \{0.1, \ldots, 0.8\}$, where $\hat{S}_{t+h}^{(\alpha)}$ is $\hat{S}_{t+h}$ with order selected by MRIC setting the penalty of $\alpha_n = n^{\alpha}$. Hyndman and Koehler [159] surveyed measures of forecast accuracy including scale-dependent measures, measures based on percentage errors, on relative errors, or relative measures. They proposed the Mean Absolute Scaled Error, useful also for multi-step forecast in univariate time series and displaying positive features for the comparison of different methods. The choice of α could benefit from these considerations.

Remark 6. *The type of penalty considered in [153] is similar to that used in [277, p. 230] for the correctly specified case.*

4.3 A MULTIVARIATE EXTENSION OF THE MRIC FRAMEWORK

In this section we extend the MRIC approach to the case where the response is a multivariate time series ($w \geq 2$) and the predictor is univariate ($m = 1$), for a generic h-step ahead forecast. Hence, Model (208) reduces to $\mathbf{y}_{t+h} = \boldsymbol{\beta}_h x_t + \boldsymbol{\varepsilon}_t^{(h)}$, namely:

$$\begin{cases} y_{t+h,1} = \beta_{h,1} x_t + \varepsilon_{t,1}^{(h)} \\ y_{t+h,2} = \beta_{h,2} x_t + \varepsilon_{t,2}^{(h)} \\ \vdots \\ y_{t+h,w} = \beta_{h,w} x_t + \varepsilon_{t,w}^{(h)}. \end{cases} \tag{219}$$

4.3.1 *Asymptotic decomposition of the MSPE matrix*

We extend the asymptotic representation of the MSPE_h defined in (214) which is the key step to derive the VMRIC in this multivariate framework. We rely upon the following assumptions, which are the natural multivariate extensions of those in [153].

Assumptions 9.

$(C1)$ $\exists\, q_1 > 5, 0 < K_1 < \infty$: *for any* $1 \leq n_1 < n_2 \leq n$,

$$\mathrm{E}\left[\left|(n_2 - n_1 + 1)^{-1/2} \sum_{t=n_1}^{n_2} x_t^2 - \mathrm{E}\left[x_t^2\right]\right|^{q_1}\right] \leq K_1.$$

$(C2)$ 1. $\mathbf{C}_{h,s} = \mathrm{E}\left[\boldsymbol{\varepsilon}_t^{(h)} x_t \left(\boldsymbol{\varepsilon}_{t+s}^{(h)} x_{t+s}\right)^\top\right] \perp t,$

2. $\mathrm{E}\left[x_1 x_n \varepsilon_{1,i}^{(h)} \varepsilon_{n,j}^{(h)}\right] = o(n^{-1})\ \forall\, i, j \in \{1, \ldots, w\}.$

$(C3)$ 1. $\sup_{-\infty<t<\infty} \mathrm{E}\left[|x_t|^{10}\right] < \infty,$

2. $\sup_{-\infty<t<\infty} \mathrm{E}\left[\left\|\boldsymbol{\varepsilon}_t^{(h)}\right\|^6\right] < \infty.$

$(C4)$ $\exists\, 0 < K_2 < \infty$: *for* $1 \leq n_1 < n_2 \leq n$,

$$\mathrm{E}\left[\left\|(n_2 - n_1 + 1)^{-\frac{1}{2}} \sum_{t=n_1}^{n_2} \boldsymbol{\varepsilon}_t^{(h)} x_t\right\|^5\right] < K_2.$$

$(C5)$ *For any* $q > 0$, $\mathrm{E}\left[\left|\hat{R}^{-1}\right|^q\right] = O(1).$

$(C6)$ $\exists\, \mathcal{F}_t \subseteq \mathcal{F}, \mathcal{F}_t$ *an increasing sequence of* σ*-fields such that:*

1. x_t *is* $\mathcal{F}_t$*-measurable,*

2. $\sup_{-\infty<t<\infty} \mathrm{E}\left[\left|\mathrm{E}\left[x_t^2 \mid \mathcal{F}_{t-k}\right] - R\right|^3\right] = o(1),$ *as* $k \to \infty,$

3. $\sup_{-\infty<t<\infty} \mathrm{E}\left[\left\|\mathrm{E}\left[\boldsymbol{\varepsilon}_t^{(h)} x_t \mid \mathcal{F}_{t-k}\right]\right\|^3\right] = o(1),$ *as* $k \to \infty.$

Theorem 1. *Under the regularity conditions (C1) – (C6), the asymptotic expression of the* $\boldsymbol{MSPE}_h$ *defined in (213) results*

$$N\left\{\mathrm{E}\left[\left(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h}\right)\left(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h}\right)^\top - \mathrm{E}\left[\boldsymbol{\varepsilon}_n^{(h)} \boldsymbol{\varepsilon}_n^{(h)\top}\right]\right]\right\} \tag{220}$$
$$= R^{-1}\,\mathrm{E}\left[\left(\boldsymbol{\varepsilon}_1^{(h)} x_1\right)\left(\boldsymbol{\varepsilon}_1^{(h)} x_1\right)^\top\right]$$
$$+ R^{-1}\,\mathrm{E}\left[\sum_{s=1}^{h-1}\left\{\left(\boldsymbol{\varepsilon}_1^{(h)} x_1\right)\left(\boldsymbol{\varepsilon}_{s+1}^{(h)} x_{s+1}\right)^\top + \left(\boldsymbol{\varepsilon}_{s+1}^{(h)} x_{s+1}\right)\left(\boldsymbol{\varepsilon}_1^{(h)} x_1\right)^\top\right\}\right]$$
$$+ o(1).$$

Remark 7. *For interpretations and details on Assumptions 9 and on the following Assumptions 10, please refer to Remark 14, Remark 16, and Section 5.4 in Chapter 5.*

4.3.2 *VMRIC and its consistent estimation*

In this section we introduce the VMRIC. Let $\{\alpha_n\}$ be the penalization term sequence defined as in Eq. (216).

$$\mathrm{VMRIC}_h = \|\mathbf{MI}_h\| + \left\|\frac{\alpha_n}{n}\mathbf{VI}_h\right\| \tag{221}$$

where

$$\begin{aligned}
\mathbf{MI}_h &= \mathrm{E}\left[\left(\boldsymbol{\varepsilon}_t^{(h)}\boldsymbol{\varepsilon}_t^{(h)\top}\right)\right], \\
\mathbf{VI}_h &= R^{-1}\left(\mathbf{C}_{h,0} + \sum_{s=1}^{h-1}\left(\mathbf{C}_{h,s} + \mathbf{C}_{h,s}^{\top}\right)\right), \\
\mathbf{C}_{h,s} &= \mathrm{E}\left[\left(x_t\boldsymbol{\varepsilon}_t^{(h)}\right)\left(x_t\boldsymbol{\varepsilon}_t^{(h)}\right)^{\top}\right].
\end{aligned}$$

The VMRIC can be estimated via the method-of-moments as to obtain:

$$\widehat{\mathrm{VMRIC}}_h \equiv \left\|\hat{\mathbf{MI}}_h\right\| + \left\|\frac{\alpha_n}{n}\hat{\mathbf{VI}}_h\right\|, \tag{222}$$

where

$$\begin{aligned}
\hat{\mathbf{MI}}_h &= N^{-1}\sum_{t=1}^{N}\left(\hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t^{\top}\right), \\
\hat{\mathbf{VI}}_h &= \hat{R}^{-1}\left[\hat{\mathbf{C}}_{h,0} + \sum_{s=1}^{h-1}\left(\hat{\mathbf{C}}_{h,s} + \hat{\mathbf{C}}_{h,s}^{\top}\right)\right],
\end{aligned}$$

and $\hat{\mathbf{C}}_{h,s} = (N-s)^{-1}\sum_{t=1}^{N-s} x_t x_{t+s}\hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_{t+s}^{\top}$, with $\hat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_{t+h} - \hat{\boldsymbol{\beta}}_n(h)x_t$ the estimated forecast error vector.
In Theorem 2 we prove that $\hat{\mathbf{MI}}_h$ and $\hat{\mathbf{VI}}_h$ are consistent estimators of $\mathbf{MI}_h$ and $\mathbf{VI}_h$, respectively. Theorem 2 relies upon the following assumptions, that are less restrictive with respect to (C1) – (C6). For further discussions on the assumptions see [153, Remark 1–3, p. 1073].

Assumptions 10. *For each* $0 \leq s \leq h-1$, *we assume the following:*

$$\begin{aligned}
&(A1) \quad n^{-1}\sum_{t=1}^{n}\left(\boldsymbol{\varepsilon}_t^{(h)}\boldsymbol{\varepsilon}_t^{(h)\top}\right) = \mathrm{E}\left[\boldsymbol{\varepsilon}_1^{(h)}\boldsymbol{\varepsilon}_1^{(h)\top}\right] + O_p\left(n^{-1/2}\right) \\
&(A2) \quad n^{-1}\sum_{t=1}^{n}\left(x_t\boldsymbol{\varepsilon}_t^{(h)}\right)\left(x_{t+s}\boldsymbol{\varepsilon}_{t+s}^{(h)}\right)^{\top} = \boldsymbol{C}_{h,s} + o_p(1), \\
&(A3) \quad n^{-1/2}\sum_{t=1}^{n} x_t\boldsymbol{\varepsilon}_t^{(h)} = O_p(1). \\
&(A4) \quad n^{-1}\sum_{t=1}^{n} x_t^2 = R + o_p(1), \\
&(A5) \quad \sup_{-\infty<t<\infty}\mathrm{E}\left[\left\|\boldsymbol{\varepsilon}_t^{(h)}\right\|^4\right] + \sup_{-\infty<t<\infty}\mathrm{E}\left[\|x_t\|^4\right] < \infty.
\end{aligned}$$

Theorem 2. *If Assumptions (A1) – (A5) hold, then for the case* $w \geq 2$, *and* $m = 1$ *we obtain:*

$$\begin{aligned}
\hat{\boldsymbol{MI}}_h &= \boldsymbol{MI}_h + O_p(n^{-1/2}), \\
\hat{\boldsymbol{VI}}_h &= \boldsymbol{VI}_h + o_p(1).
\end{aligned}$$

4.3.3 *Asymptotic efficiency*

In this section we prove the asymptotic efficiency of the VMRIC in the fixed dimensionality framework. To this end, let $\mathcal{M}$ be the set of K candidate models; each model is indicated either by ℓ or κ, $1 \leq \ell, \kappa \leq K$. Define the subsets M_1 and M_2 as follows:

$$M_1 = \left\{ \kappa : 1 \leq \kappa \leq K, \|\mathbf{MI}_h(\kappa)\| = \min_{1 \leq \ell \leq K} \|\mathbf{MI}_h(\ell)\| \right\}, \tag{223}$$

$$M_2 = \left\{ \kappa : \kappa \in M_1, \|\mathbf{VI}_h(\kappa)\| = \min_{\ell \in M_1} \|\mathbf{VI}_h(\ell)\| \right\}. \tag{224}$$

In short, for a given forecast horizon h, M_1 contains the models with the minimum $\mathbf{MI}_h$ whereas in M_2 we are minimizing $\mathbf{VI}_h$ among the candidates models in M_1. The definition of efficiency used in our framework is the same as that of [153]:

Definition 70. *Given a sample of size n, a model selection criterion is said to be asymptotically efficient if it selects the model $\hat{\ell}_h$ such that*

$$\lim_{n \to \infty} \Pr\left(\hat{\ell}_h \in M_2\right) = 1.$$

Remark 8. *Alternative definitions of asymptotic efficiency for model selection are available. For instance, in the framework of linear stationary processes, [272] defines the Mean Efficiency when a criterion attains asymptotically a lower bound for the sum of squared prediction errors. Also, the notion of Approximate Efficiency is given in [276]. In [190], a criterion that depends upon the ratio between loss functions is introduced. This latter definition is similar to the Loss Efficiency proposed in [267]. See Section 2.5.2 for a detailed discussion.*

The VMRIC selects the model with the smallest variability index among those that achieve the best goodness of fit. Hence, the selected model $\hat{\ell}_h$ is such that:

$$\text{VMRIC}_h\left(\hat{\ell}_h\right) \equiv \min_{1 \leq \ell \leq K} \left\|\hat{\mathbf{MI}}_h(\ell)\right\| + \min_{\ell \in M_1} \left\|\frac{\alpha_n}{n} \hat{\mathbf{VI}}_h(\ell)\right\|. \tag{225}$$

In the next Theorem we show that the VMRIC is an asymptotic efficient model selection criterion in the sense of Definition 70.

Theorem 3. *Assume that for each $1 \leq \ell \leq K$, $0 \leq s \leq h-1$, Theorem 2 holds and let $\hat{\ell}_h$ be the model selected by the* VMRIC. *Then we have that:*

$$\lim_{n \to \infty} \Pr\left(\hat{\ell}_h \in M_2\right) = 1,$$

namely, the VMRIC *is asymptotically efficient in the sense of Definition 70.*

4.4 EXAMPLE: A MISSPECIFIED BIVARIATE AR(2) MODELS

The aim of this section is twofold. First, we assess the goodness of the theoretical derivations and the finite sample behaviour of the Method-of-Moments Estimator (MoME) for the VMRIC. Second, we show that in presence of misspecification the VMRIC leads to selecting the best predictive model (i.e. is asymptotically efficient) whereas both the AIC and the BIC fail to do so. In order to achieve the goals we consider a bivariate AR(2) DGP and use two misspecified predictive models for it: in Model 1 there is one omitted lagged predictor, whereas Model 2 uses only one non-informative predictor. We derive theoretically the Mean Square Prediction Error matrix and the VMRIC for both models and these show that Model 1 is a better predictive model over Model 2. Based on this, we assess the ability of the VMRIC, and of the multivariate versions of the AIC and BIC to select the best model (Model 1) in finite samples and for different parameterizations.

We start by providing the definition of misspecification as in [153, p. 1084]. Consider an increasing sequence of σ-fields, $\{\mathcal{G}_t\}$ such that $\sigma(\mathbf{x}_s, s \leq t) \subseteq \mathcal{G}_t \subseteq \mathcal{F}$, where $\{\mathbf{x}_t\}$ is an m-dimensional weakly stationary process defined over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 71. *The h-step ahead forecasting model:*

$$\mathbf{y}_{t+h} = \boldsymbol{\beta}_h^\top \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(h)}, \tag{226}$$

is correctly specified with respect to an increasing sequence of σ-fields, $\{\mathcal{G}_t\}$ if

$$E[\mathbf{y}_{t+h} \mid \mathcal{G}_t] = \boldsymbol{\beta}_h^\top \mathbf{x}_t \quad a.s., \ \forall\ -\infty < t < \infty. \tag{227}$$

Otherwise, it is misspecified.

Remark 9. *For correctly specified models we have $E\left[\boldsymbol{\varepsilon}_t^{(h)} x_{t-j}\right] = \mathbf{0}$, $j \geq 0$, i.e. both simultaneous and lagged correlations are null vectors. A possible consequence of misspecification is that it may occur that $E\left[\boldsymbol{\varepsilon}_t^{(h)} x_s\right] \neq \mathbf{0}$, for $s \neq t$, while still have $E\left[\boldsymbol{\varepsilon}_t^{(h)} x_t\right] = \mathbf{0}$, i.e. to have null simultaneous correlation and non-null cross- serial- correlation between the forecasting error vector and the regressor, e.g. Remark 12. Null simultaneous correlation depends on the definition of the pseudo-true parameter $\boldsymbol{\beta}_h$.*

Remark 10. *Hansen [139, p. 268] details the consequences for a constrained estimator $\widetilde{\boldsymbol{\beta}}$ when the constraint is incorrect, indicating that it generalizes the analysis of 'omitted variable bias'. Estimators' bias is unavoidable, but the distributions centred at the pseudo-true projections (instead of being at the true parameter $\boldsymbol{\beta}$) maintain the conventional covariance estimator. An alternative approach delivering the same explanation is that of local misspecification, i.e. when the misspecification decreases with sample size n, convenient for other purposes.*

Remark 11. *In the univariate case of [153, p. 1067], references to 'single-step' model selection procedures robust to model misspecification are indicated, some of which partially reviewed in Chapter 2 and Chapter 3: TIC [292], GIC [173], GBIC, and GBIC$_p$ [200]. Hsu et al. note that it is difficult to justify their asymptotic efficiency in the Fixed Dimensionality (FD) setting, as it is shown in [154, S5]. The MRIC instead achieves this property without the help of further criteria. In particular, the TIC, from the* i.i.d. *setting, has a similar term to VI_1 for $h = 1$ in the univariate case of [153], i.e. the variability index for one-step ahead forecast. It was obtained as a bias correction as shown in Section 2.7.1. The focus is on independent observations, excluding the case of time series. The APE as shown in [328] considers this case, but focusses on the one-step prediction case, not applicable to the MSPE or multi-step prediction. A comparison against the TIC remains an interesting empirical task that will be included in successive works.*

Consider the following DGP :

$$\mathbf{y}_{t+1} = \boldsymbol{a} w_t + \boldsymbol{\varepsilon}_{t+1}, \tag{228}$$

where $\boldsymbol{a} \neq \mathbf{0}$, $\{\boldsymbol{\varepsilon}_t\}$ is a sequence of independent and identically distributed (hereafter i.i.d.) bivariate random vectors with $E[\boldsymbol{\varepsilon}_1] = \mathbf{0}$, $E\left[\boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_1^\top\right] > \mathbf{0}$ and w_t is the following scalar AR(2) process:

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \delta_t, \tag{229}$$

where $\phi_1 \phi_2 \neq 0$, $\{\delta_t\}$ a sequence of i.i.d. random variables independent of $\{\boldsymbol{\varepsilon}_t\}$ such that

$$E[\delta_1] = 0 \quad \text{and} \quad E\left[\delta_1^2\right] = 1 - \phi_2^2 - \left\{\phi_1^2 \frac{1+\phi_2}{1-\phi_2}\right\}.$$

Hence, we obtain $E[w_t^2] \equiv \gamma_w(0) = 1$, where $\gamma_w(j) = E[w_t w_{t+j}]$ is the j-th lag autocovariance of w.

We consider the correctly specified 2-step ahead forecasting model:

$$\begin{aligned} &\mathbf{y}_{t+2} = \boldsymbol{a} w_{t+1} + \boldsymbol{\varepsilon}_{t+2}, \text{ which leads to} \\ &\mathbf{y}_{t+2} = \boldsymbol{a}\phi_1 w_t + \boldsymbol{a}\phi_2 w_{t-1} + \boldsymbol{\varepsilon}_t^{*(2)}, \end{aligned} \tag{230}$$

where $\boldsymbol{\varepsilon}_t^{*(2)} = \boldsymbol{\varepsilon}_{t+2} + \boldsymbol{a}\delta_{t+1}$. It can be easily proved that $E\left[\boldsymbol{\varepsilon}_t^{*(2)} w_{t-j}\right] = \mathbf{0}$ for $j \geq 0$.

Now, consider the following *misspecified* model, Model 1:

$$\mathbf{y}_{t+2} = \boldsymbol{\beta} w_t + \boldsymbol{\varepsilon}_t^{(2)}, \quad \text{with} \quad \boldsymbol{\beta} = \frac{E[\mathbf{y}_{t+2} w_t]}{V[w_T]} = \boldsymbol{a}\left(\phi_1 + \frac{\phi_1\phi_2}{1-\phi_2}\right).$$

The forecasting error results:

$$\boldsymbol{\varepsilon}_t^{(2)} = \boldsymbol{\varepsilon}_t^{*(2)} - \boldsymbol{a}\phi_2\left[\frac{\phi_1}{1-\phi_2} w_t - w_{t-1}\right]. \tag{231}$$

Remark 12. *We show that in our case, in presence of misspecification we have* $E\left[\boldsymbol{\varepsilon}_t^{(2)} w_t\right] = \mathbf{0}$*, whereas* $E\left[\boldsymbol{\varepsilon}_t^{(2)}\, w_{t-j}\right] \neq \mathbf{0}$ *for* $j \neq 0$*:*

$$\begin{aligned} E\left[\boldsymbol{\varepsilon}_t^{(2)} w_{t-j}\right] &= -\boldsymbol{a}\frac{\phi_2}{1-\phi_2}\left\{\phi_1 E\left[w_t w_{t-j}\right] - (1-\phi_2)\, E\left[w_{t-1} w_{t-j}\right]\right\} \\ &= -\boldsymbol{a}\frac{\phi_2}{1-\phi_2}\left\{\gamma_w(j+1) - \gamma_w(j-1)\right\}, \end{aligned} \tag{232}$$

which is zero if $j = 0$*, otherwise this is generally not the case.*

We compute the theoretical value of the VMRIC by using Eq. (221). After some routine algebra, we get:

$$\mathbf{MI} = E\left[\boldsymbol{\varepsilon}_n^{(2)} \boldsymbol{\varepsilon}_n^{(2)\top}\right] = \boldsymbol{\sigma}_\varepsilon^2 + \boldsymbol{a}\boldsymbol{a}^\top\left[\sigma_\delta^2 + \phi_2^2\left(1 - \gamma_w^2(1)\right)\right], \tag{233}$$

which highlights how the variance-covariance matrix of the 2-step ahead forecast vector is equal to the DGP's variance-covariance plus a bias term that depends upon the misspecification considered.

Now we focus on the variability index **VI**. We get

$$\begin{aligned} \mathbf{C}_{2,0} = \boldsymbol{\sigma}_\varepsilon^2 + \boldsymbol{a}\boldsymbol{a}^\top\Big\{\sigma_\delta^2 + \phi_2^2\Big(&\gamma_w(1)^2 E\left[w_t^4\right] \\ &- 2\gamma_w(1) E\left[w_t^3 w_{t-1}\right] + E\left[w_t^2 w_{t-1}^2\right]\Big)\Big\} \end{aligned} \tag{234}$$

and

$$\begin{aligned} \mathbf{C}_{2,1} = \boldsymbol{a}\boldsymbol{a}^\top \gamma_w(1)\Big(&b_1 E\left[w_{t-1}^3 w_{t-2}\right] + b_2 E\left[w_{t-1} w_{t-2}^3\right] \\ &+ b_3 E\left[w_{t-1}^2 w_{t-2}^2\right]\Big), \end{aligned} \tag{235}$$

where

$$\begin{aligned} b_1 &= 2\phi_1\phi_2\gamma_w(1) - \phi_2, \qquad b_2 = -\phi_2^2, \\ b_3 &= \phi_2\left(\phi_2\gamma_w(1) - 2\phi_1 + \gamma_w(1)^{-1}\right). \end{aligned}$$

Following Eq. (221), the results from Eq. (233), (234), and (235), deliver the VMRIC for this case.

Now we consider a second misspecified model, Model 2:

$$\mathbf{y}_{t+2} = \boldsymbol{\rho} z_t + \boldsymbol{\eta}_t^{(2)}, \tag{236}$$

where z_t is a weakly stationary linear AR(1) process independent of w_t:

$$z_t = \psi_1 z_{t-1} + \upsilon_t \tag{237}$$

with $\psi_1 \in (-1, 1)$, and $\{\upsilon_t\}$ is a sequence of i.i.d. random variables independent of both the error terms $\{\delta_t\}$ and $\{\boldsymbol{\varepsilon}_t\}$ such that $E[\upsilon_t] = 0$ and $E[\upsilon_t^2] = 1 - \psi_1^2$, delivering $E[z_t] = 0$ and $E[z_t^2] = 1$. Thus, z_t is uncorrelated with both w_t and $\mathbf{y}_t$, therefore $\boldsymbol{\rho} = \mathbf{0}$. The forecasting error

Table 1: Parameters' combinations for the DGP of Eq. (228), (229), and (237).

Case	ϕ_1	ϕ_2	a_1	a_2	ψ_1
1	0.4	-0.75	1.50	-2.00	0.80
2	-0.4	-0.45	-0.75	1.25	-0.65
3	0.3	-0.80	1.00	0.50	-0.75

in this case results $\boldsymbol{\eta}_t^{(2)} = \boldsymbol{a} w_{t+1} + \boldsymbol{\varepsilon}_{t+2}$. Following similar arguments as above we obtain **MI** and **VI** for Model 2:

$$\mathbf{MI} = \boldsymbol{\sigma}_{\varepsilon}^2 + \boldsymbol{a}\boldsymbol{a}^{\top} \tag{238}$$

$$\mathbf{VI} = \boldsymbol{\sigma}_{\varepsilon}^2 + \boldsymbol{a}\boldsymbol{a}^{\top}(1 + 2\psi_1 \gamma_w(1)) \tag{239}$$

As mentioned above, Model 1 is misspecified since it omits the lagged predictor w_{t-1}, while Model 2 only includes the non-informative predictor z_t.

4.4.1 *Large and finite sample performance*

First, we compare the above theoretical derivations with their sample counterpart. We consider three different parameterizations, presented in Table 1. Also, $\alpha_n = n^{\alpha}$ with $\alpha = 0.85$. Note that, in order for Eq. (216) to hold, α must range in $(0.5, 1)$. Further experiments showed that results are fairly robust if reasonable values of α are selected. For an empirical method to determine it, see [154, Section 5]. We take the following variance/covariance matrix for the innovations:

$$E[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^{\top}] = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

We compute both the VMRIC for Model 1 and Model 2, and estimate the VMR̂IC and VMR̂IC on a large sample of $n = 10^6$ observations. The results are shown in Table 2 for the two models, where the theoretical VMRIC (rows 1 and 3) is compared with the estimated one (rows 2 and 4). The results seem to confirm the consistency of the estimator shown in Eq. (222). Clearly, the VMRIC of Model 1 is consistently smaller than that of Model 2 and indicates its superior predictive capability. The finite sample behaviour of the method of moments estimator of the VMRIC can be further appreciated in Table 3 where we show their bias and Mean Squared Error (MSE), computed as follows:

$$\text{Bias} = \left\| E\left[\left(\hat{\mathbf{MI}} - \mathbf{MI} \right) + \frac{n^{\alpha}}{n} \left(\hat{\mathbf{VI}} - \mathbf{VI} \right) \right] \right\|, \tag{240}$$

$$\text{MSE} = \left\| E\left[\left\{ \left(\hat{\mathbf{MI}} - \mathbf{MI} \right) + \frac{n^{\alpha}}{n} \left(\hat{\mathbf{VI}} - \mathbf{VI} \right) \right\}^2 \right] \right\|. \tag{241}$$

Table 2: Theoretical and estimated VMRIC of Models 1 and 2, for the three parametrizations of Table 1, computed on a data set of $n = 10^6$ observations.

	Model 1		Model 2	
Case	VMRIC	$\widehat{\text{VMRIC}}$	VMRIC	$\widehat{\text{VMRIC}}$
1	6.671	6.636	7.914	7.902
2	2.777	2.768	3.164	3.168
3	2.801	2.784	2.994	2.993

Table 3: Bias and Mean-Squared Error (MSE) for the (method of moments) estimator of the VMRIC for the three parametrizations, $\alpha = 0.85$ and different sample size n. The results are based upon 1000 Monte Carlo replications.

	Case 1		Case 2		Case 3	
n	Bias	MSE	Bias	MSE	Bias	MSE
100	0.227	1.137	0.063	0.306	0.030	0.182
250	0.117	0.455	0.022	0.107	0.032	0.076
500	0.061	0.225	0.015	0.048	0.004	0.032
1000	0.019	0.109	0.010	0.023	0.002	0.015
2500	0.008	0.044	0.001	0.009	0.001	0.006
5000	0.009	0.023	0.001	0.004	0.003	0.003
10000	0.001	0.012	0.003	0.002	0.001	0.002
15000	0.004	0.008	0.001	0.001	0.002	0.001
30000	0.002	0.004	0.001	0.001	0.001	0.001

The results are based upon 1000 Monte Carlo replications and seem to indicate a rate of convergence of the order of n^{-1}.

In Table 4, we show the percentages of correct model selection by the VMRIC, compared with the multivariate version of the AIC and BIC for the three parameterizations of Table 1. For a sample size of $n = 100$, both the AIC and BIC select the best predictive model in about 50% of the cases and relying upon them is tantamount to tossing a fair coin. In such a case, the VMRIC selects the correct model in about 80% of the cases and reaches 100% for $n = 1000$. On the contrary, for Case 3, both the AIC and BIC cannot go above 64% for a sample size as large as $n = 10000$ observations and this is a general indication of their lack of asymptotic efficiency. At the end of this chapter, Figure 1 presents the box-plots and empirical distribution of the method of moments estimators, while Tables 5, 6, and 7 present results for additional parametrizations.

Table 4: Percentages of correctly selected models by the three information criteria for the three parametrizations and varying sample size n.

	Case 1			Case 2			Case 3		
n	VMRIC	AIC	BIC	VMRIC	AIC	BIC	VMRIC	AIC	BIC
100	85.9	52.5	52.5	84.6	56.2	56.2	72.1	49.0	49.0
1000	99.9	65.6	65.6	99.9	73.7	73.7	97.0	56.8	56.8
10000	100	88.0	88.0	100	97.8	97.8	100	63.8	63.8

4.5 PROOFS

In this section we detail the proofs of the three theorems. Hereafter all the derivations hold for any fixed $h \geq 1$; for the sake of presentation we write $\boldsymbol{\varepsilon}_t$ instead of $\boldsymbol{\varepsilon}_t^{(h)}$. Remember that $\{l_n\}$ indicates an increasing sequence of positive integers such that:

$$l_n \to \infty, \qquad \frac{l_n}{\sqrt{n}} = o(1) \tag{242}$$

and define $a = n - l_n - h$ and $b = n - l_n - h + 1$.

4.5.1 *Proof of Theorem 1*

The proof of Theorem 1 relies upon four propositions.

Proposition 3. *Under assumptions of Theorem 1, it holds that:*

$$N(\mathrm{I}) = (\mathrm{III}) + o(1), \tag{243}$$

where

$$\begin{aligned}(\mathrm{I}) &= -\mathrm{E}\left[x_n \hat{R}^{-1}\left(\hat{\boldsymbol{\Sigma}}\boldsymbol{\varepsilon}_n^\top + \boldsymbol{\varepsilon}_n \hat{\boldsymbol{\Sigma}}^\top\right)\right], \\ (\mathrm{III}) &= -\mathrm{E}\left[x_n R^{-1}\left(\hat{\boldsymbol{\Sigma}}_A \boldsymbol{\varepsilon}_n^\top + \boldsymbol{\varepsilon}_n \hat{\boldsymbol{\Sigma}}_A^\top\right)\right],\end{aligned}$$

with $\hat{\boldsymbol{\Sigma}} = \left(N^{-1}\sum_{t=1}^N x_t \boldsymbol{\varepsilon}_t\right)$ *and* $\hat{\boldsymbol{\Sigma}}_A = \sum_{t=1}^N \boldsymbol{\varepsilon}_t x_t$.

Proof. Let $\boldsymbol{A}_1 = \sum_{t=1}^N (\boldsymbol{\varepsilon}_t x_t)\, \boldsymbol{\varepsilon}_n^\top$ and note that

$$\|(\mathrm{I}) - (\mathrm{III})\| = \left\|\mathrm{E}\left[x_n\left(\hat{R}^{-1} - R^{-1}\right)\left(\boldsymbol{A}_1 + \boldsymbol{A}_1^\top\right)\right]\right\|. \tag{244}$$

By using standard properties of the norm, (243) follows upon proving that

$$\left\|\mathrm{E}\left[x_n\left(\hat{R}^{-1} - R^{-1}\right)\boldsymbol{A}_1^\top\right]\right\| = o(1). \tag{245}$$

Let

$$\tilde{R} = (n - l_n)^{-1} \sum_{t=1}^{n-l_n} x_t^2. \tag{246}$$

By adding and subtracting $\varepsilon_n x_n \left(\tilde{R}^{-1} \left[\sum_{t=1}^{N} (\varepsilon_t x_t) \right]^\top \right)$, we have

$$\begin{aligned} &\mathrm{E}\left[x_n \left(\hat{R}^{-1} - R^{-1} \right) \boldsymbol{A}_1^\top \right] = \qquad (247) \\ &\mathrm{E}\left[\varepsilon_n x_n \left(\hat{R}^{-1} - \tilde{R}^{-1} \right) \sum_{t=1}^{N} \varepsilon_t^\top x_t \right] \\ &+ \mathrm{E}\left[\varepsilon_n x_n \left(\tilde{R}^{-1} - R^{-1} \right) \sum_{t=1}^{N} \varepsilon_t^\top x_t \right] \end{aligned}$$

which is equal to

$$\mathrm{E}\left[\varepsilon_n x_n \left(\hat{R}^{-1} - \tilde{R}^{-1} \right) \left(\sum_{t=1}^{N} \varepsilon_t x_t \right)^\top \right] \qquad (248)$$

$$+ \mathrm{E}\left[\varepsilon_n x_n \left(\tilde{R}^{-1} - R^{-1} \right) \left(\sum_{t=b}^{N} \varepsilon_t x_t \right)^\top \right] \qquad (249)$$

$$+ \mathrm{E}\left[\varepsilon_n x_n \left(\tilde{R}^{-1} - R^{-1} \right) \left(\sum_{t=1}^{a} \varepsilon_t x_t \right)^\top \right]. \qquad (250)$$

We show below that the norms of (248), (249) and (250) are asymptotically negligible. Focus on the first one: by combining conditions (C3), (C4), Lemma 1, and Hölder's inequality, it follows that $\|(248)\|$ is bounded by

$$\begin{aligned} &\mathrm{E}\left[\left\| \varepsilon_n x_n \left(\hat{R}^{-1} - \tilde{R}^{-1} \right) \left(\sum_{t=1}^{N} \varepsilon_t x_t \right)^\top \right\| \right] \leq \mathrm{E}\left[\|\varepsilon_n\|^6 \right]^{\frac{1}{6}} \mathrm{E}\left[|x_n|^6 \right]^{\frac{1}{6}} \\ &\times \mathrm{E}\left[\left| \hat{R}^{-1} - \tilde{R}^{-1} \right|^3 \right]^{\frac{1}{3}} \mathrm{E}\left[\left\| N^{\frac{1}{2}} N^{-\frac{1}{2}} \sum_{t=1}^{N} \varepsilon_t x_t \right\|^3 \right]^{\frac{1}{3}} = O\left(\frac{l_n}{n^{1/2}} \right), \end{aligned}$$

which converges to zero due to the definition of l_n in (242). Similarly, we have that $\|(249)\|$ is bounded by

$$\begin{aligned} &\mathrm{E}\left[\|\varepsilon_n\|^6 \right]^{\frac{1}{6}} \mathrm{E}\left[|x_n|^6 \right]^{\frac{1}{6}} \mathrm{E}\left[\left| \tilde{R}^{-1} - R^{-1} \right|^3 \right]^{\frac{1}{3}} \\ &\times \mathrm{E}\left[\left\| \left((N-b+1)^{\frac{1}{2}} (N-b+1)^{-\frac{1}{2}} \sum_{t=b}^{N} \varepsilon_t x_t \right)^\top \right\|^3 \right]^{\frac{1}{3}}. \end{aligned}$$

which is an $O\left(n^{-1/2} l_n \right)$ thereby vanishing asymptotically. Lastly, Condition (C6), Lemma 1, and Hölder's inequality imply that $\|(249)\|$ is bounded by

$$\begin{aligned} &\mathrm{E}\left[\|\mathrm{E}\left[\varepsilon_t x_t \mid \mathcal{F}_{t-l_n} \right]\|^3 \right]^{\frac{1}{3}} \mathrm{E}\left[\left| \tilde{R}^{-1} - R^{-1} \right|^3 \right]^{\frac{1}{3}} \\ &\times \mathrm{E}\left[\left\| a^{\frac{1}{2}} a^{-\frac{1}{2}} \sum_{t=1}^{a} \varepsilon_t^\top x_t \right\|^3 \right]^{\frac{1}{3}} = o(1) \end{aligned}$$

and this completes the proof. □

Proposition 4. *Under assumptions of Theorem 1, it holds that:*

$$N(\mathrm{II}) = (\mathrm{IV}) + o(1), \tag{251}$$

where

$$(\mathrm{II}) = \mathrm{E}\left[\hat{R}^{-1}\hat{\mathbf{\Sigma}}x_n x_n \hat{\mathbf{\Sigma}}^\top \hat{R}^{-1}\right], \quad (\mathrm{IV}) = \mathrm{E}\left[\hat{\mathbf{\Sigma}}_B R^{-1}\hat{\mathbf{\Sigma}}_B^\top\right],$$

with $\hat{\mathbf{\Sigma}}$ *being defined in Proposition 3 and* $\hat{\mathbf{\Sigma}}_B = N^{-\frac{1}{2}}\sum_{t=1}^{N}\boldsymbol{\varepsilon_t}x_t$.

Proof. Let $\boldsymbol{M}_1 = x_n\left(\hat{R}^{-1} - R^{-1}\right)\hat{\mathbf{\Sigma}}_B$ and $\boldsymbol{M}_2 = x_n R^{-1}\hat{\mathbf{\Sigma}}_B$. Since

$$\begin{aligned} N\,(\mathrm{II}) &= \mathrm{E}\left[(\boldsymbol{M}_1 + \boldsymbol{M}_2)\,(\boldsymbol{M}_1 + \boldsymbol{M}_2)^\top\right] \\ &= \mathrm{E}\left[\boldsymbol{M}_1\boldsymbol{M}_1^\top\right] + \mathrm{E}\left[\boldsymbol{M}_2\boldsymbol{M}_2^\top\right] + \mathrm{E}\left[\boldsymbol{M}_1\boldsymbol{M}_2^\top\right] \\ &\quad + \mathrm{E}\left[\boldsymbol{M}_2\boldsymbol{M}_1^\top\right] \end{aligned}$$

the proof of (251) reduces to show that the following conditions hold:

$$\left\|\mathrm{E}\left[\boldsymbol{M}_1\boldsymbol{M}_1^\top\right]\right\| = o\,(1)\,, \tag{252}$$

$$\left\|\mathrm{E}\left[\boldsymbol{M}_1\boldsymbol{M}_2^\top\right]\right\| = o\,(1)\,, \tag{253}$$

$$\left\|\mathrm{E}\left[\boldsymbol{M}_2\boldsymbol{M}_2^\top\right] - (\mathrm{IV})\right\| = o\,(1)\,. \tag{254}$$

Conditions (252) and (253) readily follow from Assumptions (C3) and (C4), Lemma 1, the non singularity of R and Hölder's inequality:

$$\begin{aligned} \mathrm{E}\left[\left\|\boldsymbol{M}_1\boldsymbol{M}_1^\top\right\|\right] &= \mathrm{E}\left[\left\|x_n^2\left(\hat{R}^{-1} - R^{-1}\right)^2\hat{\mathbf{\Sigma}}_B\hat{\mathbf{\Sigma}}_B^\top\right\|\right] \\ &\leq \left(\mathrm{E}\left[|x_n|^{10}\right]\right)^{\frac{1}{5}}\left(\mathrm{E}\left[\left|\hat{R}^{-1} - R^{-1}\right|^5\right]\right)^{\frac{2}{5}} \\ &\quad \times \left(\mathrm{E}\left[\left\|\hat{\mathbf{\Sigma}}_B\right\|^5\right]\right)^{\frac{2}{5}} = o\,(1)\,; \\ \mathrm{E}\left[\left\|\boldsymbol{M}_1\boldsymbol{M}_2^\top\right\|\right] &= \mathrm{E}\left[\left\|x_n^2\left(\hat{R}^{-1} - R^{-1}\right)R^{-1}\hat{\mathbf{\Sigma}}_B\hat{\mathbf{\Sigma}}_B^\top\right\|\right] \\ &\leq \left(\mathrm{E}\left[|x_n|^{10}\right]\right)^{\frac{1}{5}}\left(\mathrm{E}\left[\left|\hat{R}^{-1} - R^{-1}\right|^5\right]\right)^{\frac{1}{5}} \\ &\quad \times \left(\mathrm{E}\left[\left|R^{-1}\right|^5\right]\right)^{\frac{1}{5}}\left(\mathrm{E}\left[\left\|\hat{\mathbf{\Sigma}}_B\right\|^5\right]\right)^{\frac{2}{5}} = o\,(1)\,. \end{aligned}$$

As concerns (254), decompose the vector $\hat{\mathbf{\Sigma}}_B$ as follows:

$$\hat{\mathbf{\Sigma}}_B = N^{-\frac{1}{2}}\sum_{t=1}^{N}\boldsymbol{\varepsilon_t}x_t = \boldsymbol{u} + \boldsymbol{w},$$

with $\boldsymbol{u} = N^{-\frac{1}{2}} \sum_{t=1}^{a} \boldsymbol{\varepsilon_t} x_t$ and $\boldsymbol{w} = N^{-\frac{1}{2}} \sum_{t=b}^{N} \boldsymbol{\varepsilon_t} x_t$. Hence, we have that

$$
\begin{aligned}
&\mathrm{E}\left[\boldsymbol{M}_2 \boldsymbol{M}_2^{\top}\right] - (\mathrm{IV}) \\
&= \mathrm{E}\left[\boldsymbol{u} R^{-1} x_n x_n R^{-1} \boldsymbol{u}^{\top}\right] - \mathrm{E}\left[\boldsymbol{u} R^{-1} R R^{-1} \boldsymbol{u}^{\top}\right] \\
&+ \mathrm{E}\left[\boldsymbol{u} R^{-1} x_n x_n R^{-1} \boldsymbol{w}^{\top}\right] - \mathrm{E}\left[\boldsymbol{u} R^{-1} R R^{-1} \boldsymbol{w}^{\top}\right] \\
&+ \mathrm{E}\left[\boldsymbol{w} R^{-1} x_n x_n R^{-1} \boldsymbol{u}^{\top}\right] - \mathrm{E}\left[\boldsymbol{w} R^{-1} R R^{-1} \boldsymbol{u}^{\top}\right] \\
&+ \mathrm{E}\left[\boldsymbol{w} R^{-1} x_n x_n R^{-1} \boldsymbol{w}^{\top}\right] - \mathrm{E}\left[\boldsymbol{w} R^{-1} R R^{-1} \boldsymbol{w}^{\top}\right].
\end{aligned}
$$

The law of iterated expectations implies that:

$$
\begin{aligned}
&\left\| \mathrm{E}\left[\boldsymbol{M}_2 \boldsymbol{M}_2^{\top}\right] - (\mathrm{IV}) \right\| \\
&\leq \left\| \mathrm{E}\left[\boldsymbol{u} R^{-1} \left(\mathrm{E}\left[x_n^2 \mid \mathcal{F}_{n-l_n}\right] - R \right) R^{-1} \boldsymbol{u}^{\top}\right] \right\| && (255) \\
&+ \left\| \mathrm{E}\left[\boldsymbol{u} R^{-1} \left(\mathrm{E}\left[x_n^2 \mid \mathcal{F}_{n-l_n}\right] - R \right) R^{-1} \boldsymbol{w}^{\top}\right] \right\| && (256) \\
&+ \left\| \mathrm{E}\left[\boldsymbol{w} R^{-1} \left(\mathrm{E}\left[x_n^2 \mid \mathcal{F}_{n-l_n}\right] - R \right) R^{-1} \boldsymbol{u}^{\top}\right] \right\| && (257) \\
&+ \left\| \mathrm{E}\left[\boldsymbol{w} R^{-1} \left(\mathrm{E}\left[x_n^2 \mid \mathcal{F}_{n-l_n}\right] - R \right) R^{-1} \boldsymbol{w}^{\top}\right] \right\|. && (258)
\end{aligned}
$$

By using arguments previously developed, it is easy to see that, under Assumptions (C4) and (C6), (255) – (258) asymptotically vanish. Therefore, conditions (252) – (254) are fulfilled and the proof is completed. □

Proposition 5. *Under assumptions of Theorem 1, it holds that:*

$$
(\mathrm{III}) = -(D) + o(1), \tag{259}
$$

where

$$
(D) = \mathrm{E}\left[R^{-1} \left[\sum_{j=h}^{N-1} \left\{ (\boldsymbol{\varepsilon}_1 x_1) (\boldsymbol{\varepsilon}_{j+1} x_{j+1})^{\top} + (\boldsymbol{\varepsilon}_{j+1} x_{j+1}) (\boldsymbol{\varepsilon}_1 x_1)^{\top} \right\} \right] \right]
$$

Proof. The result readily follows upon noting that, under Assumption (C2) and the weakly stationarity of the process $\{x_t\}$, it holds that:

$$
\begin{aligned}
(\mathrm{III}) &= -\sum_{t=1}^{N} \mathrm{E}\left[R^{-1} \left\{ (\boldsymbol{\varepsilon}_t x_t) (\boldsymbol{\varepsilon}_n x_n)^{\top} + (\boldsymbol{\varepsilon}_n x_n) (\boldsymbol{\varepsilon}_t x_t)^{\top} \right\} \right] \\
&= -\sum_{j=h}^{n-1} \mathrm{E}\left[R^{-1} \left\{ (\boldsymbol{\varepsilon}_1 x_1) (\boldsymbol{\varepsilon}_{j+1} x_{j+1})^{\top} + (\boldsymbol{\varepsilon}_{j+1} x_{j+1}) (\boldsymbol{\varepsilon}_1 x_1)^{\top} \right\} \right] \\
&= -\mathrm{E}\left[R^{-1} \left(\sum_{j=h}^{N-1} \left\{ (\boldsymbol{\varepsilon}_1 x_1) (\boldsymbol{\varepsilon}_{j+1} x_{j+1})^{\top} + (\boldsymbol{\varepsilon}_{j+1} x_{j+1}) (\boldsymbol{\varepsilon}_1 x_1)^{\top} \right\} \right) \right] \\
&+ o(1).
\end{aligned}
$$

□

Proposition 6. *Under assumptions of Theorem 1, it holds that:*

$$(\mathrm{IV}) = (1) + (Q) + (D) + o(1), \tag{260}$$

where

$$(1) = N^{-1}\,\mathrm{E}\left[R^{-1}\left\{\sum_{t=1}^{N}(\boldsymbol{\varepsilon}_t x_t)\,(\boldsymbol{\varepsilon}_t x_t)^\top\right\}\right],$$

$$(Q) = \mathrm{E}\left[R^{-1}\left[\sum_{s=1}^{h-1}\left\{(\boldsymbol{\varepsilon}_1 x_1)\,(\boldsymbol{\varepsilon}_{s+1} x_{s+1})^\top + (\boldsymbol{\varepsilon}_{s+1} x_{s+1})\,(\boldsymbol{\varepsilon}_1 x_1)^\top\right\}\right]\right]$$

$$(D) = \mathrm{E}\left[R^{-1}\left[\sum_{j=h}^{N-1}\left\{(\boldsymbol{\varepsilon}_1 x_1)\,(\boldsymbol{\varepsilon}_{j+1} x_{j+1})^\top + (\boldsymbol{\varepsilon}_{j+1} x_{j+1})\,(\boldsymbol{\varepsilon}_1 x_1)^\top\right\}\right]\right]$$

Proof. Let

$$(2) = N^{-1}\,\mathrm{E}\left[R^{-1}\left\{\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}(\boldsymbol{\varepsilon}_j x_j)\,(\boldsymbol{\varepsilon}_k x_k)^\top\right\}\right],$$

and note that $(\mathrm{IV}) - (1) = (2) + (2)^\top$. Moreover

$$\begin{aligned}(2) &= N^{-1}\,\mathrm{E}\left[R^{-1}\left\{\sum_{j=1}^{N-1}(N-j)\,(\boldsymbol{\varepsilon}_1 x_1)\,(\boldsymbol{\varepsilon}_{j+1} x_{j+1})^\top\right\}\right] \\ &= \mathrm{E}\left[R^{-1}\left\{\sum_{j=1}^{N-1}(\boldsymbol{\varepsilon}_1 x_1)\,(\boldsymbol{\varepsilon}_{j+1} x_{j+1})^\top\right\}\right] \qquad (261) \\ &\quad - N^{-1}\,\mathrm{E}\left[R^{-1}\left\{\sum_{j=1}^{N-1} j\,(\boldsymbol{\varepsilon}_1 x_1)\,(\boldsymbol{\varepsilon}_{j+1} x_{j+1})^\top\right\}\right]. \qquad (262)\end{aligned}$$

Assumptions (C2) implies that (262) is $o(1)$. Since (261) can be written as

$$\begin{aligned}&\mathrm{E}\left[R^{-1}\left\{\sum_{s=1}^{h-1}(\boldsymbol{\varepsilon}_1 x_1)\,(\boldsymbol{\varepsilon}_{s+1} x_{s+1})^\top\right\}\right] \\ &+ \mathrm{E}\left[R^{-1}\left\{\sum_{j=h}^{N-1}(\boldsymbol{\varepsilon}_1 x_1)\,(\boldsymbol{\varepsilon}_{j+1} x_{j+1})^\top\right\}\right],\end{aligned}$$

then $(261) + (261)^\top = (Q) + (D)$ and this completes the proof. □

Proof of Theorem 1

We prove that:

$$N\left\{\mathrm{E}\left[(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})\,(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})^\top - \mathrm{E}\left[\boldsymbol{\varepsilon}_n^{(h)}\boldsymbol{\varepsilon}_n^{(h)\top}\right]\right]\right\}$$

is equal to

$$R^{-1}\,\mathrm{E}\left[\left(\boldsymbol{\varepsilon}_1^{(h)}x_1\right)\left(\boldsymbol{\varepsilon}_1^{(h)}x_1\right)^\top\right] \tag{263}$$

$$+\,R^{-1}\,\mathrm{E}\left[\sum_{s=1}^{h-1}\left\{\left(\boldsymbol{\varepsilon}_1^{(h)}x_1\right)\left(\boldsymbol{\varepsilon}_{s+1}^{(h)}x_{s+1}\right)^\top+\left(\boldsymbol{\varepsilon}_{s+1}^{(h)}x_{s+1}\right)\left(\boldsymbol{\varepsilon}_1^{(h)}x_1\right)^\top\right\}\right] \tag{264}$$

$$+\,o(1).$$

Since

$$\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=\hat{R}^{-1}\left(N^{-1}\sum_{t=1}^{N}x_t\boldsymbol{y}_{t+h}\right)-\boldsymbol{\beta}=\hat{R}^{-1}\left(N^{-1}\sum_{t=1}^{N}x_t\boldsymbol{\varepsilon}_t\right),$$

routine algebra implies that:

$$\mathrm{E}\left[\left(\mathbf{y}_{n+h}-\hat{\mathbf{y}}_{n+h}\right)\left(\mathbf{y}_{n+h}-\hat{\mathbf{y}}_{n+h}\right)^\top\right]-\mathrm{E}\left[\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^\top\right]=(\mathrm{I})+(\mathrm{II}). \tag{265}$$

By applying Propositions 3 – Propositions 6, we have:

$$N\left\{\mathrm{E}\left[\left(\mathbf{y}_{n+h}-\hat{\mathbf{y}}_{n+h}\right)\left(\mathbf{y}_{n+h}-\hat{\mathbf{y}}_{n+h}\right)^\top-\mathrm{E}\left[\boldsymbol{\varepsilon}_n^{(h)}\boldsymbol{\varepsilon}_n^{(h)\top}\right]\right]\right\}$$
$$=N(\mathrm{I})+N(\mathrm{II})=(\mathrm{III})+(\mathrm{IV})+o(1)=(1)+(Q)+o(1).$$

The proof is completed upon noting that $(1)=(263)$ and $(Q)=(264)$.

4.5.2 *Proof of Theorem 2*

We start proving that

$$\hat{\mathbf{MI}}_h=\mathbf{MI}_h+O_p(n^{-1/2}). \tag{266}$$

Note that

$$\hat{\mathbf{MI}}_h=N^{-1}\left(\sum_{t=1}^{N}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top\right)-\left(N^{-1}\sum_{t=1}^{N}x_t\boldsymbol{\varepsilon}_t\right)\hat{R}^{-1}\left(N^{-1}\sum_{s=1}^{N}x_s\boldsymbol{\varepsilon}_s\right)^\top$$

hence, it holds that $\hat{\mathbf{MI}}_h-\mathbf{MI}_h$ equals

$$N^{-1}\left\{\sum_{t=1}^{N}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top-\mathrm{E}\left[\boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_1^\top\right]\right)\right\} \tag{267}$$

$$-\left(N^{-1}\sum_{t=1}^{N}x_t\boldsymbol{\varepsilon}_t\right)\hat{R}^{-1}\left(N^{-1}\sum_{t=1}^{N}x_t\boldsymbol{\varepsilon}_t\right)^\top. \tag{268}$$

Assumption (A1) implies that $(334)=O_p(n^{-1/2})$ whereas, by combining Assumptions (A3) and (A4) with the non-singularity of R and Hölder's inequality, it can be shown that $(335)=O_p(n^{-1})$ and hence the proof of (266) is complete.

Next, we prove that

$$\hat{\mathbf{VI}}_h = \mathbf{VI}_h + o_p(1).$$

It suffices to show that

$$\hat{\mathbf{C}}_{h,s} = \mathbf{C}_{h,s} + o_p(1). \tag{269}$$

It holds that $\hat{\mathbf{C}}_{h,s}$ is equal to

$$(N-s)^{-1} \sum_{t=1}^{N-s} \left(x_t \boldsymbol{\varepsilon}_t^\top\right)^\top \left(x_{t+s} \boldsymbol{\varepsilon}_{t+s}^\top\right) \tag{270}$$

$$- (N-s)^{-1} \sum_{t=1}^{N-s} x_t^2 x_{t+s} \left(\hat{\boldsymbol{\beta}}_n(h) - \boldsymbol{\beta}_h\right) \boldsymbol{\varepsilon}_{t+s}^\top \tag{271}$$

$$- (N-s)^{-1} \sum_{t=1}^{N-s} x_t x_{t+s}^2 \boldsymbol{\varepsilon}_t \left(\hat{\boldsymbol{\beta}}_n(h) - \boldsymbol{\beta}_h\right)^\top \tag{272}$$

$$+ (N-s)^{-1} \sum_{t=1}^{N-s} x_t^2 x_{t+s}^2 \left(\hat{\boldsymbol{\beta}}_n(h) - \boldsymbol{\beta}_h\right) \left(\hat{\boldsymbol{\beta}}_n(h) - \boldsymbol{\beta}_h\right)^\top. \tag{273}$$

We prove that (271) is $o_p(1)$ componentwise. To this end consider:

$$\mathrm{E}\left[(N-s)^{-1} \|*\| \sum_{t=1}^{N-s} x_t^2 x_{t+s} \varepsilon_{t+s,i}\right],$$

with $\varepsilon_{t+s,i}$ being the i-th component of the vector $\boldsymbol{\varepsilon}_{t+s}$. The triangular inequality and Hölder's inequality imply that:

$$\begin{aligned}
&\mathrm{E}\left[(N-s)^{-1} \|*\| \sum_{t=1}^{N-s} x_t^2 x_{t+s} \varepsilon_{t+s,i}\right] \\
&\leq (N-s)^{-1} \sum_{t=1}^{N-s} \mathrm{E}\left[\|*\| x_t^2 x_{t+s} \varepsilon_{t+s,i}\right] \\
&\leq (N-s)^{-1} \sum_{t=1}^{N-s} \left\{\left(\mathrm{E}\left[x_t^4\right]\right)^{1/2} \left(\mathrm{E}\left[\|*\| x_{t+s} \varepsilon_{t+s,i}{}^2\right]\right)^{1/2}\right\}.
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}_n(h) - \boldsymbol{\beta}_h = \hat{R}^{-1}\left(N^{-1}\sum_{j=1}^N x_j \boldsymbol{\varepsilon}_{j,h}\right)$, by combining Assumptions (A3), (A4) and (A5) with Chebyshev's inequality we obtain that (271) is $o_p(1)$. Similarly, we can verify that (272) and (273) are $o_p(1)$. Lastly, Condition (A2) implies that $(270) = \mathbf{C}_{h,s} + o_p(1)$, hence (269) is verified and the whole proof is complete.

4.5.3 *Proof of Theorem 3*

By Theorem 2 the VMRIC_h defined in (225) can be written as:

$$\begin{aligned}
&\mathrm{VMRIC}_h\left(\hat{\ell}_h\right) = \\
&\min_{1\leq \ell \leq K} \left\|\mathbf{MI}_h + O_p(n^{-1/2})\right\| + \min_{\ell \in M_1} \left\|\frac{\alpha_n}{n}\mathbf{VI}_h + o_p\left(\frac{\alpha_n}{n}\right)\right\|.
\end{aligned} \tag{274}$$

Therefore,

$$\lim_{n\to\infty} \mathrm{VMRIC}_h\left(\hat{\ell}_h\right) = \min_{1\le\ell\le K} \|\mathbf{MI}_h\| \tag{275}$$

and hence

$$\lim_{n\to+\infty} \Pr\left(\hat{\ell}_h \in M_1\right) = 1. \tag{276}$$

Now, consider two models ℓ_1 and ℓ_2 in the candidates set $J_{\ell_1}, J_{\ell_2} \in M_1$ such that $\mathrm{VI}_h(\ell_1) \neq \mathrm{VI}_h(\ell_2)$. We show that

$$\lim_{n\to\infty} \Pr\Bigg[\mathrm{sign}\left\{\mathrm{VMRIC}_h(\ell_1) - \mathrm{VMRIC}_h(\ell_2)\right\} = \mathrm{sign}\left\{\|\mathbf{VI}_h(\ell_1)\| - \|\mathbf{VI}_h(\ell_2)\|\right\}\Bigg] = 1. \tag{277}$$

By defining $\mathbf{MI}_h^*$ to be the minimum value of $\mathbf{MI}_h$ over the family of candidate models, we have:

$$\mathrm{VMRIC}_h(\ell_1) = \left\|\mathbf{MI}_h^* + O_p(n^{-1/2})\right\| + \left\|\frac{\alpha_n}{n}\mathbf{VI}_h(\ell_1) + o_p\left(\frac{\alpha_n}{n}\right)\right\|,$$
$$\mathrm{VMRIC}_h(\ell_2) = \left\|\mathbf{MI}_h^* + O_p(n^{-1/2})\right\| + \left\|\frac{\alpha_n}{n}\mathbf{VI}_h(\ell_2) + o_p\left(\frac{\alpha_n}{n}\right)\right\|.$$

Therefore, for sufficiently large n, it holds that:

$$\mathrm{VMRIC}_h(\ell_1) - \mathrm{VMRIC}_h(\ell_2) = \left\|\frac{\alpha_n}{n}\right\| \left(\|\mathbf{VI}_h(\ell_1)\| - \|\mathbf{VI}_h(\ell_2)\|\right).$$

Thus

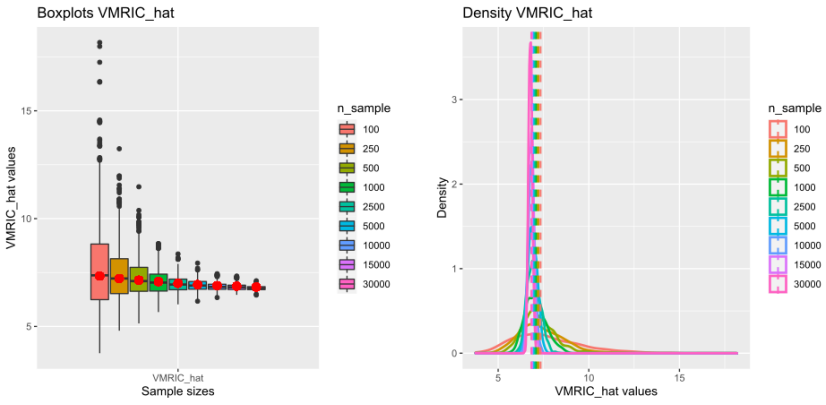
$$\mathrm{sign}\left\{\mathrm{VMRIC}_h(\ell_1) - \mathrm{VMRIC}_h(\ell_2)\right\} = \mathrm{sign}\left\{\|\mathbf{VI}_h(\ell_1)\| - \|\mathbf{VI}_h(\ell_2)\|\right\},$$

and (352) is verified and implies that

$$\lim_{n\to\infty} \Pr\left(\hat{\ell}_h \in M_2\right) = 1. \tag{278}$$

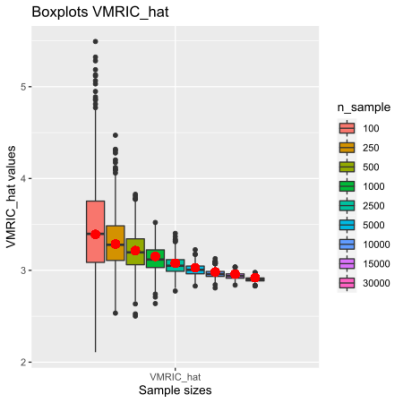
This completes the proof.

4.6 FIGURES AND TABLES

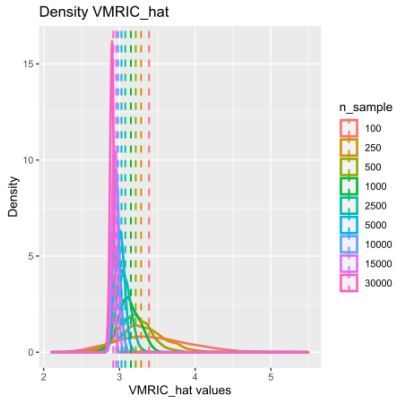


(a) Case 1 - Boxplots

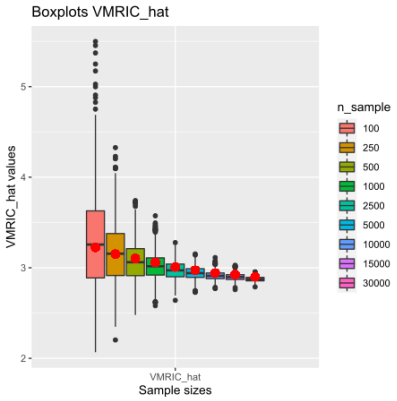
(b) Case 1 - Densities



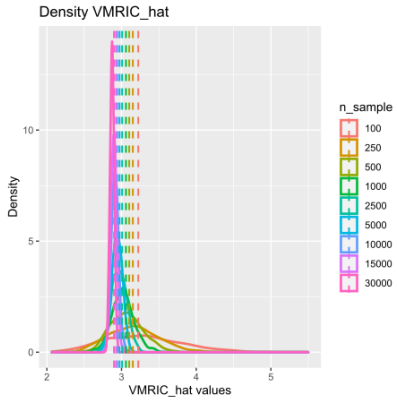
(c) Case 2 - Boxplots



(d) Case 2 - Densities



(e) Case 3 - Boxplots



(f) Case 3 - Densities

Figure 1: Consistency of the $\widehat{\text{VMRIC}}_2$ for Model 1 for three different combinations of parameters. The red dot indicates the population VMRIC_2.

Table 5: Additional parameters' combinations for the DGP of Eq. (228), (229), and (237).

Case	ϕ_1	ϕ_2	a_1	a_2	ψ_1
4	0.4	-0.75	1.5	-2	-0.25
5	-0.4	-0.45	-0.75	-1.25	0.65
6	-0.33	-0.66	1	0.5	-0.8

Table 6: Additional theoretical and estimated VMRIC of Models 1 and 2, for the three parameterizations of Table 5, computed on a data set of $n = 10^6$ observations.

	Model 1		Model 2	
Case	VMRIC	$\widehat{\text{VMRIC}}$	VMRIC	$\widehat{\text{VMRIC}}$
4	6.671	6.636	7.537	7.530
5	3.682	3.667	3.941	3.935
6	2.814	2.805	3.081	3.084

Table 7: Percentages of correctly selected models by the three information criteria for the three additional parametrizations and varying sample size n.

	Case 4			Case 5			Case 6		
n	VMRIC	AIC	BIC	VMRIC	AIC	BIC	VMRIC	AIC	BIC
100	74.5	51.6	51.6	71.8	58.7	58.7	77.9	51.1	51.1
1000	97.6	65.5	65.5	93.2	69.6	69.6	99.5	59.7	59.7
10000	100	88.0	88.0	100	95.4	95.4	100	72.7	72.7

5

THE FULL MULTIVARIATE EXTENSION OF THE MRIC: MULTIVARIATE RESPONSE AND MULTIVARIATE PREDICTOR

ABSTRACT

We extend the Vectorial MRIC (VMRIC) proposed in Chapter 4 to the case where the response is a multivariate time series with multiple predictor. We obtain an asymptotic expression for the Mean-Squared Prediction Error (MSPE) matrix which allows us to define the VMRIC, derive its Method-of-Moments Estimator (MoME), prove its asymptotic consistency, and show that the VMRIC is an asymptotically efficient criterion for h-step ahead possibly-misspecified vector time series models with multiple regressor. Remarks on the type of models satisfying the technical conditions are advanced for vector autoregressive models with exogenous variables (VARX), also known as dynamic simultaneous equations models.

Keywords: multivariate time series, multiple regressor, MSPE matrix, information criteria, vectorial MRIC, VARX
2020 MSC: Primary 62H12, Secondary 62F12

5.1 INTRODUCTION

The model selection step is a fundamental task in statistical modelling and its implementation typically depends upon the objective of the exercise. In the time series framework the focus is on either forecasting future values or describing/controlling the process that has generated the data (DGP). A good model selection criterion must feature a good ability to identify the model with the "best" fit to future values, in a specified sense. In particular, in the parametric time series framework, we can identify two main properties. The first one is consistency, i.e., the ability to select the true DGP with probability one as the sample size diverges. This assumes that a true model exists and that it is among the set of candidate models. If either the set of candidate models does not contain the true DGP, or, for some reason, a true model cannot be postulated, then a selection criterion should be asymptotically efficient, for instance, in the mean square sense, i.e. it minimizes the mean squared prediction error as the sample size diverges. Starting from the seminal work of Akaike, [6] a plethora of model selection criteria has been proposed. These include Akaike's AIC [6, 8], Schwarz's Bayesian Information Criterion (BIC) [258], and Rissanen's Minimum

Description Length (MDL) [242]. Such criteria paved the way for various extensions dealing with different unsolved issues. For instance, the AIC is efficient but not consistent (i.e. it leads to select overfitting models), whereas the BIC is consistent but not efficient, see [153] for a discussion.

A recent development for model selection in possibly misspecified parametric time series models in the fixed-dimensionality setting is given by the Misspecification-Resistant Information Criterion (hereafter MRIC) [153]. Fixed-dimensionality means that the number of observations increases to infinity while the number of 'true' parameters is finite. In this respect, the MRIC provides a solution to the original research question of Akaike: it enjoys both consistency, in case the true model is included as a candidate, and asymptotic efficiency when a true model either cannot be assumed or is not included. Moreover, when the number of variables in the model grows with the sample size, the MRIC can achieve asymptotic efficiency, without the need for additional criteria. Finally, in the high-dimensional setting, the MRIC can be used together with appropriate model selection criteria to identify the best predictive models. The MRIC is based upon the additive decomposition of the mean squared prediction error in a term that depends upon the misspecification level and a term that measures the sampling variability of the predictor. The idea is to select the model with smallest variability among those that minimize the misspecification index.

After showing in Chapter 4 that the multivariate extension of the MRIC with a univariate regressor is viable, the present chapter shows the first full extension of the MRIC, the VMRIC, to possibly-misspecified multivariate time models with multiple regressor in h-step ahead prediction. Only a few conditions had to be adapted in order to obtain the asymptotic decomposition of the MSPE, which follows similarly the structure as in the univariate case, but displays the presence of quadratic forms for the Variability Index (VI) matrix. This is detailed in the proof of Theorem 4, and in the second part of the proof of Theorem 5 for the asymptotic consistency of the method-of-moments (MoM) estimator. All the proofs are however reported for readability.

The rest of the chapter is organized as follows: in Section 5.2 we recall and update the notation; in Section 5.3 we extend the MRIC approach to multivariate time series with multiple regressor. In particular, in Section 5.3.1 we obtain the general asymptotic decomposition of the MSPE matrix into two parts, as before: the first one is linked to the goodness of fit of the model and the second one depends upon the prediction variance. In Section 5.3.2 we present the VMRIC and derive a consistent estimator for it, whereas in Section 5.3.3, we prove the asymptotic efficiency of the VMRIC as in the univariate case. Section 5.4 presents a digression on the conditions for possibly-misspecified vector autoregressive models with exogenous variables (VARX), or dynamic simul-

taneous equations models. All the proofs are detailed in Section 5.5. Appendix B.1 contains the auxiliary technical lemma.

5.2 NOTATION AND PRELIMINARIES

Let us consider two weakly stationary stochastic processes $\{\mathbf{y}_t\} \in \mathbb{R}^w$ and $\{\mathbf{x}_t\} \in \mathbb{R}^m$, with $w, m \in \mathbb{N}^+$, defined in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We observe the sample $(\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}, \{\mathbf{x}_1, \ldots, \mathbf{x}_n\})$, with $t = \{1, 2, \ldots, N, N+1, \ldots, N+h = n\}$. Define the sample means $\bar{\mathbf{y}} = n^{-1} \sum_{t=1}^{n} \mathbf{y}_t$, and $\bar{\mathbf{x}} = n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t$. We replace the unconditional expectations, $E[\mathbf{y}_{t+1}]$ and $E[\mathbf{x}_t]$, with their respective sample counterparts, considering the complete set of n observations for efficiency. This because the differences *(a)* between $(\mathbf{y}_{t+1} - E[\mathbf{y}_{t+1}])$ and $(\mathbf{y}_{t+1} - \bar{\mathbf{y}})$; and *(b)* between $(\mathbf{x}_t - E[\mathbf{x}_t])$ and $(\mathbf{x}_t - \bar{\mathbf{x}})$, vanish asymptotically. Without lack of generality, assume $E[\mathbf{y}_t] = \mathbf{0}$, and $E[\mathbf{x}_t] = \mathbf{0}$. In order to forecast $\mathbf{y}_{n+h}$, $h \geq 1$, we adopt the following h-step ahead forecasting model:

$$\mathbf{y}_{t+h} = \boldsymbol{\beta}_h \mathbf{x}_t + \boldsymbol{\varepsilon}_t^{(h)}, \tag{279}$$

where $\boldsymbol{\beta}_h$ is a $(w \times m)$ matrix parameters' matrix defined as the pseudo-true parameter for the possibly-misspecified model

$$\boldsymbol{\beta}_h = \underset{\mathbf{C} \in \mathbb{R}^{(w \times m)}}{\operatorname{argmin}} \; E\left[(\mathbf{y}_{t+h} - \mathbf{C}\mathbf{x}_t) (\mathbf{y}_{t+h} - \mathbf{C}\mathbf{x}_t)^\top \right], \tag{280}$$

and $\boldsymbol{\varepsilon}_t^{(h)}$ is the w-length vector containing the h-step ahead forecasting errors. The prediction error vector $\boldsymbol{\varepsilon}_t^{(h)}$ can be both serially and cross-correlated, and also correlated with $\mathbf{x}_s$, $s \neq t$.[1]

Remark 13. *This general definition of the model includes:* (i) *the setting of multi-step forecasting with $h < 1$;* (ii) *exogenous and endogenous variables in $\mathbf{x}$; and* (iii) *cases where the prediction error vector $\boldsymbol{\varepsilon}_t^{(h)}$ can be serially correlated, but also correlated with $\mathbf{x}_s$, $s \neq t$. Moreover, the multivariate framework differs from [153] in different key aspects. For instance, (a) the components of the error vector can be cross-correlated, and (b) $\mathbf{x}_t \boldsymbol{\varepsilon}_t^{(h)}$ and $\mathbf{x}_k \boldsymbol{\varepsilon}_k^{(h)}$, for $t \neq k$, can also be both serially and cross correlated. Besides, as in the original MRIC, $\mathbf{x}$ may vary also with h, but for notational simplicity, it is avoided.*

Define

$$\hat{\mathbf{R}} = N^{-1} \sum_{t=1}^{N} \mathbf{x}_t \mathbf{x}_t^\top \qquad \text{and} \qquad \mathbf{R} = E[\mathbf{x}_1 \mathbf{x}_1^\top]. \tag{281}$$

Then, the Ordinary least squares (OLS) estimator of $\boldsymbol{\beta}_h$ results:

$$\hat{\boldsymbol{\beta}}_n(h) = \hat{\mathbf{R}}^{-1} \left(N^{-1} \sum_{t=1}^{N} \mathbf{x}_t \mathbf{y}_{t+h}^\top \right). \tag{282}$$

1 The following Remark 13 is the same as Remark 3, included here for convenience.

The prediction of $\mathbf{y}_{n+h}$, $h \geq 1$, is given by

$$\hat{\mathbf{y}}_{n+h} = \hat{\boldsymbol{\beta}}_n(h)\mathbf{x}_n \tag{283}$$

and the corresponding Mean Squared Prediction Error matrix is

$$\mathbf{MSPE}_h = E\left[(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})(\mathbf{y}_{n+h} - \hat{\mathbf{y}}_{n+h})^\top\right]. \tag{284}$$

5.3 THE FULL MULTIVARIATE EXTENSION OF THE MRIC FRAMEWORK

In this section we extend the MRIC approach to the case where the response is a multivariate time series ($w \geq 2$) and multiple predictor ($m \geq 1$), for a generic h-step ahead forecast, $h \geq 1$. Hence, Model (279) can be written in extensive form as:

$$\begin{cases} y_{t+h,1} = \beta_{1,1}^{(h)} x_{t,1} + \beta_{1,2}^{(h)} x_{t,2} + \cdots + \beta_{1,m}^{(h)} x_{t,m} + \varepsilon_{t,1}^{(h)} \\ y_{t+h,2} = \beta_{2,1}^{(h)} x_{t,1} + \beta_{2,2}^{(h)} x_{t,2} + \cdots + \beta_{2,m}^{(h)} x_{t,m} + \varepsilon_{t,2}^{(h)} \\ \vdots \\ y_{t+h,w} = \beta_{w,1}^{(h)} x_{t,1} + \beta_{w,2}^{(h)} x_{t,2} + \cdots + \beta_{w,m}^{(h)} x_{t,m} + \varepsilon_{t,w}^{(h)}. \end{cases} \tag{285}$$

5.3.1 *Asymptotic decomposition of the MSPE matrix*

In this section we further extend the asymptotic version of the $\mathbf{MSPE}_h$ derived in Eq. (220) which allows us to write the VMRIC_h for multivariate time series responses and multiple regressor. Let

$$\mathbf{B}_t^{(h)} = \mathbf{x}_t \boldsymbol{\varepsilon}_t^{(h)\top} \tag{286}$$

and consider the following regularity conditions:

Assumptions 11.

$(C1')$ $\exists\, q_1 > 5, 0 < C_1 < \infty$:
for any $1 \leq n_1 < n_2 \leq n$, *and any* $1 \leq i, j \leq m$,
$$E\left[\left|(n_2 - n_1 + 1)^{-1/2} \sum_{t=n_1}^{n_2} x_{t,i}x_{t,j} - E\left[x_{t,i}x_{t,j}\right]\right|^{q_1}\right] \leq C_1.$$

$(C2')$ 1. $\mathbf{D}_{h,s} = E\left[\mathbf{B}_t^{(h)\top}\mathbf{B}_{t+s}^{(h)}\right] = E\left[\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t+s}^{(h)\top}\right)\operatorname{tr}\left\{\mathbf{x}_t\mathbf{x}_t^\top\right\}\right] \perp\!\!\!\perp t$,
2. $\forall\, i, j = \{1, 2, \ldots, w\}$, $\forall\, k, l = \{1, 2, \ldots, m\}$,
$E\left[\varepsilon_{1,i}^{(h)}\varepsilon_{n,j}^{(h)}x_{1,k}x_{n,l}\right] = o(n^{-1})$.

$(C3')$ 1. $\sup_{-\infty<t<\infty} E\left[\|\mathbf{x}_t\|^{10}\right] < \infty$,
2. $\sup_{-\infty<t<\infty} E\left[\left\|\boldsymbol{\varepsilon}_t^{(h)}\right\|^6\right] < \infty$.

$(C4')$ $\exists\, 0 < C_2 < \infty$: *for* $1 \leq n_1 < n_2 \leq n$,
$$E\left[\left\|(n_2 - n_1 + 1)^{-\frac{1}{2}} \sum_{t=n_1}^{n_2} \mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)\top}\right\|^5\right] < C_2.$$

$(C5')$ *For any* $q > 0$, $E\left[\left\|\hat{\mathbf{R}}^{-1}\right\|^q\right] = O(1)$.

$(C6')$ $\exists\, \mathcal{F}_t \subseteq \mathcal{F}, \mathcal{F}_t$ *an increasing sequence of* σ*-fields* :
(1) $\mathbf{x}_t$ *is* $\mathcal{F}_t$*-measurable*
(2) $\sup_{-\infty<t<\infty} E\left[\left\|E\left[\mathbf{x}_t\mathbf{x}_t^\top|\mathcal{F}_{t-k}\right] - R\right\|^3\right] = o(1)$,
(3) $\sup_{-\infty<t<\infty} E\left[\left\|E\left[\mathbf{B}_t^{(h)}|\mathcal{F}_{t-k}\right] - \mathbf{0}\right\|^3\right] = o(1)$,
as k *diverges.*

Remark 14. *To aid with the qualitative interpretation of Assumptions 11, note the following. For further details, see Section 5.4.*

(i) Condition $(C1')$ *requires the finiteness of the* q_1*-th moment, with* $q_1 > 5$*, of the difference between the sample and population second-order moment of the covariance between* $x_{t,i}$ *and* $x_{t,j}$*, i.e. component-wise. This condition depends on the square-summability of the processes composing both the dependent multivariate response and the multiple regressor. For this reason, it appears to be attainable also for vector time series under general condition. Besides, it also relies on the First Moment Bound Theorem of Findley and Wei [115], which has been proved for vector time series. For these reasons, we expect this condition to hold also in the multivariate setting. We are currently completing the proof.*

(ii) The first part of Condition $(C2')$ *involves the* s*-lag cross-(auto)covariance matrix, between the* h*-steps ahead forecast error and the regressor and requires it to be independent of time* t*. Note that the symbol* $\perp\!\!\!\perp$ *reads*

in this case "is independent of". *The first part should be ensured for processes admitting linear vectorial representation, in particular, satisfying Wold's multivariate representation theorem. Given that this is a fairly general condition, it should not present a major problem, specially if it is further assumed finite fourth-order moments of the white noise processes or martingale-difference vector sequences, considering these type of processes the building blocks of* $\{\mathbf{y}_t\}, \{\mathbf{x}_t\}, \left\{\boldsymbol{\varepsilon}_t^{(h)}\right\}$.

The second part states that this s*-lag cross-(auto)covariance matrix vanishes asymptotically, component-wise, for the maximum lag* n*, for sufficiently large sample size. In other words, it suggests that the dependence between* $\mathbf{x}_t\boldsymbol{\varepsilon}_{t,h}^\top$ *and* $\mathbf{x}_s\boldsymbol{\varepsilon}_{s,h}^\top$ *vanishes sufficiently quickly as* $|t-s|$ *diverges. This condition is fundamental to prove Propositions 1 and 5 in both univariate and multiple regressor scenario.*

(iii) Condition $(C3')$ *states uniform integrability of both the regressor and the* h*-steps ahead forecast error. This can be easily obtain for instance for possibly-misspecified VARX models with processes satisfying the conditions for Wold's representation theorem in the multivariate case (i.e. under conditions of convergence of the vector processes) and fourth-order finiteness of the white noise that compose these. This condition has been shown in Section 5.4.*

(iv) Condition $(C4')$ *requires the sample covariance between the regressor and the misspecified forecast error to be in the* $\mathcal{L}^5$ *space. In particular, it states that for a finite costant* C_2*, and for any time index* n_1 *and* n_2 *we have boundedness of the fifth-order moment of the sample covariance vector between the multiple regressor and the multivariate forecast error. For the same reasons exposed for Condition* $(C1')$*, it is expected to hold.*

(v) Condition $(C5')$ *says that for any positive order* q*, boundedness of* q*-moments of the inverse of regressors' sample variance is satisfied. In other words, it requires the inverse of the sample variance-covariance matrix to be in the* $\mathcal{L}^p$ *space, with* $p > 0$*. In the scalar case* $m = 1$*, it is shown in view of Theorem 2.1 in Chan and Ing [71] for univariate (nonlinear) stochastic regression models with applications to time series. It remains open to show that this is the case for the multivariate case, e.g. in a VARX*(p, q) *model where the regressor is composed by both processes* $\{\mathbf{s}_{t-i}\}, \{\mathbf{y}_{t-j}\}$*, with* $i = \{1, 2, \ldots, q\}$*,* $j = \{0, 1, \ldots, p\}$*, as in Section 5.4.*

(vi) Condition $(C6')$ *requires:*

- *(I)* $\mathcal{F}_t$*-measurability of* $\mathbf{x}_t$*, and*
- *(II) uniform convergence of regressors' conditional variance to their population values, and*
- *(III) uniform convergence of the conditional covariance between the* h*-steps ahead forecast error and the regressor to its population value.*

In Section 5.4 we show that under general assumptions on the underlying processes, it holds for possibly-misspecified VARX(p,q) model, where we may have $p = \infty$.

Current research is devoted to complete the proofs connected with the statements in the previous Remark 14. We are seeing promising results for a general class of multivariate time series models to satisfy these conditions. We have shown for the general case that Conditions (C3′) and (C6′) hold for vector autoregressive with exogenous variables models with infinite AR part and finite q lags for the exogenous part, i.e. VARX(∞, q), under very general conditions. Further details are included in the Section 5.4.

Remark 15. *Our natural full vectorial extension only modified conditions (C2), the second point in (C3), (C4), and the third point in (C6). The rest are the same as in the* VMRIC_h *with univariate regressor.*

Theorem 4. *Under the regularity conditions (C1′) – (C6′), the asymptotic expression of the* $\boldsymbol{MSPE}_h$ *defined in (284), for the general case $w \geq 2$ and $m \geq 2$, results*

$$
\begin{aligned}
& N\left\{E\left[\left(\mathbf{y}_{n+h}-\hat{\mathbf{y}}_{n+h}\right)\left(\mathbf{y}_{n+h}-\hat{\mathbf{y}}_{n+h}\right)^{\top}\right]-E\left[\boldsymbol{\varepsilon}_n\boldsymbol{\varepsilon}_n^{(h)\top}\right]\right\}= \\
& E\left[\boldsymbol{B}_t^{(h)\top}\boldsymbol{R}^{-1}\boldsymbol{B}_t^{(h)}\right]+\sum_{s=1}^{h-1}\left\{E\left[\boldsymbol{B}_t^{(h)\top}\boldsymbol{R}^{-1}\boldsymbol{B}_{t+s}^{(h)}\right]\right. \\
& \left.+E\left[\left(\boldsymbol{B}_t^{(h)\top}\boldsymbol{R}^{-1}\boldsymbol{B}_{t+s}^{(h)}\right)^{\top}\right]\right\} \\
& +o(1)
\end{aligned}
\tag{287}
$$

where $E\left[\boldsymbol{B}_t^{(h)\top}\boldsymbol{R}^{-1}\boldsymbol{B}_{t+s}^{(h)}\right]=E\left[\boldsymbol{\varepsilon}_1^{(h)}\boldsymbol{\varepsilon}_{t+s}^{(h)\top}\mathrm{tr}\left\{\boldsymbol{x}_1\boldsymbol{x}_{1+s}^{\top}\boldsymbol{R}^{-1}\right\}\right].$

5.3.2 *VMRIC and its consistent estimation*

In this section we introduce the VMRIC. Let $\{\alpha_n\}$ be the penalization term sequence defined as in Eq. (311).

$$
\mathrm{VMRIC}_h=\|\mathbf{MI}_h\|+\left\|\frac{\alpha_n}{n}\mathbf{VI}_h\right\| \tag{288}
$$

where

$$
\begin{aligned}
\mathbf{MI}_h &= \mathrm{E}\left[\left(\boldsymbol{\varepsilon}_t^{(h)}\boldsymbol{\varepsilon}_t^{(h)\top}\right)\right], \\
\mathbf{VI}_h &= R^{-1}\left(\mathbf{C}_{h,0}+\sum_{s=1}^{h-1}\left(\mathbf{C}_{h,s}+\mathbf{C}_{h,s}^{\top}\right)\right), \\
\mathbf{C}_{h,s} &= \mathrm{E}\left[\left(\mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)\top}\right)\left(\mathbf{x}_{t+s}\boldsymbol{\varepsilon}_{t+s}^{(h)\top}\right)^{\top}\right].
\end{aligned}
$$

The VMRIC can be estimated via the method of moments as to obtain:

$$
\widehat{\mathrm{VMRIC}}_h\equiv\left\|\hat{\mathbf{MI}}_h\right\|+\left\|\frac{\alpha_n}{n}\hat{\mathbf{VI}}_h\right\|, \tag{289}
$$

where

$$\hat{\mathbf{MI}}_h = N^{-1} \sum_{t=1}^{N} \left(\hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t^\top \right),$$
$$\hat{\mathbf{VI}}_h = \hat{R}^{-1} \left[\hat{\mathbf{C}}_{h,0} + \sum_{s=1}^{h-1} \left(\hat{\mathbf{C}}_{h,s} + \hat{\mathbf{C}}_{h,s}^\top \right) \right],$$

and $\hat{\mathbf{C}}_{h,s} = (N-s)^{-1} \sum_{t=1}^{N-s} x_t x_{t+s} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_{t+s}^\top$, with $\hat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_{t+h} - \hat{\boldsymbol{\beta}}_n(h) x_t$ the estimated forecast error vector.
In Theorem 5 below we prove that $\hat{\mathbf{MI}}_h$ and $\hat{\mathbf{VI}}_h$ are consistent estimators of $\mathbf{MI}_h$ and $\mathbf{VI}_h$, respectively. Theorem 5 relies upon the following assumptions, that are less restrictive with respect to (C1′) – (C6′). For further discussions on the assumptions see [153, Remark 1–3, p. 1073].

Assumptions 12. *For each $0 \leq s \leq h-1$, we assume the following:*

$$(A1') \quad n^{-1} \sum_{t=1}^{n} \left(\boldsymbol{\varepsilon}_t^{(h)} \boldsymbol{\varepsilon}_t^{(h)\top} \right) = \mathrm{E} \left[\boldsymbol{\varepsilon}_1^{(h)} \boldsymbol{\varepsilon}_1^{(h)\top} \right] + O_p \left(n^{-1/2} \right),$$
$$(A2') \quad n^{-1} \sum_{t=1}^{n} \left(\mathbf{x}_t \boldsymbol{\varepsilon}_t^{(h)\top} \right) \left(\mathbf{x}_{t+s} \boldsymbol{\varepsilon}_{t+s}^{(h)\top} \right)^\top = \boldsymbol{C}_{h,s} + o_p(1),$$
$$(A3') \quad n^{-1/2} \sum_{t=1}^{n} \mathbf{x}_t \boldsymbol{\varepsilon}_t^{(h)\top} = O_p(1),$$
$$(A4') \quad n^{-1} \sum_{t=1}^{n} \mathbf{x}_t \mathbf{x}_t^\top = \mathbf{R} + o_p(1),$$
$$(A5') \quad \sup_{-\infty<t<\infty} \mathrm{E} \left[\left\| \boldsymbol{\varepsilon}_t^{(h)} \right\|^4 \right] + \sup_{-\infty<t<\infty} \mathrm{E} \left[\|\mathbf{x}_t\|^8 \right] < \infty.$$

Theorem 5. *If Assumptions (A1) – (A5) hold, then for the case $w \geq 2$, and $m = 1$ we obtain:*

$$\hat{\boldsymbol{MI}}_h = \boldsymbol{MI}_h + O_p(n^{-1/2}),$$
$$\hat{\boldsymbol{VI}}_h = \boldsymbol{VI}_h + o_p(1).$$

Remark 16. *Consider the following for a qualitative aid to interpret Conditions 12:*

(i) Assumption (A1′) refers to the convergence in probability of the sample variance-covariance matrix of the misspecified forecasting error to its population value, since $O_p(n^{-1/2}) = O_p(o(1)) = o_p(1)$.[2] In the original univariate case in [153], Theorem 4.3, it is required for asymptotic efficiency across several high-dimensional time series models. Furthermore, from Remark S.2 in [154], we know that it has a similar condition for the nonlinear regression case. It has consequences on the convergence

2 The motivation behind this choice, and I conjecture also behind the original derivation in the scalar case, is because it considers also cases of boundedness in probability with specific rates, i.e. where the bounding quantity is such that itself converges to zero at a rate slower than n^{-1}.

of $\hat{\boldsymbol{M}}\boldsymbol{I}_h$ to $\boldsymbol{MI}_h$, as shown in Section 5.5.2. In our multivariate setting, it should follow from the consequences of Assumptions 13 for VARX(p, q) models, i.e. Wold's multivariate representation theorem, and from using the First Moment Bound Theorem of Findley and Wei [115], applicable to vector time series. See Section 5.4. It is the focus of current research.

(ii) Assumption (A2′) is related to Condition (iii) in Assumptions 13, and to the consequence of Wold's representation theorem.

(iii) In [153], the scalar counterparts of Assumptions (A3′), (A4′), and (A5′), are ensured by Conditions (C4′), (C1′), and (C3′).

Now, we are able to define the approximate estimate of the VMRIC for h-step ahead prediction. As we can see, it keeps the same general structure as the case with univariate regressor, but with a different composition.

Definition 72. *By Theorem 4, the* $\widehat{\mathrm{VMRIC}}_h$ *quantifying the model's performance is estimated using the MoMEs:*

$$\widehat{\mathrm{VMRIC}}_h = \left\|\hat{\boldsymbol{M}}\boldsymbol{I}_h\right\| + \left\|\frac{\alpha_n}{n}\hat{\boldsymbol{V}}\boldsymbol{I}_h\right\|, \tag{290}$$

with

$$\hat{\mathbf{M}}\mathbf{I}_h = N^{-1}\sum_{t=1}^{N}\hat{\boldsymbol{\varepsilon}}_t^{(h)}\hat{\boldsymbol{\varepsilon}}_t^{(h)\top}, \tag{291}$$

$$\hat{\mathbf{V}}\mathbf{I}_h = \hat{\boldsymbol{D}}_{h,0} + \sum_{s=1}^{h-1}\left\{\hat{\boldsymbol{D}}_{h,s} + \hat{\boldsymbol{D}}_{h,s}^{\top}\right\}, \tag{292}$$

where

$$\hat{\boldsymbol{D}}_{h,s} = (N-s)^{-1}\sum_{t=1}^{N-s}\hat{\boldsymbol{B}}_t^{(h)\top}\hat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{B}}_{t+s}^{(h)}, \tag{293}$$

$$\hat{\boldsymbol{B}}_t^{(h)} = \boldsymbol{x}_t\hat{\boldsymbol{\varepsilon}}_t^{(h)\top}. \tag{294}$$

5.3.3 *Asymptotic efficiency*

In this section we prove the asymptotic efficiency of the VMRIC in the fixed dimensionality framework. To this end, let $\mathcal{M}$ be the set of K candidate models; each model is indicated either by ℓ or κ, $1 \leq \ell, \kappa \leq K$. Define the subsets M_1 and M_2 as follows:

$$M_1 = \left\{\kappa : 1 \leq \kappa \leq K, \|\mathbf{MI}_h(\kappa)\| = \min_{1\leq\ell\leq K}\|\mathbf{MI}_h(\ell)\|\right\}, \tag{295}$$

$$M_2 = \left\{\kappa : \kappa \in M_1, \|\mathbf{VI}_h(\kappa)\| = \min_{\ell\in M_1}\|\mathbf{VI}_h(\ell)\|\right\}. \tag{296}$$

In short, for a given forecast horizon h, M_1 contains the models with the minimum $\mathbf{MI}_h$ whereas in M_2 we are minimizing $\mathbf{VI}_h$ among the candidates models in M_1. The definition of efficiency used in our framework is the same as that of [153]:

Definition 73. *Given a sample of size n, a model selection criterion is said to be asymptotically efficient if it selects the model $\hat{\ell}_h$ such that*

$$\lim_{n\to\infty} \Pr\left(\hat{\ell}_h \in M_2\right) = 1.$$

Remark 17. *Please refer to Section 2.5.2 in Chapter 2, and Remark 8 in Chapter 4 for a detailed discussion.*

The VMRIC selects the model with the smallest variability index among those that achieve the best goodness of fit. Hence, the selected model $\hat{\ell}_h$ is such that:

$$\text{VMRIC}_h\left(\hat{\ell}_h\right) \equiv \min_{1\leq\ell\leq K} \left\|\hat{\mathbf{MI}}_h(\ell)\right\| + \min_{\ell\in M_1} \left\|\frac{\alpha_n}{n}\hat{\mathbf{VI}}_h(\ell)\right\|. \tag{297}$$

In the next Theorem we show that the VMRIC is an asymptotic efficient model selection criterion in the sense of Definition 73.

Theorem 6. *Assume that for each $1 \leq \ell \leq K$, $0 \leq s \leq h-1$, Theorem 5 holds and let $\hat{\ell}_h$ be the model selected by the* VMRIC*. Then we have that:*

$$\lim_{n\to\infty} \Pr\left(\hat{\ell}_h \in M_2\right) = 1,$$

namely, the VMRIC *is asymptotically efficient in the sense of Definition 73.*

5.4 EXAMPLE: A POSSIBLY-MISSPECIFIED VECTOR AUTOREGRESSIVE WITH MULTIPLE EXOGENOUS REGRESSOR MODEL, VARX(P,Q)

We follow Lütkepohl [197], Reinsel [238], and Hansen [139] in the theoretical structure of this example.

Consider as DGP the general vector autoregressive model of order $p = \infty$ with exogenous multiple regressor with autoregressive part of order q, in structural form:

$$\mathbf{A}\mathbf{y}_t = \mathbf{A}_1^*\mathbf{y}_{t-1} + \mathbf{A}_2^*\mathbf{y}_{t-2} + \cdots + \boldsymbol{\eta}_0^*\mathbf{s}_t + \cdots + \boldsymbol{\eta}_q^*\mathbf{s}_{t-q} + \boldsymbol{\epsilon}_t, \tag{298}$$

for a w-dimensional multivariate dependent vector $\mathbf{y}_t$ of endogenous variables, an m'-dimensional multiple exogenous regressor vector $\mathbf{s}_t$ of exogenous variables, with matrix $\mathbf{A}$ of dimension $(w \times w)$ representing the instantaneous relation between the endogenous variables, coefficients' matrices $\mathbf{A}_i^*, \boldsymbol{\eta}_j^*$ of dimensions $(w \times w)$ and $(w \times m')$, with $i = \{0, 1, \dots\}$ and $j = \{1, \dots, q\}$ respectively, and a w-dimensional error vector $\boldsymbol{\epsilon}_t$.

If $\{\boldsymbol{\epsilon}_t\}$ is a White Noise vector process, i.e. $E[\boldsymbol{\epsilon}_t] = \mathbf{0}$, $E\left[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_t^\top\right] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ nonsingular, and $E\left[\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_s^\top\right] = \mathbf{0}$ for $s \neq t$, then it is defined as a VARX(∞, q) model or *dynamic simultaneous equations model*. Notice that $\mathbf{s}_t$ may contain both stochastic and nonstochastic components. Let $\mathbf{A} = \boldsymbol{I}_w$, and $\boldsymbol{\eta}_0^* = \mathbf{0}$, so we can write the reduced form version, usually employed for forecasting, multiplier analysis or control:

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\eta}_0 \mathbf{s}_t + \boldsymbol{\eta}_1 \mathbf{s}_{t-1} + \cdots + \boldsymbol{\eta}_q \mathbf{s}_{t-q} + \mathbf{u}_t, \tag{299}$$

where matrices $\mathbf{A}_i \equiv \mathbf{A}^{-1}\mathbf{A}_i^*$, $i = \{1, 2, \dots\}$, and $\boldsymbol{\eta}_j \equiv \mathbf{A}^{-1}\boldsymbol{\eta}_j^*$, $j = \{0, 1, \dots q\}$, are usually nonlinear functions of the reduced parameters, $\mathbf{u}_t \equiv \mathbf{A}^{-1}\boldsymbol{\epsilon}_t$ the transformed errors. Notice that the reduced form assumes that $\mathbf{A}^{-1}$ exists, which is guaranteed by $\mathbf{A} = \boldsymbol{I}_w$ in this example. Its reduced form with the lag operator becomes:

$$\mathbf{A}(B)\mathbf{y}_t = \boldsymbol{\eta}(B)\mathbf{s}_t + \boldsymbol{\epsilon}_t, \tag{300}$$

where $\mathbf{A}(B) = \boldsymbol{I}_w - \mathbf{A}_1 B - \mathbf{A}_2 B^2 - \cdots = \sum_{j=0}^{\infty} \mathbf{A}_j B^j$, and $\boldsymbol{\eta}(B) = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1 B + \cdots + \boldsymbol{\eta}_q B^q = \sum_{j=0}^{q} \boldsymbol{\eta}_j B^j$, with $\boldsymbol{\eta}_0 = \mathbf{0}$, the lag polynomials for the endogenous and exogenous vector variables respectively, where B is the backshift operator: $B\mathbf{y}_t = \mathbf{y}_{t-1}$.

Assume that the exogenous vector variable has VMA(∞) model representation: $\mathbf{s}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\delta}_{t-j}$, with $\boldsymbol{\Psi}_j$ $(m' \times m')$ real matrices, $\{\boldsymbol{\delta}_t\} = \left\{[\delta_{1,t}, \dots, \delta_{m',t}]^\top\right\}$ White Noise random vectors, i.e. $E[\boldsymbol{\delta}_t] = \mathbf{0}$, and $E\left[\boldsymbol{\delta}_t \boldsymbol{\delta}_t^\top\right] = \boldsymbol{\Sigma}_{m'}$ its $(m' \times m')$ nonsingular variance-covariance matrix independent of time t. Further, let $\{\boldsymbol{\epsilon}_t\}$ be independent of $\{\boldsymbol{\delta}_t\}$.

Remark 18. *In order to obtain the vector moving-average of infinite order representation, VMA(∞), of a stationary vector autoregressive of order p. For simplicity, let $\|\mathbf{A}\|_F^2 \equiv \operatorname{tr}\left\{\mathbf{A}^\top \mathbf{A}\right\}$ be the Frobenious matrix norm. Since $\|\mathbf{A}\| \leq \|\mathbf{A}\|_F$, where $\|\mathbf{A}\|$ denotes the spectral norm as before, this does not impact any of our results. Consider a purely nondeterministic weakly stationary process $\{\mathbf{y}_t\} \in \mathbb{R}^w$, with costant mean $E[\mathbf{y}_t] = \boldsymbol{\mu}$. Then, it follows that $\{\mathbf{y}_t - \boldsymbol{\mu}\}$ is the output of a causal linear filter with white noise input $\{\boldsymbol{\varepsilon}_t\}$, hence, delivering our VMA(∞) representation:*

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Phi}_j \boldsymbol{\varepsilon}_{t-j}, \tag{301}$$

with $\boldsymbol{\Phi}_0 = \boldsymbol{I}_w$, where $\boldsymbol{\Phi}_j$ are such that:

$$\sum_{j=0}^{\infty} \|\boldsymbol{\Phi}_j\|_F^2 < \infty. \tag{302}$$

Assumptions 13. *In order for Conditions (C3') and (C6') to hold, consider the following assumptions:*

(i) Let the DGP be a VARX(∞, q) defined by:

$$\mathbf{A}^*(B)\mathbf{y}_t = \boldsymbol{\eta}^*(B)\mathbf{s}_t + \boldsymbol{\epsilon}_t, \tag{303}$$

where $\mathbf{A}^{-1}(z) \equiv \boldsymbol{\theta}(z) = \sum_{j=0}^{\infty} \boldsymbol{\theta}_j z^j$ *and* $\mathbf{C}_j \equiv \sum_{k=0}^{j} [\boldsymbol{\Psi}_k \otimes \boldsymbol{\theta}_{j-k}]$ *are:*

$$\sum_{j=0}^{\infty} \|\boldsymbol{\theta}_j\|_F^2 < \infty, \tag{304}$$

$$\sum_{j=0}^{\infty} \|\mathbf{C}_j\|_F^2 < \infty. \tag{305}$$

(ii) There exist two constants $c_1 > 0$, *and* $r > 3/4$ *such that:*

$$\|\boldsymbol{\theta}_j\|_F \leq c_1 (j+1)^{-r}, \tag{306}$$

$$\|\boldsymbol{\Psi}_j\|_F + \|\mathbf{C}_j\|_F \leq c_1 (j+1)^{-r}. \tag{307}$$

(iii) The fourth moments of

$$\mathbf{v}_t \equiv \left[\epsilon_{1,t} \dots \epsilon_{w,t}\ \delta_{1,t}, \dots, \delta_{m',t}\right]^\top, \tag{308}$$

a $(m' \times 1)$*-dimensional error vector, with* $m = w + m'$ *and*

$$E\left[\mathbf{v}_t \mathbf{v}_t^\top\right] = E\begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\delta}} \end{bmatrix} = \boldsymbol{\Lambda}, \tag{309}$$

are independent of t, *and that:*

$$\sup_{-\infty<t<\infty} E\left[\|\boldsymbol{v}\|^\theta\right] < +\infty,\ \theta > 10. \tag{310}$$

Theorem 7. *If Assumptions 13 hold, then Conditions (C3') and (C6') follow.*

Remark 19. *Assumptions 13 allow for the VMA(∞) representation of the involved processes, i.e. to write each in its Wold representation version. See Subsection 5.5.4 for further details.*

5.5 PROOFS

In this section we detail the proofs of the three theorems for the multiple regressor case, and one of the VARX(∞, q) model example. As in the previous chapter, hereafter all the derivations hold for any fixed $h \geq 1$; for the sake of presentation we write ε_t instead of $\varepsilon_t^{(h)}$. Remember that $\{l_n\}$ indicates an increasing sequence of positive integers such that:

$$l_n \to \infty, \qquad \frac{l_n}{\sqrt{n}} = o(1) \tag{311}$$

and define $a = n - l_n - h$ and $b = n - l_n - h + 1$.

5.5.1 *Proof of Theorem 4*

We follow the same structure as in the previous chapter. Hereafter all the derivations hold for any fixed $h \geq 1$; for the sake of presentation we will write $\boldsymbol{\varepsilon}_t$ instead of $\boldsymbol{\varepsilon}_t^{(h)}$. Remember that $\{l_n\}$ indicates an increasing sequence of positive integers such that:

$$l_n \to \infty, \qquad \frac{l_n}{\sqrt{n}} = o\left(1\right) \tag{312}$$

and define $a = n - l_n - h$ and $b = n - l_n - h + 1$.

Proposition 7. *Under assumptions of Theorem 4, it holds that*

$$\text{(I)} = \text{(III)} + o(1), \tag{313}$$

with

$$\begin{aligned}
\text{(I)} &= -\left\{E\left[\hat{\boldsymbol{\Sigma}}^\top \hat{\mathbf{R}}^{-1} \boldsymbol{A}\right] + E\left[\boldsymbol{A}^\top \hat{\mathbf{R}}^{-1} \hat{\boldsymbol{\Sigma}}\right]\right\}, \\
\text{(III)} &= -\left\{E\left[\hat{\boldsymbol{\Sigma}}^\top \mathbf{R}^{-1} \boldsymbol{A}\right] + E\left[\boldsymbol{A}^\top \mathbf{R}^{-1} \hat{\boldsymbol{\Sigma}}\right]\right\},
\end{aligned}$$

where $\boldsymbol{A} = \mathbf{x}_n \boldsymbol{\varepsilon}_n^\top$ *and* $\hat{\boldsymbol{\Sigma}} = \sum_{t=1}^N \mathbf{x}_t \boldsymbol{\varepsilon}_t^\top$.

Proof. We need to show that:

$$\|\text{(I)} - \text{(III)}\| = o(1). \tag{314}$$

Note that the left hand side is equal to:

$$\left\|\left\{E\left[\hat{\boldsymbol{\Sigma}}^\top \left(\hat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\right) \boldsymbol{A}\right] + E\left[\boldsymbol{A}^\top \left(\hat{\mathbf{R}}^{-1} - \boldsymbol{R}^{-1}\right) \hat{\boldsymbol{\Sigma}}\right]\right\}\right\| \tag{315}$$

By using standard properties of the norm, we show that Eq. 313 follows after noticing that:

$$E\left[\hat{\boldsymbol{\Sigma}}^\top \left(\hat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\right) \boldsymbol{A}\right] = o(1). \tag{316}$$

Let

$$\tilde{\mathbf{R}}^{-1} = (n - l_n)^{-1} \sum_{t=1}^{n-l_n} \mathbf{x}_t \mathbf{x}_t^\top \tag{317}$$

Now, add and substract $\hat{\mathbf{\Sigma}}^{\top}\tilde{\mathbf{R}}^{-1}\mathbf{x}_n\varepsilon_n^{\top}$ from the left hand side of Eq. 316 to obtain:

$$
\begin{aligned}
& E\left[\hat{\mathbf{\Sigma}}^{\top}\left(\hat{\mathbf{R}}^{-1}-\mathbf{R}^{-1}\right)\boldsymbol{A}\right] \\
&= E\left[\sum_{t=1}^{N}\left(\mathbf{x}_t\varepsilon_t^{\top}\right)^{\top}\left(\hat{\mathbf{R}}^{-1}-\tilde{\mathbf{R}}^{-1}\right)\mathbf{x}_n\varepsilon_n^{\top}\right] \\
&\quad + E\left[\sum_{t=1}^{N}\left(\mathbf{x}_t\varepsilon_t^{\top}\right)^{\top}\left(\tilde{\mathbf{R}}^{-1}-\mathbf{R}^{-1}\right)\mathbf{x}_n\varepsilon_n^{\top}\right] \\
&= E\left[\sum_{t=1}^{N}\left(\boldsymbol{x}_t\varepsilon_t^{\top}\right)^{\top}\left(\hat{\boldsymbol{R}}^{-1}-\tilde{\boldsymbol{R}}^{-1}\right)\mathbf{x}_n\varepsilon_n^{\top}\right] && (318) \\
&\quad + E\left[\sum_{t=b}^{N}\left(\boldsymbol{x}_t\varepsilon_t^{\top}\right)^{\top}\left(\tilde{\boldsymbol{R}}^{-1}-\boldsymbol{R}^{-1}\right)\mathbf{x}_n\varepsilon_n^{\top}\right] && (319) \\
&\quad + E\left[\sum_{t=1}^{a}\left(\boldsymbol{x}_t\varepsilon_t^{\top}\right)^{\top}\left(\tilde{\boldsymbol{R}}^{-1}-\boldsymbol{R}^{-1}\right)\mathbf{x}_n\varepsilon_n^{\top}\right] && (320)
\end{aligned}
$$

We proceed to show that the norms of (318), (319), (320) vanish asymptotically. Let us consider the first one. By combining conditions (C3'), (C4'), Lemma 1, and Hölder's Inequality, it follows that $\|(318)\|$ is bounded by

$$
\begin{aligned}
& E\left[\left\|\sum_{t=1}^{N}\left(\boldsymbol{x}_t\varepsilon_t^{\top}\right)^{\top}\left(\hat{\boldsymbol{R}}^{-1}-\tilde{\boldsymbol{R}}^{-1}\right)\boldsymbol{x}_n\varepsilon_n^{\top}\right\|\right] \\
& \leq E\left[\left\|\left(\hat{\boldsymbol{R}}^{-1}-\tilde{\boldsymbol{R}}^{-1}\right)\right\|^3\right]^{\frac{1}{3}} E\left[\|\boldsymbol{x}_n\|^6\right]^{\frac{1}{6}} E\left[\|\varepsilon_n\|^6\right]^{\frac{1}{6}} \\
& \times E\left[\left\|N^{\frac{1}{2}}N^{-\frac{1}{2}}\sum_{t=1}^{N}\left(\boldsymbol{x}_t\varepsilon_t^{\top}\right)^{\top}\right\|^3\right]^{\frac{1}{3}} = O\left(\frac{l_n}{n^{\frac{1}{2}}}\right)
\end{aligned}
$$

which vanishes asymptotically due to the definition of l_n in (312). In the same manner, we have that $\|(319)\|$ is bounded by

$$
\begin{aligned}
& E\left[\|\varepsilon_n\|^6\right]^{\frac{1}{6}} E\left[\|\boldsymbol{x}_n\|^6\right]^{\frac{1}{6}} E\left[\left\|\left(\tilde{\boldsymbol{R}}^{-1}-\boldsymbol{R}^{-1}\right)\right\|^3\right]^{\frac{1}{3}} \\
& \times E\left[\left\|(N-b+1)^{\frac{1}{2}}(N-b+1)^{-\frac{1}{2}}\sum_{t=b}^{N}\left(\boldsymbol{x}_t\varepsilon_t^{\top}\right)^{\top}\right\|^3\right]^{\frac{1}{3}}
\end{aligned}
$$

which is an $O(n^{-1/2}l_n)$ thus vanishing in the limit. Finally, conditions (C4'), (C6'), Lemma 1, and Hölder's Inequality imply that $\|(320)\|$ is bounded by

$$
\begin{aligned}
& E\left[\left\|E\left[\boldsymbol{x}_t\varepsilon_t^{T}|\mathcal{F}_{t-l_n}\right]\right\|^3\right]^{\frac{1}{3}} E\left[\left\|\tilde{\boldsymbol{R}}^{-1}-\boldsymbol{R}^{-1}\right\|^3\right]^{\frac{1}{3}} \\
& \times E\left[\left\|(a)^{\frac{1}{2}}(a)^{-\frac{1}{2}}\sum_{t=1}^{a}\left(\boldsymbol{x}_t\varepsilon_t^{T}\right)^{T}\right\|^3\right]^{\frac{1}{3}} = o(1)
\end{aligned}
$$

completing the proof. □

Proposition 8. *Under assumptions of Theorem 4, it holds that:*

$$(\mathrm{II}) = (\mathrm{IV}) + o(1) \tag{321}$$

where

$$\begin{aligned}(\mathrm{II}) &= E\left[\hat{\mathbf{\Sigma}}^T \hat{\mathbf{R}}^{-1} \mathbf{x}_n \mathbf{x}_n^T \hat{\mathbf{R}}^{-1} \hat{\mathbf{\Sigma}}\right] \\ (\mathrm{IV}) &= E\left[\hat{\mathbf{\Sigma}}_B^T \mathbf{R}^{-1} \hat{\mathbf{\Sigma}}_B\right]\end{aligned}$$

with $\hat{\mathbf{\Sigma}}$ *defined in Proposition 7 and* $\hat{\mathbf{\Sigma}}_B = N^{-\frac{1}{2}} \sum_{t=1}^{N} \left(\mathbf{x}_t \boldsymbol{\varepsilon}_t^T\right)$.

Proof. Let $\boldsymbol{M}_1 = \hat{\mathbf{\Sigma}}_B^T \left(\hat{\boldsymbol{R}}^{-1} - \boldsymbol{R}^{-1}\right) \boldsymbol{x}_n$ and $\boldsymbol{M}_2 = \hat{\mathbf{\Sigma}}_B^T \boldsymbol{R}^{-1} \boldsymbol{x}_n$. Since

$$\begin{aligned}N\,(\mathrm{II}) &= E\left[\left(\boldsymbol{M}_1 + \boldsymbol{M}_2\right)\left(\boldsymbol{M}_1 + \boldsymbol{M}_2\right)^\top\right] \\ &= E\left[\boldsymbol{M}_1 \boldsymbol{M}_1^\top\right] + E\left[\boldsymbol{M}_2 \boldsymbol{M}_2^\top\right] + E\left[\boldsymbol{M}_1 \boldsymbol{M}_2^\top\right] + E\left[\boldsymbol{M}_2 \boldsymbol{M}_1^\top\right]\end{aligned}$$

the proof of (321) reduces to show that the following conditions hold:

$$\left\| E\left[\boldsymbol{M}_1 \boldsymbol{M}_1^\top\right] \right\| = o\left(1\right), \tag{322}$$

$$\left\| E\left[\boldsymbol{M}_1 \boldsymbol{M}_2^\top\right] \right\| = o\left(1\right), \tag{323}$$

$$\left\| E\left[\boldsymbol{M}_2 \boldsymbol{M}_2^\top\right] - (\mathrm{IV}) \right\| = o\left(1\right). \tag{324}$$

Conditions (322) and (323) readily follow from Assumptions (C3′) and (C4′), Lemma 1, the non singularity of R and Hölder's Inequality:

$$\begin{aligned}E\left[\left\|\boldsymbol{M}_1 \boldsymbol{M}_1^T\right\|\right] &= E\left[\left\|\hat{\mathbf{\Sigma}}_B^T \left(\hat{\boldsymbol{R}}^{-1} - \boldsymbol{R}^{-1}\right) \boldsymbol{x}_n \boldsymbol{x}_n^T \left(\hat{\boldsymbol{R}}^{-1} - \boldsymbol{R}^{-1}\right) \hat{\mathbf{\Sigma}}_B\right\|\right] \\ &\leq \left(E\left[\left\|\boldsymbol{x}_n\right\|^{10}\right]\right)^{\frac{1}{5}} \left(E\left[\left\|\hat{\boldsymbol{R}}^{-1} - \boldsymbol{R}^{-1}\right\|^5\right]\right)^{\frac{2}{5}} \\ &\quad \times \left(E\left[\left\|\hat{\mathbf{\Sigma}}_B\right\|^5\right]\right)^{\frac{2}{5}} = o(1) \\ E\left[\left\|\boldsymbol{M}_1 \boldsymbol{M}_2^T\right\|\right] &= E\left[\left\|\hat{\mathbf{\Sigma}}_B^T \left(\hat{\boldsymbol{R}}^{-1} - \boldsymbol{R}^{-1}\right) \boldsymbol{x}_n \boldsymbol{x}_n^T \boldsymbol{R}^{-1} \hat{\mathbf{\Sigma}}_B\right\|\right] \\ &\leq \left(E\left[\left\|\boldsymbol{x}_n\right\|^{10}\right]\right)^{\frac{1}{5}} \left(E\left[\left\|\left(\hat{\boldsymbol{R}}^{-1} - \boldsymbol{R}^{-1}\right)\right\|^5\right]\right)^{\frac{1}{5}} \\ &\quad \times \left(E\left[\left\|\boldsymbol{R}^{-1}\right\|^5\right]\right)^{\frac{1}{5}} \left(E\left[\left\|\hat{\mathbf{\Sigma}}_B\right\|^5\right]\right)^{\frac{2}{5}} = o(1)\end{aligned}$$

In relation to Eq. (324), partition matrix $\hat{\mathbf{\Sigma}}_B$ as:

$$\hat{\mathbf{\Sigma}}_B = N^{-\frac{1}{2}} \sum_{t=1}^{N} \left(\boldsymbol{x}_t \boldsymbol{\varepsilon}_t^T\right) = \boldsymbol{U} + \boldsymbol{W}$$

with $\boldsymbol{U} = N^{-\frac{1}{2}} \sum_{t=1}^{a} \left(\boldsymbol{x}_t \boldsymbol{\varepsilon}_t^T \right)$ and $\boldsymbol{W} = N^{-\frac{1}{2}} \sum_{t=b}^{N} \left(\boldsymbol{x}_t \boldsymbol{\varepsilon}_t^T \right)$. Therefore, we have that $E\left[\boldsymbol{M}_2 \boldsymbol{M}_2^\top\right] - (\mathrm{IV})$ is equal to:

$$
\begin{aligned}
& E\left[\boldsymbol{U}^T \boldsymbol{R}^{-1} \boldsymbol{x}_n \boldsymbol{x}_n^T \boldsymbol{R}^{-1} \boldsymbol{U}\right] - E\left[\boldsymbol{U}^T \boldsymbol{R}^{-1} \boldsymbol{R} \boldsymbol{R}^{-1} \boldsymbol{U}\right] \\
& \quad + E\left[\boldsymbol{W}^T \boldsymbol{R}^{-1} \boldsymbol{x}_n \boldsymbol{x}_n^T \boldsymbol{R}^{-1} \boldsymbol{U}\right] - E\left[\boldsymbol{W}^T \boldsymbol{R}^{-1} \boldsymbol{R} \boldsymbol{R}^{-1} \boldsymbol{U}\right] \\
& \quad + E\left[\boldsymbol{U}^T \boldsymbol{R}^{-1} \boldsymbol{x}_n \boldsymbol{x}_n^T \boldsymbol{R}^{-1} \boldsymbol{W}\right] - E\left[\boldsymbol{U}^T \boldsymbol{R}^{-1} \boldsymbol{R} \boldsymbol{R}^{-1} \boldsymbol{W}\right] \\
& \quad + E\left[\boldsymbol{W}^T \boldsymbol{R}^{-1} \boldsymbol{x}_n \boldsymbol{x}_n^T \boldsymbol{R}^{-1} \boldsymbol{W}\right] - E\left[\boldsymbol{W}^T \boldsymbol{R}^{-1} \boldsymbol{R} \boldsymbol{R}^{-1} \boldsymbol{W}\right]
\end{aligned}
$$

The law of iterated expectations implies that:

$$
\begin{aligned}
& \left\| E\left[\boldsymbol{M}_2 \boldsymbol{M}_2^\top\right] - (\mathrm{IV}) \right\| \\
& \leq \left\| E\left[\boldsymbol{U}^T \boldsymbol{R}^{-1} \left(E\left[\boldsymbol{x}_n \boldsymbol{x}_n^T | \mathcal{F}_{n-l_n}\right] - \boldsymbol{R} \right) \boldsymbol{R}^{-1} \boldsymbol{U}\right] \right\| && (325) \\
& + \left\| E\left[\boldsymbol{U}^T \boldsymbol{R}^{-1} \left(E\left[\boldsymbol{x}_n \boldsymbol{x}_n^T | \mathcal{F}_{n-l_n}\right] - \boldsymbol{R} \right) \boldsymbol{R}^{-1} \boldsymbol{W}\right] \right\| && (326) \\
& + \left\| E\left[\boldsymbol{W}^T \boldsymbol{R}^{-1} \left(E\left[\boldsymbol{x}_n \boldsymbol{x}_n^T | \mathcal{F}_{n-l_n}\right] - \boldsymbol{R} \right) \boldsymbol{R}^{-1} \boldsymbol{U}\right] \right\| && (327) \\
& + \left\| E\left[\boldsymbol{W}^T \boldsymbol{R}^{-1} \left(E\left[\boldsymbol{x}_n \boldsymbol{x}_n^T | \mathcal{F}_{n-l_n}\right] - \boldsymbol{R} \right) \boldsymbol{R}^{-1} \boldsymbol{W}\right] \right\| && (328)
\end{aligned}
$$

By using previously developed arguments, it is easy to see that, under Assumptions (C4′) and (C6′), (325) – (328) are negligible. Therefore, conditions (322) – (324) hold completing the proof. □

Proposition 9. *Under assumptions of Theorem 4, it holds that:*

$$
(\mathrm{III}) = -(D) + o(1), \tag{329}
$$

where

$$
(D) = E\left[\sum_{j=h}^{N-1} \left(\varepsilon_1 \boldsymbol{x}_1^T \boldsymbol{R}^{-1} \boldsymbol{x}_{1+j} \boldsymbol{\varepsilon}_{1+j}^T + \varepsilon_{1+j} \boldsymbol{x}_{1+j}^T \boldsymbol{R}^{-1} \boldsymbol{x}_1 \boldsymbol{\varepsilon}_1^T \right) \right]
$$

Proof. The result readily follows upon noting that, under Assumption (C2′) and the weakly stationarity of the process $\{x_t\}$, it holds that (III) is equal to:

$$
\begin{aligned}
& = -\sum_{t=1}^{N} E\left[\left(\boldsymbol{x}_t \boldsymbol{\varepsilon}_t^T \right)^T \boldsymbol{R}^{-1} \boldsymbol{x}_n \boldsymbol{\varepsilon}_n^T + \varepsilon_n \boldsymbol{x}_n^T \boldsymbol{R}^{-1} \boldsymbol{x}_t \boldsymbol{\varepsilon}_t^T \right] \\
& = -\sum_{j=h}^{n-1} E\left[\left(\boldsymbol{x}_1 \boldsymbol{\varepsilon}_1^T \right)^T \boldsymbol{R}^{-1} \left(\boldsymbol{x}_{1+j} \boldsymbol{\varepsilon}_{1+j}^T \right) + \left(\boldsymbol{x}_{1+j} \boldsymbol{\varepsilon}_{1+j}^T \right)^T \boldsymbol{R}^{-1} \left(\boldsymbol{x}_1 \boldsymbol{\varepsilon}_1^T \right) \right] \\
& = -E\left[\sum_{j=h}^{N-1} \left(\varepsilon_1 \boldsymbol{x}_1^T \boldsymbol{R}^{-1} \boldsymbol{x}_{1+j} \boldsymbol{\varepsilon}_{1+j}^T + \varepsilon_{1+j} \boldsymbol{x}_{1+j}^T \boldsymbol{R}^{-1} \boldsymbol{x}_1 \boldsymbol{\varepsilon}_1^T \right) \right] + o(1).
\end{aligned}
$$

□

Proposition 10. *Under assumptions of Theorem 4, it holds that:*

$$(\text{IV}) = (1) + (Q) + (D) + o(1), \tag{330}$$

where

$$(1) = N^{-1}E\left[\sum_{t=1}^{N}\left(\boldsymbol{x}_t\boldsymbol{\varepsilon}_t^T\right)^T\boldsymbol{R}^{-1}\left(\boldsymbol{x}_t\boldsymbol{\varepsilon}_t^T\right)\right],$$

$$(Q) = E\left[\sum_{s=1}^{h-1}\left(\boldsymbol{\varepsilon}_1\boldsymbol{x}_1^T\boldsymbol{R}^{-1}\boldsymbol{x}_{1+s}\boldsymbol{\varepsilon}_{1+s}^T + \boldsymbol{\varepsilon}_{1+s}\boldsymbol{x}_{1+s}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1\boldsymbol{\varepsilon}_1^T\right)\right],$$

$$(D) = E\left[\sum_{j=h}^{N-1}\left(\boldsymbol{\varepsilon}_1\boldsymbol{x}_1^T\boldsymbol{R}^{-1}\boldsymbol{x}_{1+j}\boldsymbol{\varepsilon}_{1+j}^T + \boldsymbol{\varepsilon}_{1+j}\boldsymbol{x}_{1+j}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1\boldsymbol{\varepsilon}_1^T\right)\right].$$

Proof. Let

$$(2) = N^{-1}E\left[\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\left(\boldsymbol{x}_j\boldsymbol{\varepsilon}_j^T\right)^T\boldsymbol{R}^{-1}\left(\boldsymbol{x}_k\boldsymbol{\varepsilon}_k^T\right)\right],$$

and note that $(\text{IV}) - (1) = (2) + (2)^\top$. Moreover

$$\begin{aligned}(2) &= N^{-1}E\left[\sum_{j=1}^{N-1}(N-j)\left(\boldsymbol{x}_1\boldsymbol{\varepsilon}_1^T\right)^T\boldsymbol{R}^{-1}\left(\boldsymbol{x}_{1+j}\boldsymbol{\varepsilon}_{1+j}^T\right)\right] \\ &= E\left[\sum_{j=1}^{N-1}\left(\boldsymbol{x}_1\boldsymbol{\varepsilon}_1^T\right)^T\boldsymbol{R}^{-1}\left(\boldsymbol{x}_{1+j}\boldsymbol{\varepsilon}_{1+j}^T\right)\right] \qquad (331) \\ &\quad - N^{-1}E\left[\sum_{j=1}^{N-1}j\left(\boldsymbol{x}_1\boldsymbol{\varepsilon}_1^T\right)^T\boldsymbol{R}^{-1}\left(\boldsymbol{x}_{1+j}\boldsymbol{\varepsilon}_{1+j}^T\right)\right]. \qquad (332)\end{aligned}$$

Assumptions (C2′) implies that (332) is $o(1)$. Since (331) can be written as

$$\begin{aligned}&E\left[\sum_{s=1}^{h-1}\left(\boldsymbol{\varepsilon}_1\boldsymbol{x}_1^T\boldsymbol{R}^{-1}\boldsymbol{x}_{1+s}\boldsymbol{\varepsilon}_{1+s}^T + \boldsymbol{\varepsilon}_{1+s}\boldsymbol{x}_{1+s}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1\boldsymbol{\varepsilon}_1^T\right)\right] \\ +E&\left[\sum_{j=h}^{N-1}\left(\boldsymbol{\varepsilon}_1\boldsymbol{x}_1^T\boldsymbol{R}^{-1}\boldsymbol{x}_{1+j}\boldsymbol{\varepsilon}_{1+j}^T + \boldsymbol{\varepsilon}_{1+j}\boldsymbol{x}_{1+j}^T\boldsymbol{R}^{-1}\boldsymbol{x}_1\boldsymbol{\varepsilon}_1^T\right)\right],\end{aligned}$$

then $(331) + (331)^\top = (Q) + (D)$ and this completes the proof. □

5.5.2 *Proof of Theorem 5*

We start proving that

$$\hat{\mathbf{MI}}_h = \mathbf{MI}_h + O_p(n^{-1/2}). \tag{333}$$

Note that

$$\hat{\mathbf{MI}}_h = N^{-1}\left(\sum_{t=1}^{N}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top\right) - \left(N^{-1}\sum_{t=1}^{N}\mathbf{x}_t\boldsymbol{\varepsilon}_t^\top\right)\hat{\mathbf{R}}^{-1}\left(N^{-1}\sum_{s=1}^{N}\mathbf{x}_s\boldsymbol{\varepsilon}_s^\top\right)^\top$$

hence, it holds that $\hat{\mathbf{MI}}_h - \mathbf{MI}_h$ equals

$$N^{-1}\left\{\sum_{t=1}^{N}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top - \mathrm{E}\left[\boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_1^\top\right]\right)\right\} \tag{334}$$

$$-\left(N^{-1}\sum_{t=1}^{N}\mathbf{x}_t\boldsymbol{\varepsilon}_t^\top\right)\hat{R}^{-1}\left(N^{-1}\sum_{t=1}^{N}\mathbf{x}_t\boldsymbol{\varepsilon}_t^\top\right)^\top. \tag{335}$$

Assumption (A1′) implies that $(334) = O_p(n^{-1/2})$ whereas, by combining Assumptions (A3′) and (A4′) with the non-singularity of $\mathbf{R}$ and Hölder's inequality, it can be shown that $(335) = O_p(n^{-1})$ and hence the proof of (333) is complete.
Next, we prove that

$$\hat{\mathbf{VI}}_h = \mathbf{VI}_h + o_p(1). \tag{336}$$

Focus on vector $\mathrm{vec}\left[\mathbf{VI}_h\right]$ and its MoME:

$$\mathrm{vec}\left[\mathbf{VI}_h\right] = \left[\mathbf{A}_{h,0} + \sum_{s=1}^{h-1}\left\{\mathbf{H}_{h,s} + \mathbf{C}_{h,s}\right\}\right]\mathrm{vec}\left(\mathbf{R}^{-1}\right), \tag{337}$$

$$\mathrm{vec}\left[\hat{\mathbf{VI}}_h\right] = \left[\hat{\mathbf{A}}_{h,0} + \sum_{s=1}^{h-1}\left\{\hat{\mathbf{H}}_{h,s} + \hat{\mathbf{C}}_{h,s}\right\}\right]\mathrm{vec}\left(\hat{\mathbf{R}}^{-1}\right), \tag{338}$$

where

$$\mathbf{A}_{h,0} = E\left[\left(\mathbf{B}_t^{(h)}\otimes\mathbf{B}_t^{(h)}\right)^\top\right], \tag{339}$$

$$\mathbf{H}_{h,s} = E\left[\left(\mathbf{B}_{t+s}^{(h)}\otimes\mathbf{B}_t^{(h)}\right)^\top\right], \tag{340}$$

$$\mathbf{C}_{h,s} = E\left[\left(\mathbf{B}_t^{(h)}\otimes\mathbf{B}_{t+s}^{(h)}\right)^\top\right], \tag{341}$$

$$\hat{\mathbf{A}}_{h,0} = N^{-1}\sum_{t=1}^{N}\left[\hat{\mathbf{B}}_t^{(h)}\otimes\hat{\mathbf{B}}_t^{(h)}\right]^\top, \tag{342}$$

$$\hat{\mathbf{H}}_{h,s} = (N-s)^{-1}\sum_{t=1}^{N-s}\left[\hat{\mathbf{B}}_{t+s}^{(h)}\otimes\hat{\mathbf{B}}_t^{(h)}\right]^\top, \tag{343}$$

$$\hat{\mathbf{C}}_{h,s} = (N-s)^{-1}\sum_{t=1}^{N-s}\left[\hat{\mathbf{B}}_t^{(h)}\otimes\hat{\mathbf{B}}_{t+s}^{(h)}\right]^\top, \tag{344}$$

with $\hat{\mathbf{B}}_t^{(h)}$ defined as in Eq. (294). It is easy to see for the estimated matrices $\hat{\mathbf{A}}_{h,0}, \hat{\mathbf{H}}_{h,s}$ and $\hat{\mathbf{C}}_{h,s}$ that:

$$\hat{\mathbf{B}}_t^{(h)}\otimes\hat{\mathbf{B}}_t^{(h)} = \left[\mathbf{B}_t^{(h)}\otimes\mathbf{B}_t^{(h)}\right] + \boldsymbol{\alpha}_1^{(1)} + \boldsymbol{\alpha}_2^{(1)} + \boldsymbol{\alpha}_3^{(1)}, \tag{345}$$

$$\hat{\mathbf{B}}_{t+s}^{(h)}\otimes\hat{\mathbf{B}}_t^{(h)} = \left[\mathbf{B}_{t+s}^{(h)}\otimes\mathbf{B}_t^{(h)}\right] + \boldsymbol{\alpha}_1^{(2)} + \boldsymbol{\alpha}_2^{(2)} + \boldsymbol{\alpha}_3^{(2)}, \tag{346}$$

$$\hat{\mathbf{B}}_t^{(h)}\otimes\hat{\mathbf{B}}_{t+s}^{(h)} = \left[\mathbf{B}_t^{(h)}\otimes\mathbf{B}_{t+s}^{(h)}\right] + \boldsymbol{\alpha}_1^{(3)} + \boldsymbol{\alpha}_2^{(3)} + \boldsymbol{\alpha}_3^{(3)}, \tag{347}$$

where for $\hat{\mathbf{A}}_{h,0}$ the $\boldsymbol{\alpha}$ matrices are:

$$\begin{aligned}
\boldsymbol{\alpha}_1^{(1)} &= \left[\mathbf{1}_t^{(h)} \otimes \mathbf{1}_t^{(h)}\right], \\
\boldsymbol{\alpha}_2^{(1)} &= -\left[\mathbf{1}_t^{(h)} \otimes \mathbf{B}_t^{(h)}\right], \\
\boldsymbol{\alpha}_3^{(1)} &= -\left[\mathbf{B}_t^{(h)} \otimes \mathbf{1}_t^{(h)}\right];
\end{aligned}$$

for $\hat{\mathbf{H}}_{h,s}$ these are:

$$\begin{aligned}
\boldsymbol{\alpha}_1^{(2)} &= \left[\mathbf{1}_{t+s}^{(h)} \otimes \mathbf{1}_t^{(h)}\right], \\
\boldsymbol{\alpha}_2^{(2)} &= -\left[\mathbf{1}_{t+s}^{(h)} \otimes \mathbf{B}_t^{(h)}\right], \\
\boldsymbol{\alpha}_3^{(2)} &= -\left[\mathbf{B}_{t+s}^{(h)} \otimes \mathbf{1}_t^{(h)}\right];
\end{aligned}$$

and for $\hat{\mathbf{C}}_{h,s}$:

$$\begin{aligned}
\boldsymbol{\alpha}_1^{(3)} &= \left[\mathbf{1}_t^{(h)} \otimes \mathbf{1}_{t+s}^{(h)}\right], \\
\boldsymbol{\alpha}_2^{(3)} &= -\left[\mathbf{1}_t^{(h)} \otimes \mathbf{B}_{t+s}^{(h)}\right], \\
\boldsymbol{\alpha}_3^{(3)} &= -\left[\mathbf{B}_t^{(h)} \otimes \mathbf{1}_{t+s}^{(h)}\right];
\end{aligned}$$

with $\mathbf{1}_t^{(h)} = \mathbf{x}_t\mathbf{x}_t^\top\left(\hat{\boldsymbol{\beta}}_h - \boldsymbol{\beta}_h\right) - \mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)\top}$. Hence, we need to prove that:

$$N^{-1}\sum_{t=1}^{N}\left\{\boldsymbol{\alpha}_1^{(1)} + \boldsymbol{\alpha}_2^{(1)} + \boldsymbol{\alpha}_3^{(1)}\right\}^\top = o_p(1), \tag{348}$$

$$(N-s)^{-1}\sum_{t=1}^{N-s}\left\{\boldsymbol{\alpha}_1^{(i)} + \boldsymbol{\alpha}_2^{(i)} + \boldsymbol{\alpha}_3^{(i)}\right\}^\top = o_p(1), \tag{349}$$

for $i = \{2, 3\}$. By usual matrix norm properties, the mixed-product property for Kronecker product [149, Lemma 4.2.10], Hölder's inequality, Theorem 8 in Lancaster and Farahat [182, p. 412], and Conditions (A4') and (A5'), Eq. (348) vanishes in probability, since:

$$\begin{aligned}
&E\left[\left\|N^{-1}\sum_{t=1}^{N}\left\{\boldsymbol{\alpha}_1^{(1)}\right\}\right\|^\top\right] \\
&\leq \left(E\left[\left\|\left[\left(N^{-1/2}\sum_{j=1}^{N}\mathbf{B}_j^{(h)}\right)\otimes\left(N^{-1/2}\sum_{j=1}^{N}\mathbf{B}_j^{(h)}\right)\right]^\top\right\|^4\right]\right)^{1/4} \\
&\times\left(E\left[\left\|\left[\hat{\mathbf{R}}^{-1}\otimes\hat{\mathbf{R}}^{-1}\right]^\top\right\|^4\right]\right)^{1/4} \\
&\times\left(E\left[\left\|\left[\left(\mathbf{x}_t\mathbf{x}_t^\top\right)\otimes\left(\mathbf{x}_t\mathbf{x}_t^\top\right)\right]^\top\right\|^2\right]\right)^{1/2} \leq o_p(1),
\end{aligned}$$

$$E\left[\left\|N^{-1}\sum_{t=1}^{N}\left\{\boldsymbol{\alpha}_2^{(1)}\right\}\right\|^{\top}\right] \leq \left(E\left[\left\|\left[\left(N^{-1/2}\sum_{j=1}^{N}\mathbf{B}_j^{(h)}\right)\otimes(\boldsymbol{I}_w)\right]^{\top}\right\|^4\right]\right)^{1/4}$$
$$\times\left(E\left[\left\|\left[\hat{\mathbf{R}}^{-1}\otimes\mathbf{B}_t^{(h)}\right]^{\top}\right\|^2\right]\right)^{1/2}$$
$$\times\left(E\left[\left\|\left[\left(\mathbf{x}_t\mathbf{x}_t^{\top}\right)\otimes(\boldsymbol{I}_m)\right]^{\top}\right\|^4\right]\right)^{1/4}\leq o_p(1),$$

and given that $\boldsymbol{\alpha}_3^{(1)}$ shares the same asymptotic behaviour as $\boldsymbol{\alpha}_2^{(1)}$. Similarly, we obtain that Eq. (349) with $i=2$ vanishes in probability since:

$$E\left[\left\|(N-s)^{-1}\sum_{t=1}^{N-s}\left[\boldsymbol{\alpha}_1^{(2)}\right]^{\top}\right\|\right]$$
$$\leq\left(E\left[\left\|\left[\left(N^{-1/2}\sum_{j=1}^{N}\mathbf{B}_j^{(h)}\right)\otimes\left(N^{-1/2}\sum_{j=1}^{N}\mathbf{B}_j^{(h)}\right)\right]^{\top}\right\|^4\right]\right)^{1/4}$$
$$\times\left(E\left[\left\|\left[\hat{\mathbf{R}}^{-1}\otimes\hat{\mathbf{R}}^{-1}\right]^{\top}\right\|^4\right]\right)^{1/4}$$
$$\times\left(E\left[\left\|\left[\left(\mathbf{x}_{t+s}\mathbf{x}_{t+s}^{\top}\right)\otimes\left(\mathbf{x}_t\mathbf{x}_t^{\top}\right)\right]^{\top}\right\|^2\right]\right)^{1/2}\leq o_p(1),$$

$$E\left[\left\|(N-s)^{-1}\sum_{t=1}^{N-s}\left[\boldsymbol{\alpha}_2^{(2)}\right]^{\top}\right\|\right]$$
$$\leq\left(E\left[\left\|\left[\left(N^{-1/2}\sum_{j=1}^{N}\mathbf{B}_j^{(h)}\right)\otimes(\boldsymbol{I}_w)\right]^{\top}\right\|^4\right]\right)^{1/4}$$
$$\times\left(E\left[\left\|\left[\hat{\mathbf{R}}^{-1}\otimes\mathbf{B}_t^{(h)}\right]^{\top}\right\|^2\right]\right)^{1/2}$$
$$\times\left(E\left[\left\|\left[\left(\mathbf{x}_{t+s}\mathbf{x}_{t+s}^{\top}\right)\otimes(\boldsymbol{I}_m)\right]^{\top}\right\|^4\right]\right)^{1/4}\leq o_p(1),$$

and, again, given that $\boldsymbol{\alpha}_3^{(2)}$ shares the same asymptotic behaviour than $\boldsymbol{\alpha}_2^{(2)}$. The case $i=3$ follows since $\hat{\mathbf{H}}_{h,s}=\hat{\mathbf{C}}_{h,s}^{\top}$. These results plus Condition (A4′) complete the proof.

5.5.3 *Proof of Theorem 6*

The proof is the same as in Section 4.5.3. It is reported for completeness.

Proof. By Theorem 5 the VMRIC_h defined in (297) can be written as:

$$\mathrm{VMRIC}_h\left(\hat{\ell}_h\right) = \min_{1\leq\ell\leq K}\left\|\mathbf{MI}_h + O_p(n^{-1/2})\right\| + \min_{\ell\in M_1}\left\|\frac{\alpha_n}{n}\mathbf{VI}_h + o_p\left(\frac{\alpha_n}{n}\right)\right\|.$$

Therefore,

$$\lim_{n\to\infty}\mathrm{VMRIC}_h\left(\hat{\ell}_h\right) = \min_{1\leq\ell\leq K}\|\mathbf{MI}_h\| \tag{350}$$

and hence

$$\lim_{n\to+\infty}\Pr\left(\hat{\ell}_h\in M_1\right) = 1. \tag{351}$$

Now, consider two models ℓ_1 and ℓ_2 in the candidates set $J_{\ell_1}, J_{\ell_2}\in M_1$ such that $\mathrm{VI}_h(\ell_1)\neq\mathrm{VI}_h(\ell_2)$. We show that

$$\lim_{n\to\infty}\Pr\Big[\operatorname{sign}\{\mathrm{VMRIC}_h(\ell_1)-\mathrm{VMRIC}_h(\ell_2)\} = \operatorname{sign}\{\|\mathbf{VI}_h(\ell_1)\|-\|\mathbf{VI}_h(\ell_2)\|\}\Big] = 1. \tag{352}$$

By defining $\mathbf{MI}_h^*$ to be the minimum value of $\mathbf{MI}_h$ over the family of candidate models, we have:

$$\mathrm{VMRIC}_h(\ell_1) = \left\|\mathbf{MI}_h^* + O_p(n^{-1/2})\right\| + \left\|\frac{\alpha_n}{n}\mathbf{VI}_h(\ell_1) + o_p\left(\frac{\alpha_n}{n}\right)\right\|,$$
$$\mathrm{VMRIC}_h(\ell_2) = \left\|\mathbf{MI}_h^* + O_p(n^{-1/2})\right\| + \left\|\frac{\alpha_n}{n}\mathbf{VI}_h(\ell_2) + o_p\left(\frac{\alpha_n}{n}\right)\right\|.$$

Therefore, for sufficiently large n, it holds that:

$$\mathrm{VMRIC}_h(\ell_1)-\mathrm{VMRIC}_h(\ell_2) = \left\|\frac{\alpha_n}{n}\right\|\left(\|\mathbf{VI}_h(\ell_1)\|-\|\mathbf{VI}_h(\ell_2)\|\right).$$

Thus

$$\operatorname{sign}\{\mathrm{VMRIC}_h(\ell_1)-\mathrm{VMRIC}_h(\ell_2)\} = \operatorname{sign}\{\|\mathbf{VI}_h(\ell_1)\|-\|\mathbf{VI}_h(\ell_2)\|\},$$

and (352) is verified and implies that

$$\lim_{n\to\infty}\Pr\left(\hat{\ell}_h\in M_2\right) = 1. \tag{353}$$

This completes the proof. □

5.5.4 *Proof of Theorem 7*

5.5.4.1 *Proof of Condition (C3')*

Proof. Assumptions (i), (ii), (iii) ensure that the processes involved in a possibly-misspecified h-step ahead forecasting model have VMA(∞) representation:

$$\mathbf{y}_t = \sum_{j=0}^{\infty} \mathbf{W}_{j,y} \mathbf{v}_{t-j}, \tag{354}$$

$$\mathbf{s}_t = \sum_{j=0}^{\infty} \mathbf{W}_{j,s} \mathbf{v}_{t-j}, \tag{355}$$

$$\boldsymbol{\varepsilon}_t^{(h)} = \sum_{j=0}^{\infty} \mathbf{W}_{j,0} \mathbf{v}_{t+h-j}, \tag{356}$$

with $\mathbf{W}_{j,y}, \mathbf{W}_{s,y}, \mathbf{W}_{j,0}$ nonrandom matrices such that:

$$\|\mathbf{W}_{j,y}\| \leq c^* (j+1)^{-r}, \tag{357}$$

$$\|\mathbf{W}_{j,s}\| \leq c^* (j+1)^{-r}, \tag{358}$$

$$\|\mathbf{W}_{j,0}\| \leq c^* (j+1)^{-r}. \tag{359}$$

A consequence of Proposition 11, detailed below, is that:

$$\sum_{k=0}^{\infty} \mathbf{W}_{k,y} \boldsymbol{\Lambda} \mathbf{W}_{k+h+a_1,0}^{\top} = \mathbf{0}, \; a_1 = \{0, 1, \dots\}, \tag{360}$$

$$\sum_{k=0}^{\infty} \mathbf{W}_{k,s} \boldsymbol{\Lambda} \mathbf{W}_{k+h+a_2,0}^{\top} = \mathbf{0}, \; a_2 = \{1, 2, \dots q\}. \tag{361}$$

Thus, since:

$$E\left[\left(\|\mathbf{x}_t\|^2\right)^5\right] = E\left[\left(\sum_{j=1}^{p} \|\mathbf{y}_{t+h-j}\|^2 + \sum_{k=1}^{q} \|\mathbf{s}_{t+h-k}\|^2\right)^5\right], \tag{362}$$

then it suffices to show that:

$$\sup_{-\infty<t<\infty} E\left[\|\mathbf{y}_t\|^{10}\right] + \sup_{-\infty<t<\infty} E\left[\|\mathbf{s}_t\|^{10}\right] = O(1), \tag{363}$$

which holds given Eq. (357), (358), (359), and Assumption (iii). Since same path can be employed to show the second part of (C3'), the proof is hence completed. □

Proposition 11. *Under Assumptions 13 we have:*

$$E\left[\mathbf{y}_{t-a_1} \boldsymbol{\varepsilon}_t^{(h)\top}\right] = \mathbf{0}, \; a_1 = \{0, 1, \dots\}, \; \textit{and} \tag{364}$$

$$E\left[\mathbf{s}_{t-a_2} \boldsymbol{\varepsilon}_t^{(h)\top}\right] = \mathbf{0}, \; a_2 = \{1, 2, \dots, q\}. \tag{365}$$

Proof. To proof Eq. (364) is the same as showing:

$$E\left[\mathbf{y}_{t-a_1}\mathbf{y}_{t+h}^{\top}\right] = E\left[\mathbf{y}_{t-a_1}\mathbf{x}_t^{\top}\right]\mathbf{R}^{-1}E\left[\mathbf{x}_t\mathbf{y}_{t+h}^{\top}\right]. \tag{366}$$

After basic algebraic manipulations, this reduces to show that:

$$E\left[\mathbf{x}_{t-h}\boldsymbol{\varepsilon}_t^{(h)\top}\right] = E\left[\mathbf{x}_{t-h}\mathbf{x}_t^{\top}\right]\mathbf{R}^{-1}E\left[\mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)\top}\right], \tag{367}$$

$$E\left[\boldsymbol{\varepsilon}_{t-h}^{(h)}\boldsymbol{\varepsilon}_{t-h}^{(h)\top}\right] = E\left[\boldsymbol{\varepsilon}_{t-h}^{(h)}\mathbf{x}_t^{\top}\right]\mathbf{R}^{-1}E\left[\mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)\top}\right]. \tag{368}$$

Since *(a)* $\mathbf{x}_t^{\top}\mathbf{R}^{-1}E\left[\mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)\top}\right]$ is the best linear predictor in the projector's sense and the conditional expectation function is the best linear predictor, i.e.

$$\mathcal{P}\left(\boldsymbol{\varepsilon}_t^{(h)\top}\middle|\mathbf{x}_t\right) = \mathbf{x}_t^{\top}\mathbf{R}^{-1}E\left[\mathbf{x}_t\boldsymbol{\varepsilon}_t^{(h)\top}\right] = E\left[\boldsymbol{\varepsilon}_t^{(h)\top}\middle|\mathbf{x}_t\right]; \tag{369}$$

and *(b)* given that $\sigma\left(\mathbf{x}_{t-h}\right) \subseteq \sigma\left(\mathbf{x}_t\right)$; then by the law of iterated expectation, we get that Eq. (367) hold. Eq. (368) and Eq. (365) follow identically. For details, see Hansen [139, Ch. 2-3]. □

5.5.4.2 *Proof of Condition (C6')*

Proof. For the first part, process $\{\mathbf{x}_t\}$ has to be adapted to the filtration $\sigma\left(\mathbf{x}_t\right) \subset \mathcal{F}_t \; \forall \, t$. It holds since in our case $\mathcal{F}_t = \sigma\left(\mathbf{v}_t, \mathbf{v}_{t-1}, \dots\right)$.

Then, the argument inside the matrix norm of

$$E\left[\left\|E\left[\mathbf{x}_t\mathbf{x}_t^{\top}\middle|\mathcal{F}_{t-k}\right] - \mathbf{R}\right\|^3\right]$$

is a (2×2)-block matrix of the type:

$$E\left[\mathbf{x}_t\mathbf{x}_t^{\top}\middle|\mathcal{F}_{t-k}\right] - \mathbf{R} = \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^{\top} & \mathbf{B} \end{bmatrix}, \tag{370}$$

with $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ block Toeplitz matrices:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{A}_1 & \cdots & \mathbf{A}_l \\ \mathbf{A}_1^{\top} & \mathbf{A}_0 & \cdots & \mathbf{A}_{l-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_l^{\top} & \mathbf{A}_{l-1}^{\top} & \cdots & \mathbf{A}_0 \end{bmatrix}, \tag{371}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \cdots & \mathbf{B}_l \\ \mathbf{B}_1^{\top} & \mathbf{B}_0 & \cdots & \mathbf{B}_{l-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_l^{\top} & \mathbf{B}_{l-1}^{\top} & \cdots & \mathbf{B}_0 \end{bmatrix}, \tag{372}$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_0 & \mathbf{C}_1 & \cdots & \mathbf{C}_l \\ \mathbf{D}_1^{\top} & \mathbf{C}_0 & \cdots & \mathbf{C}_{l-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_l^{\top} & \mathbf{D}_{l-1}^{\top} & \cdots & \mathbf{C}_0 \end{bmatrix}, \tag{373}$$

where in this case we set $l = p - 1$, with:

$$\mathbf{A}_0 = E\left[\mathbf{y}_t\mathbf{y}_t^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_t\mathbf{y}_t^\top\right], \tag{374}$$

$$\mathbf{B}_0 = E\left[\mathbf{s}_t\mathbf{s}_t^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{s}_t\mathbf{s}_t^\top\right], \tag{375}$$

$$\mathbf{C}_0 = E\left[\mathbf{y}_t\mathbf{s}_t^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_t\mathbf{s}_t^\top\right], \tag{376}$$

and

$$\mathbf{A}_j = E\left[\mathbf{y}_t\mathbf{y}_{t-j}^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_t\mathbf{y}_t^\top\right], \tag{377}$$

$$\mathbf{B}_j = E\left[\mathbf{s}_t\mathbf{s}_{t-j}^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{s}_t\mathbf{s}_t^\top\right], \tag{378}$$

$$\mathbf{C}_j = E\left[\mathbf{y}_t\mathbf{s}_{t-j}^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_t\mathbf{s}_t^\top\right], \tag{379}$$

$$\mathbf{D}_j = E\left[\mathbf{s}_t\mathbf{y}_{t-j}^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{s}_t\mathbf{y}_{t-j}^\top\right], \tag{380}$$

for $j = \{1, 2, \ldots, l\}$. Since the operator norm $\|\cdot\|$ is a Shatten-p norm with $p = \infty$, we can deploy Theorem 1 of Bhatia and Kittaneh [42] for partioned operators:

$$\left\|\begin{bmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^\top & \mathbf{B}\end{bmatrix}\right\|^2 \leq \|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 + 2\|\mathbf{C}\|^2, \tag{381}$$

therefore obtaining:

$$\begin{aligned} E\left[\left\|E\left[\mathbf{x}_t\mathbf{x}_t^\top\middle|\mathcal{F}_{t-k}\right] - \mathbf{R}\right\|^3\right] &\leq E\left[\left(\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 + 2\|\mathbf{C}\|^2\right)^2\right] \\ &\leq E\left[\|\mathbf{A}\|^4\right] + E\left[\|\mathbf{B}\|^4\right] + 4E\left[\|\mathbf{C}\|^4\right] \\ &+ 2E\left[\|\mathbf{A}\|^2\|\mathbf{B}\|^2\right] + 4E\left[\|\mathbf{A}\|^2\|\mathbf{C}\|^2\right] \\ &+ 4E\left[\|\mathbf{B}\|^2\|\mathbf{C}\|^2\right]. \end{aligned} \tag{382}$$

Applying again [42] to each block Toeplitz matrix and the triangle inequality we get, for instance:

$$E\left[\|\mathbf{A}\|^4\right] = E\left[\left(\|\mathbf{A}\|^2\right)^2\right] \leq \boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_2 + \boldsymbol{\Gamma}_3, \tag{383}$$

where:

$$\boldsymbol{\Gamma}_1 = E\left[p^2\left\|E\left[\mathbf{y}_t\mathbf{y}_t^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_t\mathbf{y}_t^\top\right]\right\|^4\right]$$

$$\boldsymbol{\Gamma}_2 = 4E\left[\left\{\sum_{i=0}^{l-1}\sum_{j=i+1}^{l}\left\|E\left[\mathbf{y}_{t-i}\mathbf{y}_{t-j}^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_{t-i}\mathbf{y}_{t-j}^\top\right]\right\|^2\right\}^2\right],$$

$$\begin{aligned}\boldsymbol{\Gamma}_3 = 4pE\Bigg[&\left\|E\left[\mathbf{y}_t\mathbf{y}_t^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_t\mathbf{y}_t^\top\right]\right\|^2, \\ &\times\sum_{i=0}^{l-1}\sum_{j=i+1}^{l}\left\|E\left[\mathbf{y}_{t-i}\mathbf{y}_{t-j}^\top\middle|\mathcal{F}_{t-k}\right] - E\left[\mathbf{y}_{t-i}\mathbf{y}_{t-j}^\top\right]\right\|^2\Bigg].\end{aligned}$$

Focussing on showing that matrices $\mathbf{\Gamma}_1$, $\mathbf{\Gamma}_2$, and $\mathbf{\Gamma}_3$ are o(1), see that:

$$\mathbf{\Gamma}_1 \leq p^2 E\left[\left(\|\boldsymbol{M}_1\| + 2\|\boldsymbol{M}_2\|\right)^4\right],$$

where in this case we define:

$$\boldsymbol{M}_1 = \sum_{j=k}^{\infty} \mathbf{W}_{j,y}\left\{\mathbf{v}_{t-j}\mathbf{v}_{t-j}^{\top} - \mathbf{\Lambda}\right\}\mathbf{W}_{j,y}^{\top} \tag{384}$$

$$\boldsymbol{M}_2 = \sum_{i=k}^{\infty}\sum_{j=i+1}^{\infty} \mathbf{W}_{i,y}\mathbf{v}_{t-i}\mathbf{v}_{t-j}^{\top}\mathbf{W}_{j,y}^{\top}. \tag{385}$$

It suffices to show that $E\left[\|\boldsymbol{M}_1\|^4\right]$ and $E\left[\|\boldsymbol{M}_2\|^4\right]$ vanish asymptotically. By the properties of the Frobenius norm and the Cauchy-Schwartz inequality, for the former we obtain:

$$E\left[\|\boldsymbol{M}_1\|^4\right] \leq E\left[\left(\mathrm{tr}\left\{\boldsymbol{M}_1\right\}\right)^4\right] \tag{386}$$

which by Assumption (iii) and Eq. (357) is a $o(1)$ for $k \to \infty$. Likewise, it can be shown that $E\left[\|\boldsymbol{M}_2\|^4\right] = o(1)$. The proofs for $\mathbf{\Gamma}_2$ and $\mathbf{\Gamma}_3$, and also for matrices $\mathbf{B}$ and $\mathbf{C}$ in Equations ((372), and (373), follow identically. The third part of Condition (C6′) follows similarly, thus completing the proof. □

Remark 20. *To show that (C5′) holds for VARX models, a possible path seems to study the consequences of Wold's representation and extending Theorem 2.1 in [71] to multivariate time series models.*

5.6 CURRENT AND FUTURE RESEARCH

Current research work focusses in extending Theorem 7 to show that general possibly-misspecified VARX(p, q) models, with $p = \infty$, satisfy the set of conditions (C1′-C6′) and (A1′-A5′). Promising results are being obtained in this sense, and interesting questions are arising, e.g. the possible extension of Chan and Ing [71] to multivariate time series models. Future research will include further simulations for different candidate models in the multiple regressor case, the extension to other common possibly-misspecified multivariate time series models, and to the high-dimensionality setting.

Part III

APPENDIX

APPENDIX A

A.1 CHAPTER 1 - TABLE OF IC FOR THE I.I.D. CASE

Table 8: Examples of sample estimators of criteria in the *i.i.d.* setting, in chronological order

Reference	IC
[19]	$PSS = \boldsymbol{y}^T \{\boldsymbol{I}_N - Q(j)\} \{\boldsymbol{I}_N - \Lambda(j)\}^{-2} \{\boldsymbol{I}_N - Q(j)\} \boldsymbol{y}$
[204]	$C_p = \frac{1}{\hat{\sigma}^2} RSS_P - n + 2p$
[204]	$C_L = \frac{1}{\hat{\sigma}^2} RSS_L - n + 2 + 2\mathrm{tr}\{XL\}$
[5, 7]	$AIC = -2l(\hat{\theta}_k) + 2k$
[292]	$TIC = -2l(\hat{\theta}_k) + 2\mathrm{tr}\left\{\hat{I}\hat{J}^{-1}\right\}$
[241, 259]	$BIC = -2l(\hat{\theta}_k) + k\log(n)$
[290]	$AIC_{c_1} = -2l(\hat{\theta}) + \frac{2n(k-c+2)}{n-k+c-3}$
[290]	$AIC_{c_{2a}} = -2l(\hat{\theta}) + \frac{2n(k-b+1)}{n-k+b-2}$
[290]	$AIC_{c_2b} = -2l(\hat{\theta}) + \frac{2n(k-b+(p+1)/2)p}{n-k+b-p-1}$
[290]	$AIC_{c_3a} = -2l(\hat{\theta}) + 2\left[(c+1)\frac{\sum_{i=1}^{c} n_{j_i}}{\sum_{i=1}^{c} n_{j_i}-c-2} + \left\{2\sum_{i=c+1}^{k} \frac{n_{j_i}}{n_{j_i}-3}\right\}\right]$
[290]	$AIC_{c_3b} = -2l(\hat{\theta}) + 2p\left[\left(c+\frac{p+1}{2}\right)\frac{\sum_{i=1}^{c} n_{j_i}}{\sum_{i=1}^{c}(n_{j_i}-c-p-1)} + \left\{\frac{p+3}{2}\sum_{i=c+1}^{k} \frac{n_{j_i}}{n_{j_i}-p-2}\right\}\right]$
[21]	$PC = \hat{\sigma}^2 * \left(1 + \frac{K_1}{T}\right)$
[214]	$GIC = N\log\hat{\sigma}^2 + a_N k$
[52]	$CAIC = -2l(\hat{\theta}_k) + k(\log(n)+1)$
[52]	$CAIFC = -2l(\hat{\theta}_k) + k(\log(n)+2) + \log\lvert I(\hat{\theta}_k)\rvert$
[52]	$KC = -2l(\hat{\theta}_k) - \log f(\theta_k^*) + k\log(n) + \log\lvert B(\hat{\theta}_k)\rvert$
[278]	$RIC = -2l(\hat{\theta}_k(\lambda)) + 2\mathrm{tr}\left\{\hat{I}(\lambda)\hat{J}(\lambda)^{-1}\right\}$
[235]	$D_n(k) = -2l(\hat{\theta}_k) + 2C_n$
[235]	$GIC = S_k + k\hat{\sigma}_m C_n$
[116]	$RIC = RSS_\gamma + \lvert\gamma\rvert\hat{\sigma}^2_{LS}(2\log p)$
[173]	$GIC = -2l(\hat{\theta}_k) + 2b_1(\hat{G})$
[173]	$GIC_R = -2l(\hat{\theta}_k) + 2b_M^{(1)}(\hat{G})$
[173]	$GAIC = -2l(\hat{\theta}_k) + 2\mathrm{tr}\left\{I(\hat{G})J(\hat{G})^{-1}\right\}$
[173]	$GIC_\lambda = -2l(\hat{\theta}(\lambda)) + 2\mathrm{tr}\left\{\hat{I}_\lambda(T(\hat{G}))\hat{J}_\lambda(T(\hat{G}))^{-1}\right\}$
[173]	$GBIC = -2l(h(x,n)) + 2\mathrm{tr}\left\{T^{(1)}(X;\hat{G})\frac{\partial\log f(X\lvert\theta)}{\partial\theta^T}\right\}$
[266]	$GIC_{\lambda_n} = \frac{S_n(\alpha)}{n} + \frac{\lambda_n\hat{\sigma}_n^2 p_n(\alpha)}{n}$
[349]	$IC = -\sum_{i=1}^{n}\log f_k(X_i, \hat{\theta}^{(k)}) + \lambda_k m_k$
[283]	$DIC = D(\bar{\theta}) + 2p_D$
[315]	$TIC_{CL} = -2l_{CL}(\hat{\theta}_k^{CL}) + 2\mathrm{tr}\left\{\hat{I}_{CL}\hat{J}_{CL}^{-1}\right\}$
[324, 326]	$WAIC(n) = B_t L(n) + \frac{\beta}{n}V(n)$
[200]	$GAIC = -2l_n(\boldsymbol{y}, \hat{\beta}_n) + 2\mathrm{tr}\left\{\hat{\boldsymbol{H}}_n\right\}$
[200]	$GBIC = -2l_n(\boldsymbol{y}, \hat{\beta}_n) + \log(n)\lvert\mathcal{M}\rvert + \mathrm{tr}\left\{\hat{\boldsymbol{H}}_{\boldsymbol{n}}\right\} - \log\lvert\hat{\boldsymbol{H}}_{\boldsymbol{n}}\rvert$
[325]	$WBIC = nL_n(w_0) + \lambda\log n$

A.2 CHAPTER 3 - BIBLIOGRAPHIC NOTES

A.2.1 *Table of IC and PC for time series models*

Table 9: Examples of approximate estimates of IC and PC in parametric time series, in chronological order

Reference	IC
[2]	$\text{FPE}_\alpha = (1+N^{-\alpha}(M+1))(1-N^{-1}(M+1))^{-1}S(M)$
[41]	$\text{FPE}^\beta = \hat{\sigma}_k^2(1+\beta k/T)$
[137]	$\text{HQ} = -2l(\hat{\theta}_k) + 2kc\log\log(n)$
[217]	$\text{AIC}_{AR} = n\log\hat{s}_p^2 + 2(p+1)$
[275]	$\text{FPE}_\gamma(k) = n\hat{\sigma}^2(k) + \gamma k\tilde{\sigma}^2(K)$
[243]	$\text{APE} = n^{-1}\sum_{t=0}^{n-1}(x_{t+1}-\hat{x}_{t+1})^2$
[356]	$I(k,C_n) = -2l(\hat{\theta}_k) + kC_n$
[157]	$\text{AIC}_c = n\log\hat{\sigma}^2 + n\frac{1+m/n}{1-(m+2)/n}$
[227]	$\text{GIC}_A = \log\hat{\sigma}_T^2(M) + \text{size}(M)C(T)/T$
[227]	$\text{GIC}_B = \hat{\sigma}_T^2(M) + \text{size}(M)C(T)/T$
[328]	$\text{FIC} = n\hat{\sigma}_n^2 + \tilde{\sigma}_n^2\log\left\lvert\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T\right\rvert$
[183]	$\text{QAIC} = -2\frac{l\hat{\theta}_k}{\hat{c}} + 2K$
[183]	$\text{QAIC}_c = QAIC + \frac{2k(k+1)}{n-k-1}$
[354]	$\text{ODQ} = n\hat{\sigma}_n^2(p,q) - n\hat{\sigma}_n^2(p^*,q^*) - a_n$
[180]	$\text{FIC}\ (k) = n\hat{\sigma}_{n,k}^2 + \hat{\sigma}_{n,K^*}^2\log\left\lvert\sum_{i=1}^n\left(\bigtriangledown g_{i,k}\left(\hat{\theta}_n^{(k)}\right)\right)\left(\bigtriangledown g_{i,k}\left(\hat{\theta}_n^{(k)}\right)\right)^T\right\rvert$
[341]	$\text{WIC} = n\log\left(\hat{\sigma}^2\right) + \frac{(2n(p+1)/(n-p-2))^2(p\log(n))^2}{2n(p+1)/(n-p-2)+p\log(n)}$
[269]	$\text{RIC} = (n-k)\log(\tilde{\sigma}^2) + \log\lvert\hat{W}\rvert + k\log(n) - k + \frac{4}{n-k-2}$
[43]	$\text{EIC} = -2l(\hat{\theta}) + 2k_q q$

A.2.2 *From parametric to nonparametric regression*

If the parameter space is of infinite dimension, then a parametric model considering a finite parameter space would lead to model misspecification and further issues. The Generalized Likelihood Ratio (GLR) test [111] allows to compare a parametric versus a nonparametric specification, and exhibits the Wilks' phenomenon, i.e. its distribution is independent of nuisance parameters, also in nonparametric regression. It is a generalization of the types of Kolmogorov-Smirnov and the Cramér-von Mises statistics. See Fan and Jiang [108] for further details.

THE GENERALIZED LR TEST According to Fan and Yao [110, p. 406], the idea in the GLR test of Fan et al. [111] (developed for independent observations), can be extended to dependent data, although acknowledging that (at their time) there were very few developments on model validation with dependent data for nonparametric regression. They expected that under some mixing conditions[1] the results for dependent data will hold. See Zhou [358] for an analysis of the GLR for time varying coefficient models with cross-correlated non-stationary regressors and errors, and Niu et al. [215] for a bias reduction proposal and an enhancement including dimension reduction adaptive to the model to improve power performance.

PARAMETRIC OR NONPARAMETRIC? The debate between parametric *versus* nonparametric has reappeared many times in literature, without a definitive result, i.e. it depends on whether the object of study has a parametric form or not, and on its objective (description-identification; prediction-selection). As recalled in the introduction, an important fact is that we continue to experience exponential growth of our processors [95], which allows us to implement algorithms that five or ten years before were not feasible in terms of computational times. From the theoretical and applied works in the involved fields, we can see combination of methods to solve the specific problems. These solutions are coherent to principles such as Akaike's entropy-maximization, Rissanen's minimum description length, or parsimony. Nevertheless, caution is advised whenever a new problem is being assessed and MS is needed. We need to understand the specific application and assumptions where the automated criterion is being employed, and need of critical understanding if a specific criterion needs of adaptations or even if it fits our practical purpose. This mainly in the light of possible automated-bias (as in the cases presented in the the previous chapter, Section 2.1). Further interesting considerations can be found respectively in Wiener [333], Bynum [62], and in the introductory chapter of Machol and Gray [202].

1 See Appendix A.2.4.

In his monograph, Wasserman [322] acknowledged that the definition of nonparametric inference is problematic in itself. He defined that nonparametric inference is a set of modern statistical methods to solve real-world statistical problems that aim to keep the number of underlying assumptions as weak as possible. CV is an example of this method [20, 24, 87, 264, 285]. For instance, to perform bandwidth selection in nonparametric regression, CV or Mallow's C_p can be employed.

DIMENSION REDUCTION AND VARIABLE SELECTION Further examples include variable selection with decision trees, adaptive bandwidth selection (Fan and Gijbels [107]). Principal Components Analysis (PCA), independent component analysis, or projection pursuit. It should be clear that the nonparametric approach offers less constraints to the researcher. See Fan [106] for a review of nonparametric methods in financial econometrics, Wasserman [322] for an introduction to nonparametric statistics, Li [191] for the study of nonparametric methods in econometrics, Robinson [248] for asymptotic theory of nonparametric regression with spatial data, and Racine et al. [231] for applied nonparametric econometrics and statistics.

A.2.3 *A sequence of developments in modelling nonlinear time series*

The beginning of spectral analysis is sometimes attributed to the work of Schuster [257], where the periodogram was used to study hidden periodicities in meteorological phenomena. Almost one hundred year later, Robinson [247] presented kernel estimators for the multivariate probability density and for regression, applied to strictly stationary univariate time series. For MS with nonlinear models, Peña Sánchez de Rivera [246] proposed a procedure based on Fisher's Efficient Score principle [233] for classes of nonlinear models, known in the econometric literature as LM test. The author analysed IBM time series and discussed operational aspects of their implementation in the context of diagnostic analysis for an integrated autoregressive moving-average (ARIMA) model.

To the best of our knowledge, Cheng and Tong [77] established the first rigorous work on the theory of nonparametric regression for time series, where the asymptotic consistency of a CV method as MS technique is proven. Their approach connected deterministic chaos and stochastic time series models, starting from the consideration that in any systematic study of chaos it is natural to determine the embedding dimension in a noisy environment first. The setting is of stochastic modelling within the framework of nonlinear regression. They introduced a generalized partial autocorrelation statistic and estimated the embedding dimension relying on order determination of an unknown nonlinear regression via CV, and proved its consistency as MS method under global boundedness. At the same time, they showed how this

method served as theoretical justification of the FPE approach from Auestad and Tjøsthein [26], an extension of which will be explained in the following subsection 3.3.1.

Yao and Tong [350] used a CV method based on the kernel estimate of the conditional mean for subset selection of stochastic regressors within the framework of nonlinear stochastic regression. The assumptions include strict stationarity and absolutely regular processes. They showed that the CV selection method is asymptotically consistent. Two kinds of asymptotic efficiency of the selected model are proved, and the results are illustrated with simulations and real data.

Yao and Tong [351] also studied the determination of bandwidth in kernel regression through a generalized CV method. They proved asymptotic optimality under the assumption that observations are strictly stationary and ρ-mixing (see Appendix A.2.4). Their simulations compared the performance of various CV bandwidth selector for dependent data. These show that the ordinary CV method is quite stable in regression estimation with random design, even if data are highly correlated.

Imoto and Konishi [160] derived IC (a type of GIC) for evaluating B-spline nonparametric regression models estimated by PMLE, in the context of misspecification, and for Generalized Linear Models. Their proposed a criterion that was later applied to select the optimal value of the smoothing parameter and the number of knots. See Fan and Yao [110], Gao [119], and Teräsvirta [294] for monographs from 2003-2010 summarizing various results in this area.

A.2.4 *Mixing conditions in stochastic processes*

Doukhan [104] treated meticulously different notions of mixing related to underlying measures of dependence between σ-fields of processes, stronger than ergodicity.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\mathcal{A}$ and $\mathcal{C}$ two sub σ-algebras of $\mathcal{F}$. Then, define the following measures of dependence between $\mathcal{A}$ and $\mathcal{C}$ as:

$$\begin{aligned}
\alpha(\mathcal{A},\mathcal{C}) &= \sup\{|P(A)P(C) - P(A\cap C)|\,, A\in\mathcal{A}, C\in\mathcal{C}\},\\
\beta(\mathcal{A},\mathcal{C}) &= \mathbb{E}\left[\text{ess sup}\,\{|P(C|\mathcal{A}) - P(C)|\,, C\in\mathcal{C}\}\right],\\
\phi(\mathcal{A},\mathcal{C}) &= \sup\left\{\left|P(C) - \frac{P(A\cap C)}{P(A)}\right|, A\in\mathcal{A}, P(A)\neq 0, C\in\mathcal{C}\right\},\\
\psi(\mathcal{A},\mathcal{C}) &= \sup\left\{\left|1 - \frac{P(A\cap C)}{P(A)P(C)}\right|, A\in\mathcal{A}, P(A)\neq 0, C\in\mathcal{C},\right.\\
&\qquad\left. C\neq 0\right\},\\
\rho(\mathcal{A},\mathcal{C}) &= \sup\left\{|\text{Corr}(X,Y)|\,, X\in\mathcal{L}^2(\mathcal{A}), Y\in\mathcal{L}^2(\mathcal{C})A\in\mathcal{A},\right.\\
&\qquad\left. P(A)\neq 0, C\in\mathcal{C}, C\neq 0\right\},
\end{aligned}$$

In particular, coefficient $\beta(\mathcal{A},\mathcal{C})$ is called absolute regularity or β-mixing coefficient, and it may be also written as:

$$\beta(\mathcal{A},\mathcal{C}) = \sup\left\{\frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}|P(A_i)P(C_j) - P(A_i\cap C_j)|\right\}$$

The supremum is taken over all the partitions (A_i), (C_i) of Ω, with $A_i\in\mathcal{A}$, $C_j\in\mathcal{C}$. A more recent use of the β-mixing definition to study time dependence is followed by Chen et al. [75], citing Davydov [91]:

Definition 74. *The process $\{x_t\}$ is β-mixing if*

$$\lim_{t\to\infty}\beta_t = 0 \tag{387}$$

and β-mixing with exponential decay rate if

$$\beta_t \leq \gamma\exp(-\delta t) \tag{388}$$

for some $\delta > 0$ and $\gamma > 0$.

A.2.5 *Local estimation: linear and polynomial*

Let us follow [309, Ch. 7]. Local estimation aims at obtaining an estimate of $\mu(x)$ by estimating it separately for each $(m \times 1)$ vector. Since $\mu(x)$ is not observable, a first-order Taylor expansion of $\mu(x_t)$ at x is considered:

$$\mu(x_t) = \mu(x) + \frac{\partial \mu(x)}{\partial x'}(x_t - x) + R(x_t, x).$$

We can substitute in:

$$Y_t = \mu(x) \cdot 1 + \frac{\partial \mu(x)}{\partial x'}(x_t - x) + R(x_t, x) + \epsilon_t,$$

where ϵ_t is the stochastic error term. Note that if $R(x_t, x) = 0$ then the usual OLS is obtained. In this case the usual estimated function value $\hat{\mu}(x)$ at x, and the estimated vector $\frac{\partial \hat{\mu}(x)}{\partial x'}$ of first partial derivatives at x, follow. But if the conditional mean function, $\mu(x_t)$, is nonlinear then $R(x_t, x) \neq 0$, thus the OLS with biased estimates is required. In nonparametric estimation, kernel estimation requires a kernel function $K(u)$ such that it is:

(i) symmetric;

(ii) compactly supported;

(iii) nonnegative;

(iv) univariate probability density s.t. $\int K(u)du = 1$.

Furthermore, a bandwidth h has to be defined. This will help to adjust a neighbourhood around x, and can be thought as a smoothing parameter, since higher h is associated with smoother function estimates. For a scalar x, this would be of the form: $\frac{1}{h}K(\frac{x_t - x}{h})$. If we are in the vector setting, with $m > 1$, and $x = (x_1, \ldots, x_m)^\top$, then we use Product Kernel, e.g.:

$$K_h(x_t - x) = \prod_{i=1}^{m} \frac{1}{h^m} K(\frac{x_{t_i} - x_i}{h}). \tag{389}$$

The particular choice of the kernel function influences the asymptotic behaviour of the Local Linear Estimator via:

1. the Kernel variance: $\sigma_K^2 = \int u^2 K(u)du$, and
2. the Kernel constant: $||K||_2^2 := \int K(u)^2 du$.

The strategy proposed in [309], is to consider estimation as a weighted LS problem:

$$\{\hat{c}, \hat{c}_1, \ldots, \hat{c}_m\} = \\ \underset{c, c_1, \ldots, c_m}{\operatorname{argmin}} \sum_{t=i_m+1}^{T} \{y_t - c - \sum_{i=1}^{m} c_i(x_{ti} - x_i)\}^2 \cdot K_h(x_t - x). \tag{390}$$

In this way, the Local Linear function estimate at point x is given by $\hat{\mu}(x,h) = \hat{c}$.

Summarizing: the local linear estimator uses a first-order Taylor approximation, useful in the situation of nonparametric analysis. Finally, note that estimating $\mu(\cdot)$ on complete support would imply to perform infinitely many estimations. For this reason, the estimation is usually performed on a specified grid or at specific values.

In this context, the NW estimator [211, 327] is a local constant function estimator:

$$\begin{aligned}\hat{\mu}_{NW}(x,h) &= \{\mathbf{Z}'_{NW}W(x,h)\mathbf{Z}_{NW}\}^{-1}\mathbf{Z}^{\top}_{NW}\mathbf{W}(x,h)\mathbf{Y} \\ &= \frac{\sum_{t=i_m+1}^{T} K_h(x_t - x)y_t}{\sum_{t=i_m+1}^{T} K_h(x_t - x)}\end{aligned} \tag{391}$$

with $\mathbf{Z}_{NW} = (1,\ldots,1)^{\top}_{1\times(T-i_m)}$, $\mathbf{W}(x,h) = diag\left\{K_h\frac{(x_t-x)}{T}\right\}^{\top}_{t=i_{m+1}}$, and $\mathbf{Y} = (Y_{i_{m+1}},\ldots,Y_T)^{\top}$.

The LLE can be seen as a Generalized LS estimator:

$$\hat{\mu}(x,h) = e\{\mathbf{Z}^{\top}(x)W(x,h)\mathbf{Z}(x)\}^{-1}\mathbf{Z}^{\top}(x)W(x,h)\mathbf{Y}, \tag{392}$$

where $\mathbf{Y} = (Y_{i_{m+1}},\ldots,Y_T)^{\top}$, $e = (1, 0_{1\times m})^{\top}$,

$$\mathbf{Z}(x) = \begin{bmatrix} 1 & \ldots & 1 \\ x_{i_m+1} - x & \ldots & x_T - x \end{bmatrix}^{\top},$$

and $W(x,h) = diag\left\{K_h\frac{(x_t-x)}{T}\right\}^{T}_{t=i_{m+1}}$. If regularity conditions are satisfied, then it can be shown that the LLE has asymptotic Gaussian distribution:

$$\sqrt{Th^m}\{\hat{\mu}(x,h) - \mu(x) - b(x)h^2\} \xrightarrow{d} N(0, v(x)), \tag{393}$$

where the asymptotic bias is defined as:

$$b(x) = \frac{\sigma_K^2}{2}\mathrm{tr}\left\{\frac{\partial^2\mu(x)}{\partial x \partial x^{\top}}\right\}, \tag{394}$$

and the asymptotic variance is equal to:

$$v(x) = \frac{\sigma^2(x)||K||_2^{2m}}{f(x)}. \tag{395}$$

The LLE is preferred for its asymptotic properties, since as h increases it can increase the asymptotic bias, while reducing h allows to decrease the asymptotic variance.

B

APPENDIX B

B.1 TECHNICAL LEMMA

The following lemma is a general result that holds for the multivariate predictor case (i.e., $m \geq 1$). It relies upon Assumption (C1) of [153, p. 1068] when $w = 1, m \geq 1$, identical to Assumption (C1′) when $w \geq 1m \geq 2$, which reduces to Assumption (C1) of Chapter 4.

Lemma 1. *Let* $\mathbf{R}$ *and* $\hat{\mathbf{R}}$ *be defined in (210) and* $\tilde{\mathbf{R}}$ *be the multivariate version of Eq. (246). Then, for* $0 < \gamma \leq 5$*, it holds that:*

$$\mathrm{E}\left[\left\|\tilde{\mathbf{R}}^{-1} - \hat{\mathbf{R}}^{-1}\right\|^{\gamma}\right] = O\left[\left(\frac{l_n}{n}\right)^{\gamma}\right], \tag{396}$$

$$\mathrm{E}\left[\left\|\mathbf{R}^{-1} - \hat{\mathbf{R}}^{-1}\right\|^{\gamma}\right] = O\left[n^{-\gamma/2}\right], \tag{397}$$

$$\mathrm{E}\left[\left\|\mathbf{R}^{-1} - \tilde{\mathbf{R}}^{-1}\right\|^{\gamma}\right] = O\left[n^{-\gamma/2}\right], \tag{398}$$

with l_n *being defined in (242).*

Proof. Triangle inequality implies that

$$\begin{aligned}
\left\|\tilde{\mathbf{R}}^{-1} - \hat{\mathbf{R}}^{-1}\right\| &\leq \left\|\hat{\mathbf{R}}^{-1}\right\| \left\|\hat{\mathbf{R}} - \tilde{\mathbf{R}}\right\| \left\|\hat{\mathbf{R}}^{-1}\right\|, \\
\left\|\mathbf{R}^{-1} - \hat{\mathbf{R}}^{-1}\right\| &\leq \left\|\hat{\mathbf{R}}^{-1}\right\| \left\|\hat{\mathbf{R}} - \mathbf{R}\right\| \left\|\mathbf{R}^{-1}\right\|, \\
\left\|\mathbf{R}^{-1} - \tilde{\mathbf{R}}^{-1}\right\| &\leq \left\|\tilde{\mathbf{R}}^{-1}\right\| \left\|\hat{\mathbf{R}} - \mathbf{R}\right\| \left\|\mathbf{R}^{-1}\right\|.
\end{aligned}$$

Since $\mathbf{R}$ is invertible we have that $\|\mathbf{R}^{-1}\| = O(1)$; under Assumption (C5′), $\left\|\hat{\mathbf{R}}^{-1}\right\| = O(1)$. Moreover, it can be easily proved that also $\|\tilde{\mathbf{R}}^{-1}\| = O(1)$. Therefore, by deploying Hölder's inequality, the results will be verified if we prove the following three conditions:

$$\mathrm{E}\left[\left\|\hat{\mathbf{R}} - \tilde{\mathbf{R}}\right\|\right] = O\left(\frac{l_n}{n}\right), \tag{399}$$

$$\mathrm{E}\left[\left\|\hat{\mathbf{R}} - \mathbf{R}\right\|\right] = O\left(n^{-1/2}\right), \tag{400}$$

$$\mathrm{E}\left[\|\tilde{\mathbf{R}} - \mathbf{R}\|\right] = O\left(n^{-1/2}\right). \tag{401}$$

Let $a = n - l_n$ and $b = a + 1$. As for (399) note that

$$\begin{aligned}
\mathrm{E}\left[\left\|\hat{\mathbf{R}} - \tilde{\mathbf{R}}\right\|\right] &\leq \mathrm{E}\left[\left\|\left(\frac{1}{N} - \frac{1}{a}\right)\sum_{t=1}^{a}\mathbf{x}_t\mathbf{x}_t^{\top}\right\|\right] + \mathrm{E}\left[\left\|\frac{1}{N}\sum_{t=b}^{N}\mathbf{x}_t\mathbf{x}_t^{\top}\right\|\right] \\
&= O\left(\frac{l_n}{n}\right).
\end{aligned}$$

Conditions (400) and (401) readily derive from Assumption (C1) and hence the proof is completed. □

BIBLIOGRAPHY

[1] Hirotugu Akaike. "Fitting autoregressive models for prediction." In: *Annals of the Institute of Statistical Mathematics* 21.1 (1969), pp. 243–247.

[2] Hirotugu Akaike. "Statistical predictor identification." In: *Annals of the Institute of Statistical Mathematics* 22.1 (1970), pp. 203–217.

[3] Hirotugu Akaike. "Autoregressive model fitting for control." In: *Annals of the Institute of Statistical Mathematics* 23 (1971), 163—180.

[4] Hirotugu Akaike. *Determination of the Number of Factors by an Extended Maximum Likelihood Principle*. Research memorandums / Institute of Statistical Mathematics. Tokyo. Institute of Statistical Mathematics, 1971.

[5] Hirotugu Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle." In: *Proceeding of the Second International Symposium on Information Theory* (1973). Ed. by B. Petrov and F. Caski.

[6] Hirotugu Akaike. "Information theory and an extension of the maximum likelihood principle." In: *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. 1973, pp. 267–281.

[7] Hirotugu Akaike. "A new look at the statistical model identification." In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723.

[8] Hirotugu Akaike. "A new look at the statistical model identification." In: *IEEE Transactions on Automatic Control* AC-19 (1974), pp. 716–723. ISSN: 0018-9286. DOI: 10.1109/tac.1974.1100705.

[9] Hirotugu Akaike. "Canonical correlation analysis of time series and the use of an information criterion." In: *Mathematics in Science and Engineering*. Vol. 126. Elsevier, 1976, pp. 27–96.

[10] Hirotugu Akaike. "On entropy maximization principle, Applications of Statistics." In: *Proceedings of the Symposium held at Wright State University*. North-Holland Publishing Company. 1977, pp. 27–41.

[11] Hirotugu Akaike. "A new look at the Bayes procedure." In: *Biometrika* 65.1 (1978), pp. 53–59.

[12] Hirotugu Akaike. "On the likelihood of a time series model." In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 27.3-4 (1978), pp. 217–235.

[13] Hirotugu Akaike. "A Bayesian extension of the minimum AIC procedure of autoregressive model fitting." In: *Biometrika* 66.2 (1979), pp. 237–242.

[14] Hirotugu Akaike. "Likelihood and the Bayes procedure." In: *Bayesian Statistics. Proceedings of the First International Meeting held in Valencia (Spain), May 28 to June 2*. University Press (Valencia, Spain), 1980, pp. 115–149.

[15] Hirotugu Akaike. "Prediction and entropy." In: *A Celebration of Statistics: The ISI Centenary Volume*. Ed. by Anthony Atkinson and Stephen Fienberg. Springer, 1985, pp. 1–24.

[16] Hirotugu Akaike. "Information Theory and an Extension of the Maximum Likelihood Principle." In: *Breakthroughs in Statistics: foundations and basic theory*. Ed. by Samuel Kotz and Norman Lloyd Johnson. Springer, 1992, pp. 610–624.

[17] Hirotugu Akaike. "Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes." In: *Selected Papers of Hirotugu Akaike*. Springer, 1998, pp. 223–247.

[18] Abdelkamel Alj, Kristján Jónasson, and Guy Mélard. "The exact Gaussian likelihood estimation of time-dependent VARMA models." In: *Computational Statistics & Data Analysis* 100 (2016), pp. 633–644.

[19] David Allen. "The prediction sum of squares as a criterion for selecting predictor variables." In: *Technical Report* 23 (1971).

[20] David Allen. "The relationship between variable selection and data agumentation and a method for prediction." In: *Technometrics* 16.1 (1974), pp. 125–127.

[21] Takeshi Amemiya. "Selection of regressors." In: *International Economic Review* (1980), pp. 331–354.

[22] Theodore Anderson and AM Walker. "On the asymptotic distribution of the autocorrelations of a sample from a linear stochastic process." In: *The Annals of Mathematical Statistics* 35.3 (1964), pp. 1296–1303.

[23] Serkan Aras and İpek Deveci Kocakoç. "A new model selection strategy in time series forecasting with artificial neural networks: IHTS." In: *Neurocomputing* 174 (2016), pp. 974–987.

[24] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection." In: *Statistics Surveys* 4 (2010), pp. 40–79.

[25] Anthony Atkinson. "A note on the generalized information criterion for choice of a model." In: *Biometrika* 67.2 (1980), pp. 413–418.

[26] Bjørn Auestad and Dag Tjøstheim. "Identification of nonlinear time series: first order characterization and order determination." In: *Biometrika* 77.4 (1990), pp. 669–687.

[27] Rina Foygel Barber and Emmanuel Candès. "Controlling the false discovery rate via knockoffs." In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[28] Jean-Marc Bardet, Kare Kamila, and William Kengne. "Consistent model selection criteria and goodness-of-fit test for common time series models." In: *Electronic Journal of Statistics* 14.1 (2020), pp. 2009–2052.

[29] Andrew Barron, Lucien Birgé, and Pascal Massart. "Risk bounds for model selection via penalization." In: *Probability Theory and Related Fields* 113.3 (1999), pp. 301–413.

[30] Edward Bedrick and Chih-Ling Tsai. "Model selection for multivariate regression in small samples." In: *Biometrics* (1994), pp. 226–231.

[31] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57.1 (1995), pp. 289–300.

[32] Joseph Berkson. "Some difficulties of interpretation encountered in the application of the chi-square test." In: *Journal of the American Statistical Association* 33.203 (1938), pp. 526–536.

[33] Joseph Berkson. "Estimation by least squares and by maximum likelihood." In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. University of California Press, Berkeley. 1956, pp. 1–11.

[34] Emily Berman. "A government of laws and not of machines." In: *Boston University Law Review* 98.5 (2018), pp. 1277–1356.

[35] Emily Berman. "Individualized Suspicion in the Age of Big Data." In: *Iowa Law Review* 105 (2019), p. 463.

[36] Dimitris Bertsimas, Angela King, and Rahul Mazumder. "Best subset selection via a modern optimization lens." In: *The Annals of Statistics* 44.2 (2016), pp. 813–852.

[37] Dimitris Bertsimas, Jean Pauphilet, and Bart Van Parys. "Sparse regression: Scalable algorithms and empirical performance." In: *Statistical Science* 35.4 (2020), pp. 555–578.

[38] Wolfgang Bessler, Alexander Leonhardt, and Dominik Wolff. "Analyzing hedging strategies for fixed income portfolios: A Bayesian approach for model selection." In: *International Review of Financial Analysis* 46 (2016), pp. 239–256. ISSN: 1057-5219. DOI: https://doi.org/10.1016/j.irfa.2015.11.013. URL:

https://www.sciencedirect.com/science/article/pii/S1057521915002100.

[39] Rajendra Bhansali. "The criterion autoregressive transfer function of Parzen." In: *Journal of Time Series Analysis* 7.2 (1986), pp. 79–104.

[40] Rajendra Bhansali. "Model specification and selection for multivariate time series." In: *Journal of Multivariate Analysis* 175 (2020), p. 104539.

[41] Rajendra Bhansali and David Downham. "Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion." In: *Biometrika* 64.3 (1977), pp. 547–551.

[42] Rajendra Bhatia and Fuad Kittaneh. "Norm inequalities for partitioned operators and an application." In: *Mathematische Annalen* 287 (1990), pp. 719–726. DOI: 10.1007/BF01446925.

[43] Baki Billah, Rob Hyndman, and Anne Koehler. "Empirical information criteria for time series forecasting model selection." In: *Journal of Statistical Computation and Simulation* 75.10 (2005), pp. 831–840.

[44] Nicholas Bingham. "Multivariate prediction and matrix Szegö theory." In: *Probability Surveys* 9 (2012), pp. 325–339.

[45] Richard Blahut. "Hypothesis testing and information theory." In: *IEEE Transactions on Information Theory* 20.4 (1974), pp. 405–417.

[46] Arnoud den Boer and Dirk Sierag. "Decision-based model selection." In: *European Journal of Operational Research* 290.2 (2021), pp. 671–686.

[47] Małgorzata Bogdan, Ewout Van Den Berg, Chiara Sabatti, Weijie Su, and Emmanuel Candès. "SLOPE—adaptive variable selection via convex optimization." In: *The Annals of Applied Statistics* 9.3 (2015), p. 1103.

[48] Aurélie Boisbunon, Stéphane Canu, Dominique Fourdrinier, William Strawderman, and Martin Wells. "Akaike's information criterion, Cp and estimators of loss for elliptically symmetric distributions." In: *International Statistical Review* 82.3 (2014), pp. 422–439.

[49] Yacouba Boubacar Mainassara. "Selection of weak VARMA models by modified Akaike's information criteria." In: *Journal of Time Series Analysis* 33.1 (2012), pp. 121–130.

[50] Anton Bovier. "Stochastic processes I." Online. Sommer 2009, Universität Bonn. Bonn, 2011. URL: https://wt.iam.uni-bonn.de/bovier/lecture-notes.

[51] George Box, Gwilym Jenkins, Gregory Reinsel, and Greta Ljung. *Time Series Analysis: Forecasting and Control*. 2016.

[52] Hamparsum Bozdogan. "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions." In: *Psychometrika* 52.3 (1987), pp. 345–370.

[53] Hamparsum Bozdogan. "On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models." In: *Communications in Statistics – Theory and Methods* 19.1 (1990), pp. 221–278.

[54] Leo Breiman. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." In: *Statistical Science* 16.3 (2001), pp. 199–231.

[55] Heinz-Peter Breuer and Francesco Petruccione. *The theory of open quantum systems*. Oxford Academic, 2002.

[56] Peter Brockwell, Richard Davis, and Stephen Fienberg. *Time series: theory and methods*. Springer Science & Business Media, 1991.

[57] Steven Buckland, Kenneth Burnham, and Nicole Augustin. "Model selection : an integral part of inference." In: *Biometrics* 53 (2 1997), pp. 603 –618.

[58] Peter Bühlmann, Bin Yu, Yoram Singer, and Larry Wasserman. "Sparse Boosting." In: *Journal of Machine Learning Research* 7.6 (2006).

[59] Kenneth Burnham and David Anderson. *Model selection and multimodel inference: a practical information- theoretic approach*. 2nd ed. Springer, New York, 2002.

[60] Kenneth Burnham and David Anderson. "Multimodel inference: understanding AIC and BIC in model selection." In: *Sociological Methods & Research* 33.2 (2004), pp. 261–304.

[61] Adolf Buse. "The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note." In: *The American Statistician* 36.3a (1982), pp. 153–157.

[62] Terrell Ward Bynum. "Norbert Wiener's Vision: the Impact of the 'Automatic Age' on our Moral Lives." In: *The impact of the Internet on our moral lives*. Ed. by Robert Cavalier. State University of New York Press, 2005, pp. 11–25.

[63] Emmanuel Candès. "Conformal prediction in 2020." Bernoulli-IMS One World Symposium 2020. 2020. URL: https://www.worldsymposium2020.org/program/live-plenary-talks.

[64] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. "Panning for gold:'model-X' knockoffs for high dimensional controlled variable selection." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3 (2018), pp. 551–577.

[65] Mehmet Caner and Marcelo Medeiros. "Model Selection and Shrinkage: An Overview." In: *Econometric Reviews* 35.8-10 (2016), pp. 1343–1346.

[66] Joseph Cavanaugh. "A large-sample model selection criterion based on Kullback's symmetric divergence." In: *Statistics & Probability Letters* 42.4 (1999), pp. 333–343. ISSN: 0167-7152. DOI: https://doi.org/10.1016/S0167-7152(98)00200-4. URL: https://www.sciencedirect.com/science/article/pii/S0167715298002004.

[67] Joseph Cavanaugh and Andrew Neath. "The Akaike Information Criterion: background, derivation, properties, application, interpretation, and refinements." In: *Wiley Interdisciplinary Reviews: Computational Statistics* 11.3 (2019), e1460.

[68] Arijit Chakrabarti and Jayanta Ghosh. "Optimality of AIC in inference about Brownian motion." In: *Annals of the Institute of Statistical Mathematics* 58.1 (2006), pp. 1–20.

[69] Joshua Chan, Eric Eisenstat, and Gary Koop. "Large Bayesian VARMAs." In: *Journal of Econometrics* 192.2 (2016), pp. 374–390.

[70] Kung Sik Chan and Howell Tong. "On estimating thresholds in autoregressive models." In: *Journal of Time Series Analysis* 7.3 (1986), pp. 179–190.

[71] Ngai Hang Chan and Ching-Kang Ing. "Uniform moment bounds of Fisher's Information with applications to time series." In: *The Annals of Statistics* 39.3 (2011), pp. 1526–1550. ISSN: 00905364, 21688966. URL: http://www.jstor.org/stable/23033606.

[72] Jia Chen, Degui Li, Oliver Linton, and Zudi Lu. "Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series." In: *Journal of the American Statistical Association* 113.522 (2018), pp. 919–932.

[73] Min Chen, Michael Dunn, Amos Golan, and Aman Ullah. *Advances in Info-metrics: Information and Information Processing Across Disciplines*. Oxford University Press, 2020.

[74] Shaobing Chen and David Donoho. "Basis pursuit." In: *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*. Vol. 1. IEEE. 1994, pp. 41–44.

[75] Xiaohong Chen, Lars Peter Hansen, and Marine Carrasco. "Nonlinearity and temporal dependence." In: *Journal of Econometrics* 155.2 (2010), pp. 155–169.

[76] Yuansi Chen, Armeen Taeb, and Peter Bühlmann. "A Look at Robustness and Stability of l_1 versu l_2-Regularization: Discussion of Papers by Bertsimas et al. and Hastie et al." In: *Statistical Science* 35.4 (2020), pp. 614–622.

[77] Bing Cheng and Howell Tong. "On consistent nonparametric order determination and chaos." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 54.2 (1992), pp. 427–449.

[78] Hai-Tang Chiou, Meihui Guo, and Ching-Kang Ing. "Variable selection for high-dimensional regression models with time series and heteroscedastic errors." In: *Journal of Econometrics* 216.1 (2020), pp. 118–136.

[79] ByoungSeon Choi. *ARMA model identification*. Springer Science & Business Media, 1992.

[80] Yuan Chow. "Local convergence of martingales and the law of large numbers." In: *The Annals of Mathematical Statistics* 36.2 (1965), pp. 552–558.

[81] Ole Christensen. *Functions, spaces, and expansions: mathematical tools in physics and engineering*. Springer Science & Business Media, 2010.

[82] Gerda Claeskens and Nils Lid Hjort. "The focused information criterion." In: *Journal of the American Statistical Association* 98.464 (2003), pp. 900–916.

[83] Gerda Claeskens and Nils Lid Hjort. Cambridge Books. Cambridge University Press, 2008.

[84] Jennifer Cobbe. "Administrative law and the machines of government: judicial review of automated public-sector decision-making." In: *Legal Studies* 39.4 (2019), pp. 636–655.

[85] David Cooper and Eric Wood. "Identifying multivariate time series models." In: *Journal of Time Series Analysis* 3.3 (1982), pp. 153–164.

[86] David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. 2017 edition. CRC Press, 1979.

[87] Peter Craven and Grace Wahba. "Smoothing noisy data with spline functions." In: *Numerische Mathematik* 31.4 (1978), pp. 377–403.

[88] Imre Csiszár and Paul Shields. "The consistency of the BIC Markov order estimator." In: *The Annals of Statistics* 28.6 (2000), pp. 1601–1619.

[89] Imre Csiszár and Zsolt Talata. "Context tree estimation for not necessarily finite memory processes, via BIC and MDL." In: *IEEE Transactions on Information Theory* 52.3 (2006), pp. 1007–1016.

[90] Simon Davies, Andrew Neath, and Joseph Cavanaugh. "Estimation optimality of corrected AIC and modified Cp in linear regression." In: *International Statistical Review* 74.2 (2006), pp. 161–168.

[91] Yu Davydov. "Mixing conditions for Markov chains." In: *Theory of Probability & Its Applications* 18.2 (1973), pp. 312–328.

[92] Frauke Degenhardt, Stephan Seifert, and Silke Szymczak. "Evaluation of variable selection methods for random forests and omics data sets." In: *Briefings in Bioinformatics* 20.2 (2019), pp. 492–503.

[93] Edwards Deming. "On probability as a basis for action." In: *The American Statistician* 29.4 (1975), pp. 146–152.

[94] Emre Demirkaya, Yang Feng, Pallavi Basu, and Jinchi Lv. "Large-scale model selection in misspecified generalized linear models." In: *Biometrika* (2020), pp. 1–21.

[95] Peter Denning and Ted Lewis. "Exponential Laws of Computing Growth." In: *Communications of the ACM* 60.1 (2017), pp. 54–65.

[96] PH Diananda and Maurice Bartlett. "Some probability limit theorems with statistical applications." In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 49. 2. Cambridge University Press. 1953, pp. 239–246.

[97] Gery Andrés Díaz Rubio, Simone Giannerini, and Greta Goracci. "On the asymptotic mean-squared prediction error for multivariate time series." In: *Book of Short Papers SIS 2021*. Ed. by Cira Perna, Nicola Salvati, and Francesco Schirripa Spagnolo. Pearson, 2021, pp. 1599 –1604.

[98] Gery Andrés Díaz Rubio, Simone Giannerini, and Greta Goracci. "A multivariate extension of the Misspecification-Resistant Information Criterion." In: *arXiv preprint* (2022). URL: https://arxiv.org/abs/2202.09225.

[99] Jie Ding, Vahid Tarokh, and Yuhong Yang. "Bridging AIC and BIC: a new criterion for autoregression." In: *IEEE Transactions on Information Theory* 64.6 (2017), pp. 4024–4043.

[100] Jie Ding, Vahid Tarokh, and Yuhong Yang. "Model selection techniques: An overview." In: *IEEE Signal Processing Magazine* 35.6 (2018), pp. 16–34.

[101] Mamadou Lamine Diop and William Kengne. "Inference and model selection in general causal time series with exogenous covariates." In: *Electronic Journal of Statistics* 16.1 (2022), pp. 116–157.

[102] Matthew Dixon and Tyler Ward. "Takeuchi's Information Criteria as a form of Regularization." In: *arXiv preprint* (2018). DOI: 10.48550/arXiv.1803.04947.

[103] Joseph Doob. "Probability and statistics." In: *Transactions of the American Mathematical Society* 36.4 (1934), pp. 759–775.

[104] Paul Doukhan. *Mixing: Properties and Examples*. Lecture Notes in Statistics. Springer, 1994.

[105] Bradley Efron. "The estimation of prediction error: covariance penalties and cross-validation." In: *Journal of the American Statistical Association* 99.467 (2004), pp. 619–632.

[106] Jianqing Fan. "A selective overview of nonparametric methods in financial econometrics." In: *Statistical Science* (2005), pp. 317–337.

[107] Jianqing Fan and Irene Gijbels. "Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57.2 (1995), pp. 371–394.

[108] Jianqing Fan and Jiancheng Jiang. "Nonparametric inference with generalized likelihood ratio tests." In: *Test* 16.3 (2007), pp. 409–444.

[109] Jianqing Fan and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." In: *Journal of the American Statistical Association* 96.456 (2001), pp. 1348–1360.

[110] Jianqing Fan and Qiwei Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Series in Statistics (SSS). Springer-Verlag, New York, 2003.

[111] Jianqing Fan, Chunming Zhang, and Jian Zhang. "Generalized likelihood ratio statistics and Wilks phenomenon." In: *The Annals of Statistics* (2001), pp. 153–193.

[112] Vicky Fasen and Sebastian Kimmig. "Information criteria for multivariate CARMA processes." In: *Bernoulli* 23.4 (A 2017), pp. 2860 –2886.

[113] David Findley. "On the unbiasedness property of AIC for exact or approximating linear stochastic time series models." In: *Journal of Time Series Analysis* 6.4 (1985), pp. 229–252.

[114] David Findley. "Counterexamples to parsimony and BIC." In: *Annals of the Institute of Statistical Mathematics* 43.3 (1991), pp. 505–514.

[115] David Findley and Ching-Zong Wei. "Moment bounds for deriving time series CLT's and model selection procedures." In: *Statistica Sinica* 3 (1993), pp. 453–480.

[116] Dean Foster and Edward George. "The risk inflation criterion for multiple regression." In: *The Annals of Statistics* (1994), pp. 1947 –1975.

[117] Rina Foygel Barber, Emmanuel Candès, Aaditya Ramdas, and Ryan Tibshirani. "The limits of distribution-free conditional predictive inference." In: *Information and Inference: A Journal of the IMA* 10.2 (2021), pp. 455–482.

[118] Racnar Frisch. "From Utopian Theory to Practical Applications: The Case of Econometrics." In: *Nobel Memorial Lecture, in: (Nobel Lectures, Economic Sciences, 1969–1980, Assar Lindbeck ed.) Singapore, World Scientific* (1992).

[119] Jiti Gao. *Nonlinear time series: semiparametric and nonparametric methods*. CRC Press, 2007.

[120] Jesús García and Verónica González-López. "Consistent estimation of partition Markov models." In: *Entropy* 19.4 (2017), p. 160.

[121] Seymour Geisser. "The predictive sample reuse method with applications." In: *Journal of the American Statistical Association* 70.350 (1975), pp. 320–328.

[122] Andrew Gelman and Aki Vehtari. "What are the most important statistical ideas of the past 50 years?" In: *Journal of the American Statistical Association* 116.536 (2021), pp. 2087–2097.

[123] John Geweke and Richard Meese. "Estimating regression models of finite but unknown order." In: *International Economic Review* (1981), pp. 55–70.

[124] Christophe Giraud. *Introduction to high dimensional statistics*. Vol. 138. CRC Press, 2014.

[125] Jan de Gooijer, Bovas Abraham, Ann Gould, and Lecily Robinson. "Methods for determining the order of an autoregressive-moving average process: A survey." In: *International Statistical Review/Revue Internationale de Statistique* (1985), pp. 301–329.

[126] Clive Granger and Timo Teräsvirta. "Modelling non-linear economic relationships." In: *OUP Catalogue* (1993).

[127] Peter Grünwald. *The minimum description length principle*. MIT Press, 2007.

[128] Valérie Haggan and Tohru Ozaki. "Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model." In: *Biometrika* 68.1 (1981), pp. 189–196.

[129] Alastair Hall. *Generalized method of moments*. Oxford University Press, 2005.

[130] Marc Hallin, Davide La Vecchia, and Hang Liu. "Center-outward R-estimation for semiparametric VARMA models." In: *Journal of the American Statistical Association* (2020), pp. 1–14.

[131] Paul Richard Halmos. *Measure Theory*. Vol. 18. Graduate Texts in Mathematics. 1974.

[132] Edward Hannan. *Time series analysis*. 1960.

[133] Edward Hannan. *Multiple time series*. John Wiley & Sons, 1970.

[134] Edward Hannan. "The estimation of the order of an ARMA process." In: *The Annals of Statistics* (1980), pp. 1071–1081.

[135] Edward Hannan. "Testing for autocorrelation and Akaike's criterion." In: *Journal of Applied Probability* 19.A (1982), pp. 403–412.

[136] Edward Hannan and Manfred Deistler. *The statistical theory of linear systems*. SIAM, 2012.

[137] Edward Hannan and Barry Quinn. "The determination of the order of an autoregression." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 41.2 (1979), pp. 190–195.

[138] Edward Hannan and Jorma Rissanen. "Recursive estimation of mixed autoregressive-moving average order." In: *Biometrika* 69.1 (1982), pp. 81–94.

[139] Bruce Hansen. *Econometrics*. Online version (unpublished), 2020.

[140] Mark Hansen and Bin Yu. "Bridging AIC and BIC: an MDL model selection criterion." In: *Proceedings of IEEE Information Theory Workshop on Detection, Estimation, Classification and Imaging*. Vol. 63. IEEE Information Theory Society, Santa Fe. 1999.

[141] Mark Hansen and Bin Yu. "Model selection and the principle of minimum description length." In: *Journal of the American Statistical Association* 96.454 (2001), pp. 746–774.

[142] Wolfgang Härdle, Helmut Lütkepohl, and Rong Chen. "A review of nonparametric time series analysis." In: *International Statistical Review* 65.1 (1997), pp. 49–72.

[143] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[144] Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons." In: *Statistical Science* 35.4 (2020), pp. 579–592.

[145] Henry Helson and David Lowdenslager. "Vector-valued processes." In: *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. II*. 1961, pp. 203–212.

[146] EM Hemerly and MHA Davis. "Strong consistency of the PLS criterion for order determination of autoregressive processes." In: *The Annals of Statistics* (1989), pp. 941–946.

[147] Ronald Hocking. "Criteria for selection of a subset regression: which one should be used?" In: *Technometrics* 14.4 (1972), pp. 967–976.

[148] Arthur Hoerl and Robert Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." In: *Technometrics* 12.1 (1970), pp. 55–67.

[149] Roger Horn and Charles Johnson. *Topics in matrix analysis*. Cambridge University Presss, Cambridge, 1991.

[150] Roger Horn and Charles Johnson. *Matrix analysis*. 2nd ed. Cambridge University Press, Cambridge, 2013. ISBN: 978-0-521-54823-6.

[151] Jonathan Hosking. "Lagrange-multiplier tests of time-series models." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 42.2 (1980), pp. 170–181.

[152] Harold Hotelling. "Relations Between Two Sets of Variates." In: *Biometrika* 28.3/4 (1936), pp. 321–377. ISSN: 00063444. URL: http://www.jstor.org/stable/2333955 (visited on 04/26/2022).

[153] Hsiang-Ling Hsu, Ching-Kang Ing, and Howell Tong. "On model selection from a finite family of possibly misspecified time series models." In: *The Annals of Statistics* 47.2 (2019), pp. 1061–1087. ISSN: 0090-5364. DOI: 10.1214/18-AOS1706.

[154] Hsiang-Ling Hsu, Ching-Kang Ing, and Howell Tong. "Supplement to "On model selection from a finite family of possibly misspecified time series models"." In: *The Annals of Statistics* 47.2 (2019), pp. 1061–1087.

[155] Peter Huber. "The behavior of maximum likelihood estimates under nonstandard conditions." In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 1. University of California Press. 1967, pp. 221–233.

[156] Peter Huber and Elvezio Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2009.

[157] Clifford Hurvich and Chih-Ling Tsai. "Regression and time series model selection in small samples." In: *Biometrika* 76.2 (1989), pp. 297–307.

[158] Clifford Hurvich and Chih-Ling Tsai. "A corrected Akaike information criterion for vector autoregressive model selection." In: *Journal of Time Series Analysis* 14.3 (1993), pp. 271–279.

[159] Rob Hyndman and Anne Koehler. "Another look at measures of forecast accuracy." In: *International Journal of Forecasting* 22.4 (2006), pp. 679–688.

[160] Seiya Imoto and Sadanori Konishi. "Selection of smoothing parameters in B-spline nonparametric regression models using information criteria." In: *Annals of the Institute of Statistical Mathematics* 55.4 (2003), pp. 671–687.

[161] Ching-Kang Ing and Tze Leung Lai. "A stepwise regression method and consistent model selection for high-dimensional sparse linear models." In: *Statistica Sinica* (2011), pp. 1473–1513.

[162] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[163] Jiming Jiang. *Large sample techniques for statistics*. Springer Texts in Statistics. Springer, New York, 2010. ISBN: 978-1-4419-6826-5. DOI: 10.1007/978-1-4419-6827-2. URL: https://doi-org.ezproxy.unibo.it/10.1007/978-1-4419-6827-2.

[164] Joseph Kadane and Nicole Lazar. "Methods and criteria for model selection." In: *Journal of the American Statistical Association* 99.465 (2004), pp. 279–290.

[165] George Kapetanios. "The asymptotic distribution of the cointegration rank estimator under the Akaike information criterion." In: *Econometric Theory* (2004), pp. 735–742.

[166] Christian Kascha and Carsten Trenkler. "Simple identification and specification of cointegrated VARMA models." In: *Journal of Applied Econometrics* 30.4 (2015), pp. 675–702.

[167] Rangasami Kashyap. "A Bayesian comparison of different classes of dynamic models using empirical data." In: *IEEE Transactions on Automatic Control* 22.5 (1977), pp. 715–727.

[168] Rangasami Kashyap. "Optimal choice of AR and MA parts in autoregressive moving average models." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (1982), pp. 99–104.

[169] Lutz Kilian and Helmut Lütkepohl. *Structural vector autoregressive analysis*. Cambridge University Press, 2017.

[170] Genshiro Kitagawa. "On the use of AIC for the detection of outliers." In: *Technometrics* 21.2 (1979), pp. 193–199.

[171] Genshiro Kitagawa and Sadanori Konishi. "Statistical Model Evaluation by Generalized Information Criteria | Bias and Variance Reduction Techniques." In: *52nd ISI Session*. 1999.

[172] Martin Klein. "Gibbs on Clausius." In: *Historical Studies in the Physical Sciences* 1 (1969), pp. 127–149.

[173] Sadanori Konishi and Genshiro Kitagawa. "Generalised information criteria in model selection." In: *Biometrika* 83.4 (1996), pp. 875–890.

[174] Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[175] Solomon Kullback. *Information theory and statistics*. Dover Books on Mathematics. Dover Publications, 1978.

[176] Solomon Kullback. "Letter to the editor: The Kullback-Leibler distance." In: *American Statistician* 41 (4 1987), pp. 338–341.

[177] Solomon Kullback and Richard Leibler. "On Information and Sufficiency." In: *Annals of Mathematical Statistics* 22.1 (Mar. 1951), pp. 79–86. DOI: 10.1214/aoms/1177729694. URL: https://doi.org/10.1214/aoms/1177729694.

[178] Naoto Kunitomo and Taku Yamamoto. "Properties of predictors in misspecified autoregressive time series models." In: *Journal of the American Statistical Association* 80.392 (1985), pp. 941–950.

[179] Michele La Rocca and Cira Perna. "Model selection for neural network models: a statistical perspective." In: *Computational Network Theory: Theoretical Foundations and Applications* (2015).

[180] Tze Leung Lai and Chang Ping Lee. "Information and prediction criteria for model selection in stochastic regression and ARMA models." In: *Statistica Sinica* (1997), pp. 285–309.

[181] Tze Leung Lai and Hongsong Yuan. "Stochastic Approximation: From Statistical Origin to Big-Data, Multidisciplinary Applications." In: *Statistical Science* 36.2 (2021), pp. 291–302.

[182] Peter Lancaster and Hanafi Farahat. "Norms on direct sums and tensor products." In: *Mathematics of Computation* 26 (118 1972), pp. 401–414.

[183] Jean-Dominique Lebreton, Kenneth Burnham, Jean Clobert, and David Anderson. "Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies." In: *Ecological Monographs* 62.1 (1992), pp. 67–118.

[184] Eun Ryung Lee, Jinwoo Cho, and Kyusang Yu. "A systematic review on model selection in high-dimensional regression." In: *Journal of the Korean Statistical Society* 48.1 (2019), pp. 1–12.

[185] Youngjo Lee and Jan Bjørnstad. "Extended likelihood approach to large-scale multiple testing." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2013), pp. 553–575.

[186] Hannes Leeb. "Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process." In: *Bernoulli* 14.3 (2008), pp. 661–690.

[187] Hannes Leeb and Benedikt Pötscher. "Model selection." In: *Handbook of Financial Time Series*. Springer, 2009, pp. 889–925.

[188] Hannes Leeb, Benedikt Pötscher, and Karl Ewald. "On various confidence intervals post-model-selection." In: *Statistical Science* 30.2 (2015), pp. 216–227.

[189] Ker-Chau Li. "Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set." In: *The Annals of Statistics* (1987), pp. 958–975.

[190] Ker-Chau Li. "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set." In: *Ann. Statist.* 15.3 (1987), pp. 958–975. ISSN: 0090-5364. DOI: 10.1214/aos/1176350486. URL: https://doi-org.ezproxy.unibo.it/10.1214/aos/1176350486.

[191] Qi Li and Jeffrey Scott Racine. *Nonparametric econometric methods*. Emerald Group Publishing, 2009.

[192] Heinz Linhart and Walter Zucchini. *Model selection*. John Wiley & Sons, 1986.

[193] Xialu Liu and Rong Chen. "Threshold factor models for high-dimensional time series." In: *Journal of Econometrics* 216.1 (2020), pp. 53–70.

[194] Lennart Ljung. *System identification: Theory for the user*. 2nd ed. Prentice-Hall, Upper Saddle River, NJ, 1999.

[195] Jan Lohmeyer, Franz Palm, Hanno Reuvers, and Jean-Pierre Urbain. "Focused information criterion for locally misspecified vector autoregressive models." In: *Econometric Reviews* 38.7 (2019), pp. 763–792.

[196] James P Long. "A note on parameter estimation for misspecified regression models with heteroskedastic errors." In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1464–1490.

[197] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

[198] Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer-Verlag, Berlin, 2005. ISBN: 3-540-40172-5. DOI: 10.1007/978-3-540-27752-1. URL: https://doi-org.ezproxy.unibo.it/10.1007/978-3-540-27752-1.

[199] Helmut Lütkepohl and Aleksei Netšunajev. "Structural vector autoregressions with heteroskedasticity: A review of different volatility models." In: *Econometrics and Statistics* 1 (2017), pp. 2–18.

[200] Jinchi Lv and Jun Liu. "Model selection principles in misspecified models." In: *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2014), pp. 141–167.

[201] David MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[202] Robert Engel Machol and Paul Gray, eds. *Recent developments in information and decision processes*. MacMillan, New York, 1962.

[203] Colin Mallows. "Choosing a Subset Regression." In: *Unpublished report* (1967).

[204] Colin Mallows. "Some remarks of Cp." In: *Technometrics* 15 (1973), pp. 661–675.

[205] Sebastiano Manzan. "Model selection for nonlinear time series." In: *Empirical Economics* 29.4 (2004), pp. 901–920.

[206] Pesi Masani and Norbert Wiener. "On bivariate stationary processes and the factorization of matrix-valued functions." In: *Theory of Probability & Its Applications* 4.3 (1959), pp. 300–308.

[207] Allan McQuarrie and Chih-Ling Tsai. *Regression and time series model selection*. World Scientific, 1998.

[208] Guy Mélard. "An indirect proof for the asymptotic properties of VARMA model estimators." In: *Econometrics and Statistics* 21 (2022), pp. 96–111.

[209] Edwin Moise. *Introductory Problem Courses in Analysis and Topology*. Springer-Verlag, New York, 1982.

[210] Arthur Munson and Rich Caruana. "On feature selection, bias-variance, and bagging." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2009, pp. 144–159.

[211] Elizbar Nadaraya. "On estimating regression." In: *Theory of Probability & Its Applications* 9.1 (1964), pp. 141–142.

[212] Naveen Naidu Narisetty. "Bayesian model selection for high-dimensional data." In: *Handbook of Statistics*. Vol. 43. Elsevier, 2020, pp. 207–248.

[213] Serena Ng and Pierre Perron. "A note on the selection of time series models." In: *Oxford Bulletin of Economics and Statistics* 67.1 (2005), pp. 115–134.

[214] Ryuei Nishii. "Asymptotic properties of criteria for selection of variables in multiple regression." In: *The Annals of Statistics* (1984), pp. 758–765.

[215] Cuizhen Niu, Xu Guo, and Lixing Zhu. "Enhancements of Nonparametric Generalized Likelihood Ratio Test: Bias Correction and Dimension Reduction." In: *Scandinavian Journal of Statistics* 45.2 (2018), pp. 217–254.

[216] Tinsley Oden and Leszek Demkowicz. *Applied functional analysis*. Textbooks in Mathematics. Third edition of [MR1384069]. CRC Press, Boca Raton, FL, 2018. ISBN: 978-1-4987-6114-7.

[217] Yosihiko Ogata. "Maximum Likelihood Estimates of Incorrect Markov Models for Time Series and the Derivation of AIC." In: *Journal of Applied Probability* 17.1 (1980), pp. 59–72. ISSN: 00219002. URL: http://www.jstor.org/stable/3212924.

[218] Arash Owrang and Magnus Jansson. "A model selection criterion for high-dimensional linear regression." In: *IEEE Transactions on Signal Processing* 66.13 (2018), pp. 3436–3446.

[219] Emanuel Parzen. "Some recent advances in time series modeling." In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 723–730.

[220] Daniel Peña and Ismael Sánchez. "Multifold predictive validation in ARMAX time series models." In: *Journal of the American Statistical Association* 100.469 (2005), pp. 135–146.

[221] Luis Raúl Pericchi. "Model selection and hypothesis testing based on objective probabilities and Bayes factors." In: *Handbook of Statistics*. Vol. 25. Elsevier, 2005, pp. 115–149.

[222] Peter Phillips. "Econometric model determination." In: *Econometrica: Journal of the Econometric Society* (1996), pp. 763–812.

[223] Donald Poskitt. "A modified Hannan—Rissanen strategy for mixed autoregressive-moving average order determination." In: *Biometrika* 74.4 (1987), pp. 781–790.

[224] Donald Poskitt. "Precision, complexity and Bayesian model determination." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 49.2 (1987), pp. 199–208.

[225] Donald Poskitt. "On the specification of cointegrated autoregressive moving-average forecasting systems." In: *International Journal of Forecasting* 19.3 (2003), pp. 503–519.

[226] Donald Poskitt. "Vector autoregressive moving average identification for macroeconomic modeling: A new methodology." In: *Journal of Econometrics* 192.2 (2016), pp. 468–484.

[227] Benedikt Potscher. "Model selection under nonstationarity: Autoregressive models and stochastic linear regression models." In: *The Annals of Statistics* (1989), pp. 1257–1274.

[228] Zhongjun Qu and Pierre Perron. "A modified information criterion for cointegration tests based on a VAR approximation." In: *Econometric Theory* 23.4 (2007), pp. 638–685.

[229] Maurice Quenouille. "Note on the elimination of insignificant variates in discriminatory analysis." In: *Annals of Eugenics* 14.1 (1947), pp. 305–308.

[230] Barry Quinn. "Order determination for a multivariate autoregression." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 42.2 (1980), pp. 182–185.

[231] Jeffrey Racine, Liangjun Su, and Aman Ullah. *The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics*. Oxford University Press, 2013.

[232] Arni Srinivasa Rao, Calyampudi Radhakrishna Rao, and Angelo Plastino, eds. *Information Geometry*. Vol. 45. Handbook of Statistics. 2021.

[233] Calyampudi Radhakrishna Rao. "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation." In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 44. 1. Cambridge University Press. 1948, pp. 50–57.

[234] Calyampudi Radhakrishna Rao, ed. *Bayesian thinking, modeling and computation*. Vol. 25. Handbook of Statistics. 2005.

[235] Calyampudi Radhakrishna Rao and Yuehua Wu. "A strongly consistent procedure for model selection in a regression problem." In: *Biometrika* 76.2 (1989), pp. 369–374.

[236] Calyampudi Radhakrishna Rao, Yuehua Wu, Sadanori Konishi, and Rahul Mukerjee. "On model selection." In: *Institute of Mathematical Statistics, Lecture Notes - Monograph Series*. Ed. by Partha Lahiri. Institute of Mathematical Statistics, 2001, pp. 1–64.

[237] Gregory Reinsel. *Elements of multivariate time series analysis*. Springer Series in Statistics. Springer - Verlag, New York, 1993. ISBN: 0-387-94063-4. DOI: 10.1007/978-1-4684-0198-1.

[238] Gregory Reinsel. *Multivariate Time Series Analysis*. John Wiley & Sons New York, NY, 1993.

[239] Yunwen Ren and Xinsheng Zhang. "Subset selection for vector autoregressive processes via adaptive Lasso." In: *Statistics & Probability Letters* 80.23-24 (2010), pp. 1705–1712.

[240] Pietro Rigo. "Basi probabilistiche dell'inferenza statistica." Unpublished. PhD Lectures notes, University of Bologna. Bologna, 2020.

[241] Jorma Rissanen. "Modeling by shortest data description." In: *Automatica* 14.5 (1978), pp. 465–471.

[242] Jorma Rissanen. "Modeling by shortest data description." In: *Automatica* 14 (5 1978), pp. 465–471. ISSN: 0005-1098. DOI: https://doi.org/10.1016/0005-1098(78)90005-5.

[243] Jorma Rissanen. "Order estimation by accumulated prediction errors." In: *Journal of Applied Probability* (1986), pp. 55–61.

[244] Jorma Rissanen. *Stochastic complexity in statistical inquiry*. Vol. 15. World Scientific Series in Computer Science. World Scientific, 1989. ISBN: 9971-50-859-1.

[245] Jorma Rissanen. *Information and complexity in statistical modeling*. Springer Science & Business Media, 2007.

[246] Juan Peña Sanchez de Rivera. "Procedimientos de selección para modelos no lineales de series temporales." In: *Estadística Española* 107 (1985), pp. 39–53.

[247] Peter Robinson. "Nonparametric estimators for time series." In: *Journal of Time Series Analysis* 4.3 (1983), pp. 185–207.

[248] Peter Robinson. "Asymptotic theory for nonparametric regression with spatial data." In: *Journal of Econometrics* 165.1 (2011), pp. 5–19.

[249] Nikolay Robinzonov, Gerhard Tutz, and Torsten Hothorn. "Boosting techniques for nonlinear time series models." In: *AStA Advances in Statistical Analysis* 96.1 (2012), pp. 99–122.

[250] Joseph Romano and Michael Wolf. "Balanced control of generalized error rates." In: *The Annals of Statistics* 38.1 (2010), pp. 598–633.

[251] Anindya Roy, Tucker McElroy, and Peter Linton. "Constrained estimation of causal invertible VARMA." In: *Statistica Sinica* 29.1 (2019), pp. 455–478.

[252] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. "Detecting and quantifying causal associations in large nonlinear time series datasets." In: *Science Advances* 5.11 (2019), eaau4996.

[253] Yosiyuki Sakamoto, Makio Ishiguro, and Genshiro Kitagawa. *Akaike information criterion statistics*. Vol. 81. Taylor & Francis, Dordrecht, The Netherlands, 1986.

[254] Takamitsu Sawa. "Information criteria for discriminating among alternative regression models." In: *Econometrica: Journal of the Econometric Society* (1978), pp. 1273–1291.

[255] Mark Schmidt. "Least squares optimization with L1-norm regularization." In: *CS542B Project Report* 504 (2005), pp. 195–221.

[256] Frank Schorfheide. "VAR forecasting under misspecification." In: *Journal of Econometrics* 128.1 (2005), pp. 99–136.

[257] Arthur Schuster. "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena." In: *Terrestrial Magnetism* 3.1 (1898), pp. 13–41.

[258] Gideon Schwarz. "Estimating the dimension of a model." In: *The Annals of Statistics* 6 (2 1978), pp. 461–464. ISSN: 0090-5364. URL: http://links.jstor.org.ezproxy.unibo.it/sici?sici=0090-5364(197803)6:2<461:ETDOAM>2.0.CO;2-5&origin=MSN.

[259] Gideon Schwarz. "Estimating the dimension of a model." In: *The Annals of Statistics* 6.2 (1978), pp. 461–464.

[260] George A. F. Seber. *A matrix handbook for statisticians*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2008. ISBN: 978-0-471-74869-4.

[261] Abd-Krim Seghouane and Maiza Bekara. "A small sample model selection criterion based on Kullback's symmetric divergence." In: *IEEE Transactions on Signal Processing* 52.12 (2004), pp. 3314–3323.

[262] Claude Shannon. "A mathematical theory of communication." In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423.

[263] Claude Shannon and Warren Weaver. *The mathematical theory of communication*. University of Illinois Press, 1949.

[264] Jun Shao. "Linear model selection by cross-validation." In: *Journal of the American Statistical Association* 88.422 (1993), pp. 486–494.

[265] Jun Shao. "Bootstrap model selection." In: *Journal of the American Statistical Association* 91.434 (1996), pp. 655–665.

[266] Jun Shao. "An asymptotic theory for linear model selection." In: *Statistica Sinica* (1997), pp. 221–242.

[267] Jun Shao. "An asymptotic theory for linear model selection." In: *Statistica Sinica* 7 (2 1997). With comments and a rejoinder by the author, pp. 221–264. ISSN: 1017-0405.

[268] Alexander Shapiro. "On concepts of directional differentiability." In: *Journal of Optimization Theory and Applications* 66.3 (1990), pp. 477–487.

[269] Peide Shi and Chih-Ling Tsai. "Regression model selection—a residual likelihood approach." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.2 (2002), pp. 237–252.

[270] Ritei Shibata. "Selection of the order of an autoregressive model by Akaike's Information Criterion." In: *Biometrika* 63.1 (1976), pp. 117–126.

[271] Ritei Shibata. "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process." In: *The Annals of Statistics* (1980), pp. 147–164.

[272] Ritei Shibata. "Asymptotically efficient selection of the order of the model for estimating parameters of a linear process." In: *The Annals of Statistics* 8.1 (1980), pp. 147–164. ISSN: 0090-5364. URL: http://links.jstor.org.ezproxy.unibo.it/sici?sici=0090-5364(198001)8:1<147:AESOTO>2.0.CO;2-N&origin=MSN.

[273] Ritei Shibata. "An optimal selection of regression variables." In: *Biometrika* 68.1 (1981), pp. 45–54.

[274] Ritei Shibata. "Asymptotic mean efficiency of a selection of regression variables." In: *Annals of the Institute of Statistical Mathematics* 35.3 (1983), pp. 415–423.

[275] Ritei Shibata. "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables." In: *Biometrika* 71.1 (1984), pp. 43–49. ISSN: 00063444. URL: http://www.jstor.org/stable/2336395.

[276] Ritei Shibata. "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables." In: *Biometrika* 71.1 (1984), pp. 43–49. ISSN: 00063444. URL: http://www.jstor.org/stable/2336395.

[277] Ritei Shibata. "Statistical Aspects of Model Selection." In: *From Data to Model*. Ed. by Jan Willems. Berlin, Heidelberg: Springer Berlin, Heidelberg, 1989, pp. 215–240. ISBN: 978-3-642-75007-6. DOI: 10.1007/978-3-642-75007-6_5. URL: https://doi.org/10.1007/978-3-642-75007-6_5.

[278] Ritei Shibata. "Statistical aspects of model selection." In: *From data to model*. Ed. by Jan Willems. Springer, 1989, pp. 215–240.

[279] Galit Shmueli. "To explain or to predict?" In: *Statistical Science* 25.3 (2010), pp. 289–310.

[280] Chor-Yiu Sin and Halbert White. "Information criteria for selecting possibly misspecified parametric models." In: *Journal of Econometrics* 71.1-2 (1996), pp. 207–225.

[281] Jaime Lynn Speiser, Michael Miller, Janet Tooze, and Edward Ip. "A comparison of random forest variable selection methods for classification prediction modeling." In: *Expert Systems with Applications* 134 (2019), pp. 93–101.

[282] David Spiegelhalter, Nicola Best, Bradley Carlin, and Angelika Van der Linde. "The deviance information criterion: 12 years on." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.3 (2014), pp. 485–493.

[283] David Spiegelhalter, Nicola Best, Bradley Carlin, and Angelika Van Der Linde. "Bayesian measures of model complexity and fit." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4 (2002), pp. 583–639.

[284] Charles Stone. "Local asymptotic admissibility of a generalization of Akaike's model selection rule." In: *Annals of the Institute of Statistical Mathematics* 34.1 (1982), pp. 123–133.

[285] Mervyn Stone. "Cross-validatory choice and assessment of statistical predictions." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 36.2 (1974), pp. 111–133.

[286] Mervyn Stone. "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 39.1 (1977), pp. 44–47.

[287] Mervyn Stone. "Asymptotics for and against cross-validation." In: *Biometrika* (1977), pp. 29–35.

[288] William Stout. "The Hartman-Wintner law of the iterated logarithm for martingales." In: *The Annals of Mathematical Statistics* 41.6 (1970), pp. 2158–2160.

[289] Luis Enrique Sucar. *Probabilistic graphical models*. Advances in Computer Vision and Pattern Recognition. Springer Nature Switzerland, 2021.

[290] Nariaki Sugiura. "Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by Akaike's." In: *Communications in Statistics-Theory and Methods* 7.1 (1978), pp. 13–26.

[291] Yoshikazu Takada. "Admissibility of some variable selection rules in linear regression model." In: *Journal of the Japan Statistical Society, Japanese Issue* 12.1 (1982), pp. 45–49.

[292] Kei Takeuchi. "Distribution of information statistic and validity criterion of models." In: *Mathematical Science* 153 (1976), pp. 12–18.

[293] Timo Teräsvirta and Ilkka Mellin. "Model selection criteria and model selection tests in regression models." In: *Scandinavian Journal of Statistics* (1986), pp. 159–171.

[294] Timo Teräsvirta, Dag Tjøstheim, and Clive Granger. *Modelling nonlinear economic time series*. Oxford University Press, Oxford, 2010.

[295] Mary Thompson. "Selection of variables in multiple regression: Part I. A review and evaluation." In: *International Statistical Review/Revue Internationale de Statistique* (1978), pp. 1–19.

[296] Mary Thompson. "Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples." In: *International Statistical Review/Revue Internationale de Statistique* (1978), pp. 129–146.

[297] George Tiao and Ruey Tsay. "Model specification in multivariate time series." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 51.2 (1989), pp. 157–195.

[298] Robert Tibshirani. "Regression shrinkage and selection via the lasso." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58.1 (1996), pp. 267–288.

[299] Robert Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3 (2011), pp. 273–282.

[300] Dag Tjøstheim and Bjørn Auestad. "Nonparametric identification of nonlinear time series: projections." In: *Journal of the American Statistical Association* 89.428 (1994), pp. 1398–1409.

[301] Howell Tong. "Determination of the order of a Markov chain by Akaike's information criterion." In: *Journal of Applied Probability* 12.3 (1975), pp. 488–497.

[302] Howell Tong. "A note on a local equivalence of two recent approaches to autoregressive order determination." In: *International Journal of Control* 29.3 (1979), pp. 441–446.

[303] Howell Tong. *Threshold models in non-linear time series analysis*. Vol. 21. Lecture Notes in Statistics. Springer-Verlag New York, 1983.

[304] Howell Tong. *Non-linear time series: a dynamical system approach*. Oxford University Press, 1990.

[305] Ruey Tsay. "Two canonical forms for vector ARMA processes." In: *Statistica Sinica* (1991), pp. 247–269.

[306] Ruey Tsay. *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, 2013.

[307] Ruey Tsay. *Multivariate time series analysis*. Wiley Series in Probability and Statistics. With R and financial applications. John Wiley & Sons, Hoboken, NJ, 2014. ISBN: 978-1-118-61790-8.

[308] Ruey Tsay and Rong Chen. *Nonlinear time series analysis*. Vol. 891. John Wiley & Sons, 2019.

[309] Rolf Tschernig. "Nonparametric time series modelling." In: *Applied Time Series Econometrics*. Ed. by Helmut Lütkepohl, Markus Krätzig, and Peter Phillips. Cambridge University Press, 2004, pp. 243–288.

[310] Rolf Tschernig and Lijian Yang. "Nonparametric lag selection for time series." In: *Journal of Time Series Analysis* 21.4 (2000), pp. 457–487.

[311] John Tukey. "Discussion, emphasizing the connection between analysis of variance and spectrum analysis." In: *Technometrics* 3.2 (1961), pp. 191–219.

[312] John Tukey. "Discussion of 'Topics in the investigation of linear relations fitted by the method of least squares' by F.J. Anscombe." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 29 (1 1967-01), pp. 47–48. ISSN: 00359246.

[313] American University. *Recent Advances in Info-Metrics Research*. 2021. URL: https://www.american.edu/cas/economics/info-metrics/workshop/.

[314] Tim Van Erven, Peter Grünwald, and Steven De Rooij. "Catching up faster by switching sooner: a predictive approach to adaptive estimation with an application to the AIC-BIC dilemma [with Discussion]." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* (2012), pp. 361–417.

[315] Cristiano Varin and Paolo Vidoni. "A note on composite likelihood inference and model selection." In: *Biometrika* 92.3 (2005), pp. 519–528.

[316] Philippe Vieu. "Choice of regressors in nonparametric estimation." In: *Computational Statistics & Data Analysis* 17.5 (1994), pp. 575–594.

[317] Philippe Vieu. "Order choice in nonlinear autoregressive models." In: *Statistics: A Journal of Theoretical and Applied Statistics* 26.4 (1995), pp. 307–328.

[318] Quang Vuong. "Likelihood ratio tests for model selection and non-nested hypotheses." In: *Econometrica: Journal of the Econometric Society* (1989), pp. 307–333.

[319] Abraham Wald. "Tests of statistical hypotheses concerning several parameters when the number of observations is large." In: *Transactions of the American Mathematical Society* 54.3 (1943), pp. 426–482.

[320] Ari Ezra Waldman. "Power, process, and automated decision-making." In: *Fordham Law Review* 88 (2019), p. 613.

[321] Larry Wasserman. "Bayesian model selection and model averaging." In: *Journal of Mathematical Psychology* 44.1 (2000), pp. 92–107.

[322] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[323] Satosi Watanabe. *Knowing and Guessing a Quantitative Study of Inference and Information*. John Wiley & Sons, 1969.

[324] Sumio Watanabe. *Algebraic geometry and statistical learning theory*. 25. Cambridge University Press, 2009.

[325] Sumio Watanabe. "A widely applicable Bayesian information criterion." In: *Journal of Machine Learning Research* 14 (2013), pp. 867–897.

[326] Sumio Watanabe and Manfred Opper. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." In: *Journal of Machine Learning Research* 11.12 (2010).

[327] Geoffrey Watson. "Smooth regression analysis." In: *Sankhyā: The Indian Journal of Statistics, Series A* (1964), pp. 359–372.

[328] Ching-Zong Wei. "On predictive least squares principles." In: *The Annals of Statistics* (1992), pp. 1–42.

[329] William Wei. *Multivariate time series analysis and applications*. John Wiley & Sons, 2018.

[330] Halbert White. "Maximum likelihood estimation of misspecified models." In: *Econometrica: Journal of the Econometric Society* (1982), pp. 1–25.

[331] Peter Whittle. *Hypothesis testing in time series analysis*. Vol. 4. Almqvist & Wiksells boktr., 1951.

[332] Peter Whittle. "The analysis of multiple stationary time series." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 15.1 (1953), pp. 125–139.

[333] Norbert Wiener. "The mathematics of self-organising systems." In: *Recent Developments in Information and Decision Processes*. Ed. by Robert E. Machol and Paul Gray. Macmillan, New York, 1962, pp. 1–22.

[334] Norbert Wiener and Pesi Masani. "The prediction theory of multivariate stochastic processes." In: *Acta Mathematica* 98.1-4 (1957), pp. 111–150.

[335] Norbert Wiener and Pesi Masani. "The prediction theory of multivariate stochastic processes, II." In: *Acta Mathematica* 99.1 (1958), pp. 93–137.

[336] Elin Wihlborg, Hannu Larsson, and Karin Hedström. ""' The Computer Says No!"–A Case Study on Automated Decision-Making in Public Authorities." In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. IEEE. 2016, pp. 2903–2912.

[337] Samuel Wilks. "The large-sample distribution of the likelihood ratio for testing composite hypotheses." In: *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62.

[338] Ines Wilms, Sumanta Basu, Jacob Bien, and David Matteson. "Sparse identification and estimation of large-scale vector autoregressive moving averages." In: *Journal of the American Statistical Association* (2021), pp. 1–12.

[339] Herman Wold. "A large-sample test for moving averages." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 11.2 (1949), pp. 297–305.

[340] Chun-Shan Wong and Wai Keung Li. "A note on the corrected Akaike information criterion for threshold autoregressive models." In: *Journal of Time Series Analysis* 19.1 (1998), pp. 113–124.

[341] Tiee-Jian Wu and Alfred Sepulveda. "The weighted average information criterion for order selection in time series and regression models." In: *Statistics & Probability Letters* 39.1 (1998), pp. 1–10.

[342] Zongben Xu, Hai Zhang, Yao Wang, XiangYu Chang, and Yong Liang. "$L_1/2$ regularization." In: *Science China Information Sciences* 53.6 (2010), pp. 1159–1169.

[343] Yishu Xue and Guanyu Hu. "Online updating of information based model selection in the big data setting." In: *Communications in Statistics – Simulation and Computation* (2019), pp. 1–14.

[344] Hirokazu Yanagihara and Chihiro Ohmoto. "On distribution of AIC in linear regression models." In: *Journal of Statistical Planning and Inference* 133.2 (2005), pp. 417–433.

[345] Hirokazu Yanagihara, Risa Sekiguchi, and Yasunori Fujikoshi. "Bias correction of AIC in logistic regression models." In: *Journal of Statistical Planning and Inference* 115.2 (2003), pp. 349–360.

[346] Yuhong Yang. "Comparing learning methods for classification." In: *Statistica Sinica* (2006), pp. 635–657.

[347] Yuhong Yang. "Consistency of cross validation for comparing regression procedures." In: *The Annals of Statistics* 35.6 (2007), pp. 2450–2473.

[348] Yuhong Yang. "Prediction/estimation with simple linear models: Is it really that simple?" In: *Econometric Theory* (2007), pp. 1–36.

[349] Yuhong Yang and Andrew Barron. "An asymptotic property of model selection criteria." In: *IEEE Transactions on Information Theory* 44.1 (1998), pp. 95–116.

[350] Qiwei Yao and Howell Tong. "On subset selection in non- parametric stochastic regression." In: *Statistica Sinica* (1994), pp. 51–70.

[351] Qiwei Yao and Howell Tong. "Cross-validatory bandwidth selections for regression estimation based on dependent data." In: *Journal of Statistical Planning and Inference* 68.2 (1998), pp. 387–415.

[352] Yuanxiang Ying, Juanyong Duan, Chunlei Wang, Yujing Wang, Congrui Huang, and Bixiong Xu. "Automated Model Selection for Time-Series Anomaly Detection." In: *arXiv preprint arXiv:2009.04395* (2020).

[353] Arnold Zellner and Franz Palm. "Time series analysis and simultaneous equation econometric models." In: *Journal of Econometrics* 2.1 (1974), pp. 17–54.

[354] Hu-Ming Zhang and Ping Wang. "A new way to estimate orders in time series." In: *Journal of Time Series Analysis* 15.5 (1994), pp. 545–559.

[355] Ting Zhang and Wei Biao Wu. "Time-varying nonlinear regression models: nonparametric estimation and model selection." In: *The Annals of Statistics* 43.2 (2015), pp. 741–768.

[356] Lin Cheng Zhao, Paruchuri Krishnaiah, and Zhidong Bai. "On detection of the number of signals in presence of white noise." In: *Journal of Multivariate Analysis* 20.1 (1986), pp. 1–25.

[357] Peng Zhao and Bin Yu. "On model selection consistency of Lasso." In: *The Journal of Machine Learning Research* 7 (2006), pp. 2541–2563.

[358] Zhou Zhou. "Nonparametric specification for non-stationary time series regression." In: *Bernoulli* 20.1 (2014), pp. 78–108.

[359] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net." In: *Journal of the Royal Statistical Society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[360] Hui Zou and Hao Helen Zhang. "On the adaptive elastic-net with a diverging number of parameters." In: *The Annals of Statistics* 37.4 (2009), p. 1733.

[361] Jan deLeeuw. "Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle." In: *Breakthroughs in Statistics*. Springer, 1992, pp. 599–609.

DECLARATION

I, Gery Andrés DÍAZ RUBIO, declare that this thesis titled, "Model Selection and the Vectorial Misspecification-Resistant Information Criterion in Multivariate Time Series " and the work presented in it are my own. I confirm that:

- This work was done wholly and mainly while in candidature for a research degree at this University.
- No part of this thesis was previously been submitted for any degree or any other qualification at this University or any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Bologna, 30 September 2022

Gery Andrés Díaz Rubio

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Final Version as of September 30, 2022 (`classicthesis`).