A Time-reversed Model Identification Approach to Time Series Forecasting

M.W. Sibeijn

Master of Science Thesis

TU Delft Delft University of Technology

Delft Center for Systems and Control

A Time-reversed Model Identification Approach to Time Series Forecasting

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

M.W. Sibeijn

September 16, 2021

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

Abstract

In this thesis, I introduce a novel model identification approach to time series forecasting. For linear stationary processes, such as AR processes, the direction of time is independent of the model parameters. By combining theoretical principles of time-reversibility in time series with conventional modeling approaches such as information criteria, I design a criterion that employs the backwards prediction (backcast) as a proxy for the forecast. Hereby, I aim to adopt a finite sample, data-driven approach to model selection. The novel criterion is named the *backwards validated information criterion (BVIC)*. The BVIC quantifies forecasting performance of models by trading off a measure of goodness-of-fit and a models ability to predict backwards. I test the performance of the BVIC by conducting experiments on synthetic and real data. In each experiment, the BVIC is examined in contrast to conventionally employed criteria.

Experimental results suggest that the BVIC has comparable performance to conventional information criteria. Specifically, in most of the experiments performed, I did not find statistically significant differences between the forecast error of the BVIC under certain parameterizations and that of the different information criteria. Nonetheless, it is worth emphasizing that the BVIC guarantees are established by design where the model order penalization term depends on strong mathematical properties of time-reversibility and a finite data assessment. In particular, the penalization term is replaced by a weighted trade-off between functional dimensions pertaining to forecasting.

That said, I observed that the BVIC recovered more accurately the real order of the underlying process than the other criteria, which rely on a static penalization of the model order. Lastly, leveraging the latter property, I perform the assessment of the order model (or, memory) of time series pertaining to epileptic seizures recorded using electrocorticographic data. The results provide converging evidence that the order of the model increases during the epileptic events.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This document concludes my MSc Systems & Control at TU Delft. I would like to thank my supervisor dr.ir. Sérgio Pequito for his assistance during the writing of this thesis. Over the past year, Sérgio has continually provided his guidance during weekly meetings and I greatly appreciate the time he dedicated for this. I learned a lot from working together.

Delft, University of Technology
September 16, 2021

M.W. Sibeijn

Chapter 1

Introduction

Time series describe a multitude of real-life processes that evolve over time. For instance, consider yearly economic growth, a monthly account of the average temperature, a daily series of the number of new COVID-19 cases, and so on. Time series may be used to in fields such as economics, healthcare, engineering, natural sciences, and social sciences [1]. In time series analysis, one primarily seeks to determine the underlying rule (or mathematical model) that provides a plausible description for sampled data [2]. Among the different goals, we often want to forecast how the process is going to evolve within a specified horizon (i.e., up to h-steps ahead) [3].

Due to uncertainty, time series are generally considered as the realization of stochastic processes [4], which are often modeled as autoregressive moving-average (ARMA) processes, or their variations that accommodate seasonality (seasonal autoregressive moving-average (SARMA)), long-term memory with linear and power-law decaying autocorrelation (autoregressive integrated moving-average (ARIMA) and autoregressive fractionally integrated moving-average (ARFIMA), respectively). Nonetheless, while many generalizations and variations of the ARMA model exist, the majority of time series applications use ARMA for its simplicity and effectiveness [4]. In fact, ARMA models can be simplified even further as they are already a generalization of autoregressive (AR) models.

Usually, for autoregressive processes, the method of finding a suitable model is done by performing system identification. In particular, the Box-Jenkins methodology [1] is often used for this purpose. This method consists of three steps: (*i*) identification of the model order (*model identification*), (*ii*) estimation of the model parameters (*parameter estimation*), and (*iii*) statistical model checking (*model validation*) – see Figure 1-1. During model identification the Box-Jenkins method evaluates and ranks different models on the basis of certain mathematical criteria, computed from data. Hence, data is used for both estimation and evaluation. Therefore, one runs the risk of fitting models too closely to a particular set of data (*over-fitting*), which may result in a failure to make reliable predictions on additional data (*generalization*).

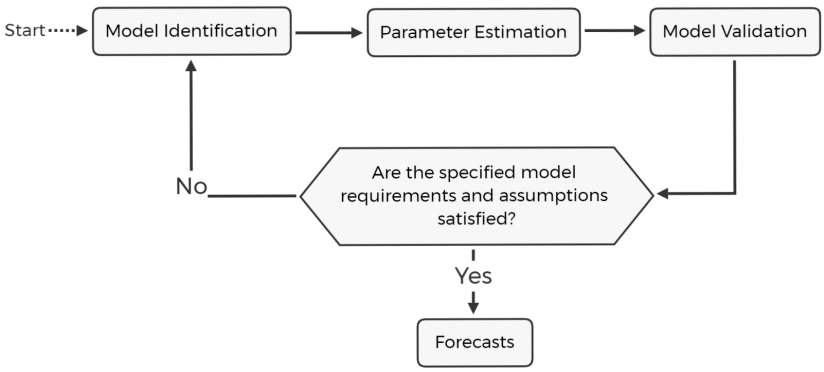


Figure 1-1: Box-Jenkins methodology. A schematic depiction of the steps required in identifying a system using the Box-Jenkins approach.

That said, model selection for time series forecasting generally involves the use of information criteria [1]. Information criteria are metrics to assess the discrepancy in information (i.e., the error) between a specified model and the true process [5]. Often, information criteria employ regression techniques to obtain the model parameters (e.g., least-squares, maximum likelihood). However, in addition to regression analysis, information criteria also penalize the model order in its objective function, with the aim of reducing over-fitting. Hereby, these methods often increase their ability to generalize outside their dataset, improving the reliability of forecasts.

Popular information criteria include Akaike's information criterion (AIC) [6], and Bayesian information criterion (BIC) [7], where the goodness-of-fit of a model is weighed against the model parsimony, but also cross-validation, where models are evaluated on their generalization capabilities by splitting the data into a training and validation set. In fact, it is also possible to extend this even further to leave-multiple-out cross-validation [8], yet in practice the improvement of performance often does not justify its use.

An implicit assumption to all previously discussed techniques is that of stationarity (i.e., a time series is said to be stationary if it has similar statistical properties to its time-shifted series) [4]. Real-life processes are almost never stationary in the strict sense, but weaker forms of stationarity exist (e.g. wide-sense stationarity), and may be used to analyse time series that satisfy this assumption. In some cases, such as with an (intracranial) electroencephalography recording, the stationarity assumption holds for limited periods of time [9]. Therefore, the available data is constrained to the duration of stationarity. Additionally, there are many other practical reasons that constrain the amount of data that can be obtained (e.g., time constraints, limited equipment).

As a result, in practice one is often required to work with small to medium sized samples of data. Here, information criteria are useful to reduce over-fitting in model selection. However, in small samples, the asymptotic properties of information criteria no longer hold [10], therefore, these criteria lack theoretical basis or proof to support their choice on how to penalize model complexity (hence, the large number of different criteria found in literature). Furthermore, extensive empirical studies [10, 11] argue that the predictive performance of information criteria depends largely on the type of data and its characteristics, and that not one criteria performs uniformly better than the others. Nonetheless, most existing information criteria

have a static penalty term (i.e., independent of data), except for criteria such as the empirical information criterion (EIC) [12].

Altogether, conventional information criteria (AIC, AICc and BIC) could leave room for improvement from a theoretical perspective and potentially also from an empirical perspective. In my thesis, I explore the possibility of extending the model selection framework by introducing a first principle analysis of a novel data-driven finite-sample method that combines aspects from information criteria and theoretical principles of time-reversibility in time series. Moreover, in favor of simplicity and generalizability, the scope of models I evaluate is set to autoregressive models. This offers a general basis for comparison of model selection techniques that may be extended in future research.

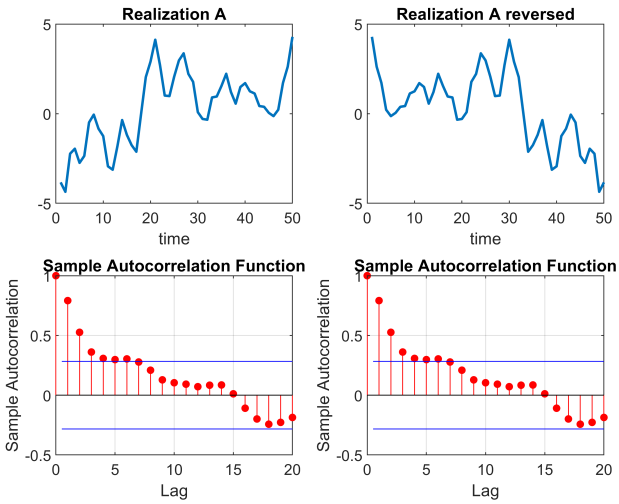


Figure 1-2: Time-reversibility. An example of how time-reversibility in time series works. A time series and its reverse are illustrated to demonstrate that the autocorrelation is not affected by the direction of time.

Specifically, for stationary time series that can be modeled as autoregressive processes, I present a novel model selection technique that employs time reversibility properties of autoregressive processes to validate the ability of a model to forecast by evaluating the backwards prediction. In Figure 1-2, an illustration of the concept of time-reversibility is shown. Considering past data is accessible, the technique aims to minimize prediction error and uncertainty in models for the backward prediction. Theoretically, I show that backward prediction achieves the same objective as forward prediction.

That said, the new method will be referred to as *backwards validated information criterion (BVIC)*. To perform model selection and determine the model parameters, I formulate an optimization problem that explicitly weighs three features: (a) regression, (b) generalization, and (c) uncertainty. The first component, regression, enables evaluation of the goodness-of-fit of a specified model with respect to the observed data, and can be quantified through maximum likelihood estimation or least-squares regression. Secondly, generalization represents a models ability to predict outside the sample of given observed data, and can be seen analogous to a measure of accuracy of prediction. Lastly, the uncertainty feature of the criterion can be considered as a level of precision exhibited by a prediction. That said, I quantify the features

of generalization and uncertainty through the metrics of mean square prediction error and theoretical prediction variance, respectively.

Consequently, to test the novel model selection criterion I conduct Monte Carlo simulations using experiments with both real and synthetic data. Here, the first two experiments consist in generating synthetic time series via specified autoregressive models with different model orders and sets of parameters. The third experiment assesses the different criteria when considering intracranial electroencephalographic data. In each experiment, I assess the quality of the BVIC, and other conventional information criteria, on the basis of a selection of performance metrics. Additionally, I evaluate the effects of noise on model selection by varying the noise variance throughout the experiments with synthetic data.

Significance. From a theoretical viewpoint I use time-reversibility to explicitly trade-off regression, generalization, uncertainty, and forecasting, formulated as an optimization problem. This approach is suitable for locally stationary processes and when limited data is available, which is opposite to the underlying assumptions in the established information criteria (e.g., AIC, AICc, BIC). It is remarkable to see that in practice the performance with respect to commonly used metrics is statistically indistinguishable, yet the behavior exhibited by the proposed criterion (i.e., BVIC) is rather distinct and allows uniform performance for different orders of the underlying model and equips the designer with the possibility to explicitly decide between accuracy and/or precision of the forecast, possibly within a specific window of time in the forecasting horizon, which is not possible with the aforementioned mainstream information criteria.

Structure. The thesis is divided into three main chapters. Firstly, I outline fundamental concepts and definitions in Chapter 2. Here, I discuss topics such as stationarity, regression, forecasting, the bias-variance trade-off, information criteria, and forecast metrics.

Secondly, in Chapter 3, I present the problem solution of my research. Here, I describe the methods and theory that directly contribute to the definition of the BVIC (i.e., time-reversibility and information criteria). Subsequently, I detail the components of the BVIC and I discuss the implications it has with respect to forecasting.

In Chapter 4, I discuss the experimental setup to assess the performance of the BVIC. Consecutively, I report the results of the three experiments that were conducted.

Finally, the thesis is concluded with a discussion on the implications of the BVIC for time series forecasting.

Chapter 2

Preliminaries

2-1 Time Series

In this section, an overview of the theoretical foundation for time series modeling is presented. The most important concepts and definitions will be highlighted such that they can be built upon in later chapters.

Definition 2.1 (Time Series [1])**.** A time series is a set of observations generated sequentially over time. If the set is discrete, the time series is said to be discrete. Thus, the observations from a discrete time series made at times $\tau_1, \tau_2, \ldots, \tau_t, \ldots, \tau_N$ may be denoted by $x(\tau_1), x(\tau_2), \ldots, x(\tau_t), \ldots, x(\tau_N)$, where $x(\tau_t) \in \mathbb{R}\ \forall t = 1, 2, \ldots, N$.

In this review, I consider discrete time series with fixed interval between observations mathematically described as $x(\tau), x(\tau + i), x(\tau + 2i), \ldots, x(\tau + Ni)$, where $\tau, i \in \mathbb{Z}$ with i typically equal to 1.

2-1-1 Stochastic Modeling

Stochastic models are important in the modeling of time series as almost all real-life processes are nondeterministic. These processes contain uncertainties in the form of external factors that are out of our control. A consequence of this unpredictable nature is that a realization of a certain process will be different when it is measured at a different point in time [13]. The nondeterministic property is caused by inherent randomness in the system that produces the data. Due to this lack of information of the specific process to be analysed, it is not possible to estimate a variable as a fixed value using a deterministic function. Therefore, the term *stochastic process* is introduced.

Definition 2.2 (Stochastic Process [1])**.** A statistical phenomenon that evolves in time according to probabilistic laws is called a *stochastic process*. The time series to be analyzed may then be thought of as a particular *realization*, produced by the underlying *probability mechanism*, of the system under study.

A stochastic process is defined as a collection of random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$, where Ω is a sample space, $\mathcal{F}$ is a σ -algebra, and P is a probability measure [14]. For a given probability space (Ω,$\mathcal{F}$,P) and a measurable space (S, Σ), a stochastic process is a collection of S -valued random variables, which can be written as:

$$\{X_t : t \in \mathbb{T}\},$$

where $\mathbb{T}$ is the set time indices for which there are observations [15].

In order to model stochastic processes, a collection of mappings is required from a set of probabilistic outcomes into measurable quantities. A single mapping between an outcome $\omega \in \Omega$, into a measurable space $\mathbb{R}$, is called a random variable. A formal definition of a random variable is;

Definition 2.3 (Random Variable [16])**.** A random variable is a function $X : \Omega \to \mathbb{R}$ with the property that $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for each $x \in \mathbb{R}$, where $\mathcal{F}$ denotes the set of possible events.

Essentially, a random variable simply represents a functional mapping between Ω into $\mathbb{R}$. From this point onward, upper case letters, such as X, Y, Z, will be used to describe generic random variables, whilst lowercase letters , such as x, y, z represent possible numerical values of these variables.

When modeling stochastic processes, the value of x cannot be described as a fixed value using a deterministic function. In other words, for different realizations of the same stochastic process, x might take on different values. In this case, a mathematical function is required that gives probabilities of occurrence of different possible outcomes of an experiment. This function is called a probability distribution function (PDF).

Definition 2.4 (Probability Distribution Function [5])**.** Given a random variable X defined on the sample space Ω, for any real value $x \in \mathbb{R}$, the probability $\Pr(\{\omega \in \Omega; X(\omega) \leq x\})$ of an event such that $X(\omega) \leq x$ can be determined. If we regard such a probability as a function of x and express it as:

$$\begin{aligned} P(x) &= \Pr(\{\omega \in \Omega; X(\omega) \leq x\}) \\ &= \Pr(X \leq x), \end{aligned}$$

then the function $P(x)$ is referred to as the probability distribution function of X.

The probability distribution fundamentally describes a random variable. Some examples of commonly used probability distributions are the Gaussian normal distribution, the Cauchy distribution and the Laplace distribution. For mathematical descriptions as well as further explanation of these distributions I refer to Konishi [5]. The normal distribution is the most widely used distribution. It is completely described by the mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$ of a random process, and is commonly described functionally as $\mathcal{N}(\mu, \sigma^2)$.

In practice, statistical models are often modeled as a type of probability distribution called conditional distributions. The conditional distribution represents a functional mapping between random variables when one random variable is known to be a particular value. In other words, a conditional distribution is a way to express the probability distribution of a random variable Y given X.

2-1-2 Time Series Models

Time series can be modeled by statistical models. Time series analysis is aimed at identifying the structure that predicts future observations based on past and present measurements. The conditional distribution that describes a time series model functionally is:

$$F_X\left(x_t \mid x_{t-1}, x_{t-2}, \ldots\right), \tag{2-1}$$

where $x_t \in \mathbb{R}$ is the prediction variable and all observations up to time $t-1$ are known.

Autoregressive Models. A particular time series model structure is the autoregressive model. This model describes a random process that predicts an output value $X = \{X_t : t \in \mathbb{T}\}$, with a linear combination of regressed terms of x_t and a white noise term. The autoregressive model of order p, abbreviated as AR(p), is formulated as follows [2]:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{2-2}$$

where c denotes a constant term, and p denotes the number of lags of past observations that contain information that must be used to determine a prediction of X_t. The vector $\varphi = \{\varphi_1, \varphi_2, \ldots, \varphi_p\} \in \mathbb{R}^p$, denotes the set of linear parameters that relates the output value at time t to its lagged terms. In the particular case where the order $p = 0$, the distribution of X_t is uncorrelated with its past observation, also referred to as white noise.

Autoregressive Moving Average (ARMA) Models. In some cases it might be necessary to add a moving average structure to the model if the time series has a complex structure. This model structure assumes that there is correlation in the errors of past observations. Autoregressive moving average models, abbreviated by ARMA(p, q), are a class of models that can be used to model a weakly stationary stochastic process $\{X_t\}$. Similar to the autoregressive model, ARMA models make use of past observations $\{X_{t-1}, \ldots, X_{t-p}\}$ to describe the output variable X_t. Additionally, ARMA models incorporate a linear combination of lagged error terms $\{\varepsilon_t, \ldots, \varepsilon_{t-p}\}$, called the moving average. Mathematically, an ARMA(p, q) model is written as follows [2]:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \tag{2-3}$$

where c is a constant term, p denotes the number of lags, $\varphi = \{\varphi_1, \varphi_2, \ldots, \varphi_p\} \in \mathbb{R}^p$, and $\theta = \{\theta_1, \theta_2, \ldots, \theta_p\} \in \mathbb{R}^q$.

A property of both autoregressive and autoregressive moving-average models is that they are stationary. The concept of stationarity will be discussed in the next section.

2-1-3 Stationarity

Stationarity is a property of a stochastic process. Within time series analysis, stationarity is a common assumption that is often required to analyze and make certain claims about a time series. Time series forecasting relies on the assumption that, to some degree, the statistical

properties of a signal remain uniform over time. Therefore, in order to perform forecasting on time series, the assumption of stationarity is required. Loosely speaking, stationarity implies that the statistical properties of a time series do not change over time. There are different notions of stationarity as strict stationarity is often not realized in real-life stochastic processes. Two of the main notions of stationarity are defined.

Definition 2.5 (Strict Stationarity [17])**.** Formally, let $\{X_t\}$ be a stochastic process and let $F_X\left(x_{t_1+\tau}, \ldots, x_{t_n+\tau}\right)$ represent the cumulative distribution function of the unconditional (i.e., with no reference to any particular starting value) joint distribution of $\{X_t\}$ at times $t_1+\tau, \ldots, t_n+\tau$. Then, $\{X_t\}$ is said to be strictly stationary, strongly stationary or strict-sense stationary if:

$$F_X\left(x_{t_1+\tau}, \ldots, x_{t_n+\tau}\right) = F_X\left(x_{t_1}, \ldots, x_{t_n}\right), \quad \forall \tau, t_1, \ldots, t_n \in \mathbb{R}, \quad \forall n \in \mathbb{N}. \tag{2-4}$$

An example of a stochastic process that is strictly stationary is a white noise process. In many real-life applications the notion of strict stationarity is too restrictive. Therefore, a weaker form of stationarity is commonly used in practice. This form of stationarity is known as weak stationarity or wide-sense stationarity (WSS).

Definition 2.6 (Wide-sense Stationarity [13])**.** A stochastic process $\{X_t\}$ is wide-sense stationary if it satisfies the following three criteria on its mean $m_X(t) \triangleq \mathrm{E}[X_t]$ and autocovariance function $\gamma_X(t_1, t_2) \triangleq \mathrm{E}[(X_{t_1} - m_X(t_1))(X_{t_2} - m_X(t_2))]$:

$$\begin{array}{ll} m_X(t) = m_X(t+\tau) & \text{for all } \tau \in \mathbb{R} \\ \gamma_X\left(t_1, t_2\right) = \gamma_X\left(t_1 - t_2, 0\right) & \text{for all } t_1, t_2 \in \mathbb{R} \\ \mathrm{E}\left[X_t^2\right] < \infty & \text{for all } t \in \mathbb{R}. \end{array} \tag{2-5}$$

Even though wide-sense stationarity is a weaker notion of stationarity than strong stationarity, the latter does not imply the former. This is derived from the fact that strictly stationary processes do not have the requirement of finite variance.

2-2 Regression and Forecasting: What is the difference?

The use of the words *regression* and *forecasting* can sometimes be misleading, and misconceptions about their use are not uncommon. While regression may have multiple meanings depending on the context, I am concerned with regression in a statistical context, also called regression analysis.

Definition 2.7 (Regression analysis)**.** A set of statistical methods for estimating the relationships between a dependent variable and one or more independent variables.

In other words, regression analysis is concerned with finding the parameters of a regression model. Among different methods, the most common type of regression analysis is *linear regression*. Here, one finds the linear combination of coefficients that, according to a specified metric (e.g., ordinary least-squares), most closely fits the data.

In time series, the most commonly used regression model is the autoregressive model (2-2), which is a form of linear regression. Hence, the models used in forecasting are often the same type of models used in regression analysis. However, the models only describe a part of the method. To clarify what distinguishes forecasting from regression, I first give the definition of forecasting.

Definition 2.8 (Time series forecasting)**.** Time series forecasting consists in the use of models to predict the future based on past observations.

Simply speaking, the difference between forecasting and regression analysis can be explained as the difference between *extrapolation* and *interpolation*. Forecasting requires an extra component to assess and reduce the bias induced by the data, such that out-of-sample predictions can be made. On the other hand, regression analysis is concerned only with the relationship between two or more variables.

Specifically, the difference is found in the estimation techniques in either process. Regression analysis applies methods such as ordinary least-squares, where the objective is to minimize the sum of squared errors. Thereby, finding the set of parameters that most closely fits the data. In forecasting, one has to account for bias and uncertainty in the model, which translates into dealing with the risk of underfitting and overfitting – see Subsection 2-3-1. Hence, to address this issue, one often uses information criteria for modeling of time series.

Furthermore, due to the uncertainty associated with extrapolation, forecasting also requires of a measure of uncertainty on a prediction. On the contrary, regression analysis only estimates a point, despite some efforts to quantify uncertainty in regression, as is the case with Bayesian regression ([18], p.20). In Bayesian regression, the uncertainty estimate can grow to be infinite, rendering the prediction practically useless. A forecast should always consist of a point estimate and a uncertainty measure.

2-3 Model Selection

Statistics and machine learning are focused on analysis of increasingly larger volumes of data. Algorithms must be able to learn and make predictions based on data, this is typically done by fitting parametric or nonparametric models. However, there is not a single model that suits all data and all possible goals. Selection of an unsuitable model can lead to incorrect conclusions or disappointing predictions [19]. Therefore, the selection of a model or method can be seen as a vital step in data analysis.

During model selection, candidate models of a class are tested to some performance metric or criterion. The model that is deemed most appropriate, according to a specific criterion, is selected. Model classes may be the variables used in linear regression, the order of an autoregressive process or the number of layers in a neural network. Even though there has been much research into the topic of model selection and many different criteria have been proposed as a systematic tool to select models, there is still no common agreement on the most appropriate method to select models. This is partly due to the fact that, no matter how model selection is conducted, it will always be exploratory in nature and cannot be confirmatory [19]. However, if the right tools are available and correctly used, model selection can provide valuable information with regards to forecasting.

This section discusses model selection within a class of models. Specifically, the class of autoregressive models is highlighted as this is the focus of my thesis.

2-3-1 Bias-Variance Trade-off

The bias-variance trade-off stands central in the problem of model selection. The total error achieved by a model can be broken down into two main components, bias and variance. *Bias* refers to the inherent error of a model with respect to the true model, often caused by simplistic assumptions built into the model, where high bias can cause underfitting. *Variance* refers to the error that is due to sensitivity from fluctuations in the set used to train the model, and is often found in models that are too complex. High variance can cause overfitting. Therefore, the bias-variance trade-off is in fact a trade-off between underfitting and overfitting.

In most cases, model selection for forecasting aims to find the model that achieves the smallest *generalization error*. The generalization error is the error between predicted values and their true values, also called out-of-sample error. This error is often measured using the mean square error (MSE) metric. The model that achieves the smallest mean squared error is therefore considered as the optimal model. However, minimizing either the bias or variance results in maximization of the other. Thus, in order to minimize the generalization error, a trade-off has to be done. Mathematically, the bias-variance trade-off can be derived from the decomposition of the mean squared error.

Let us consider a time series that is described by the model:

$$Y_t = f(X_t) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right) \tag{2-6}$$

where $Y_t \in \mathbb{R}$ and $X_t \in \mathbb{R}^m$ are random variables. The error between a single prediction using predictor $\hat{f}(\cdot)$ and the true observation at time t is defined as:

$$e_t = Y_t - \hat{f}(X_t). \tag{2-7}$$

The expected squared error becomes [20]:

$$\begin{aligned} \mathrm{E}[e_t^2] &= \mathrm{E}[(Y_t - \hat{f}(X_t))^2] \\ &= \mathrm{E}[(f(X_t) + \varepsilon_t - \hat{f}(X_t))^2] \\ &= (f(X_t) - \mathrm{E}[\hat{f}(X_t)])^2 + \mathrm{E}[\varepsilon_t^2] + \mathrm{E}[(\mathrm{E}[\hat{f}(X_t)] - \hat{f}(X_t))^2] \\ &= \mathrm{Bias}[\hat{f}(X_t)]^2 + \sigma_\varepsilon^2 + \mathrm{Var}[\hat{f}(X_t)]. \end{aligned} \tag{2-8}$$

The error consists of the *bias* term, the *variance* term, and an *irreducible noise* term.

This trade-off can also be visualized as a curve as in Figure 2-1. In this figure the difference between in-sample and generalization error is visualized. Increasing the complexity of the model leads to a better fit on the training data, and thus a decrease in bias error. Models that are too complex will overfit, resulting in an increase in variance.

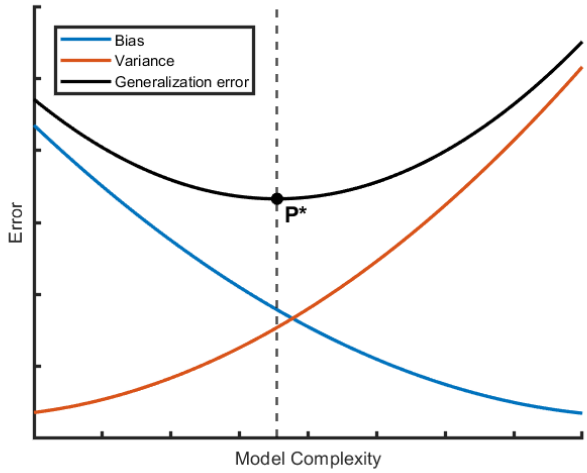


Figure 2-1: The bias-variance trade-off. The point where the generalization error is minimal is denoted by **P***, the dotted line represents the complexity of the model at which the smallest generalization error is achieved.

2-3-2 Information Criteria

Information criteria are measures of information loss. The introduction of an information criterion as a model selection approach was first motivated by Akaike [6]. Akaike proposed the use of Kullback-Leibler (K-L) information [21] for model selection. The measure of K-L information depends on the probability distribution function $G(x)$ that represents the true data generating process, and the probability distribution function $F(x)$ of an arbitrary specified model [5], defined by:

$$I(G;F) = E_G\left[\log\left\{\frac{G(X)}{F(X)}\right\}\right], \tag{2-9}$$

where E_G represents the expectation with respect to $G(x)$.

If the probability distribution functions are discrete, for which probabilities can be expressed as $\{g(x_i) : i = 1, 2, \dots\}$ and $\{f(x_i) : i = 1, 2, \dots\}$, then K-L information can be given as:

$$I(g;f) = \sum_{i=1}^{\infty} g\left(x_i\right) \log\left\{\frac{g\left(x_i\right)}{f\left(x_i\right)}\right\}. \tag{2-10}$$

The K-L information has two important properties [5]:

1. $I(g;f) \geq 0$
2. $I(g;f) = 0 \iff g(x) = f(x)$.

From the properties, one can derive that the model that best estimates the true data is the model for which the K-L information is minimal. This is further clarified when the K-L information is decomposed into:

$$I(g;f) = E_G\left[\log\left\{\frac{g(X)}{f(X)}\right\}\right] = E_G[\log g(X)] - E_G[\log f(X)]. \tag{2-11}$$

Since the first term on the right hand side of the equation is a constant that depends on the true model, it suffices to only consider the second term when comparing models. This term is the expected log-likelihood of the specified density function $f(x)$. Maximizing this term for a set of candidate models will find the model closest to the true model.

In practice, the true probability distribution is unknown. Therefore it is not possible to compute the expected log-likelihood based on $G(x)$. However, an estimate of the expected log-likelihood can be obtained if observed data from the true distribution is available. Then, the true distribution function $G(x)$ can be replaced by an empirical distribution $\hat{G}(x)$ that is based on the observed data. As such, the estimated expected log-likelihood can be written as:

$$E_{\hat{G}}[\log f(X)]. \tag{2-12}$$

Consider a sample of independently observed discrete observations $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$. Then, the expectation can be written as the weighted sum of the logarithms of the density function of each observation:

$$E_{\hat{G}}[\log f(X)] = \sum_{\alpha=1}^{n} \hat{g}(x_\alpha) \log f\left(x_\alpha\right), \tag{2-13}$$

where $f(x)$ is a density function and $x_\alpha \in \boldsymbol{x}$. If each observation has equal probability (i.e., identically distributed), then the empirical distribution function $\hat{g}(x_\alpha) = 1/n$. For the number of observations n tending to infinity, the value of the estimated expected log-likelihood converges to the true expected log-likelihood as a consequence of the law of large numbers [5].

In this case, the log-likelihood may be expressed as:

$$\ell(\theta) = nE_G[\log f(X)] = \sum_{\alpha=1}^{n} \log f_c\left(x_\alpha|\theta\right), \tag{2-14}$$

where $f_c(x|\theta)$ denotes a conditional distribution, having a set of unknown parameters $\theta \in \Theta$. Also, the expression has been multiplied by n. The maximum likelihood estimator (MLE) is found by computing the set of parameters that maximize the log-likelihood function.

However, when modeling time series, the samples are generally not independent and identically distributed, and cannot be expressed as the sum of logarithms of density functions. For instance, the successive (up to an order p) samples of an autoregressive model are correlated. Thus, the observations of the time series cannot be sampled independently.

Nonetheless, there is a way to express the likelihood of a time series by using the conditional distribution for time series models previously mentioned in (2-1). The likelihood can be expressed as:

$$L(\theta) = f_c(x_1, \ldots, x_n|\theta) = \prod_{k=1}^{n} f_c(x_k|x_1, \ldots, x_{k-1}). \tag{2-15}$$

For an autoregressive model of order p, written as:

$$x_k = \sum_{j=1}^{p} a_j x_{k-j} + \varepsilon_k, \quad \varepsilon_k \sim N\left(0, \sigma^2\right), \tag{2-16}$$

the conditional distribution for $k > p$ can be obtained as follows:

$$f_c\left(x_k \mid x_1, \ldots, x_{k-1}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x_k - \sum_{j=1}^{m} a_j x_{k-j}\right)^2\right\}. \tag{2-17}$$

Then, log-likelihood for the time series is obtained by taking the logarithm of the term in (2-15), excluding the first p observations:

$$\ell(\theta) = -\frac{n-p}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\underbrace{\sum_{k=p+1}^{n}\left(x_k - \sum_{j=1}^{p} a_j x_{k-j}\right)^2}_{RSS(a)}, \tag{2-18}$$

where $\theta = \{a, \sigma^2\}$ and $a = \{a_1, \ldots, a_p\}$. The variance term σ^2 can be approximated by computing for which value the derivative of the log-likelihood equals zero as follows:

$$\begin{aligned} \frac{\partial \ell(\theta)}{\partial \sigma^2} &= -\frac{n-p}{2\sigma^2} + \frac{1}{2\sigma^4} RSS(a), \\ \frac{\partial \ell(\theta)}{\partial \sigma^2} &= 0 \implies \hat{\sigma}^2 = \frac{RSS(a)}{n-p}. \end{aligned} \tag{2-19}$$

If the result from (2-19) is substituted back into the log-likelihood equation as expressed in (2-18), then it reduces to:

$$\ell(a|\hat{\sigma}^2) = -\frac{n-p}{2}\log(2\pi\hat{\sigma}^2) - \frac{n-p}{2}. \tag{2-20}$$

In the final step, constant terms may be removed from the equation as these will not affect the maximum likelihood estimate. The set of parameters is obtained by maximizing the log-likelihood as follows:

$$\hat{\boldsymbol{a}} = \max_a \ell(a|\hat{\sigma}^2) = \max_a\{-\frac{n-p}{2}\log\hat{\sigma}^2\}. \tag{2-21}$$

The log-likelihood measure is used in all information criteria that will be discussed in this survey. In statistics, the measure of log-likelihood is understood as an approximation of the K-L information. Therefore, the model with the maximum log-likelihood out of all candidate models is the model that has the smallest loss of information with respect to the data.

Bias Correction. Let us now consider the selection of models for the purpose of prediction. In an ideal case, where the number of observations is infinite, the log-likelihood measure is a approximation of the information loss, and thereby a good metric to determine the best model among candidate models. In practice, the number of observations available are always limited. Consider a set of data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^n$, generated by the true probability distribution $g(x)$. The goal is to measure the goodness-of-fit of an estimated model $f_h(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})$ used to predict future independent data $x_{n+h} \in \mathbb{R}$. In other words, information criteria for forecasting aim to estimate the out-of-sample information of a model.

From the arguments presented in the previous section it would seem that the log-likelihood measure is a good approximation of the information loss in an estimated model. However, this is generally not a fair metric when it comes to forecasting, as the log-likelihood estimator contains a bias towards the observed data $\boldsymbol{x}_n$. This follows from the fact that the estimated log-likelihood uses observed data $\boldsymbol{x}_n$ as a validation method while $\boldsymbol{x}_n$ is also used to estimate the model parameters. The double use of the same data results in a biased prediction of future state x_{n+h}.

Due to this inherent biased estimation of information, a necessary component of an information criterion is a correction for the bias. This gives rise to the general form of an information criterion:

$$\begin{aligned} IC(\boldsymbol{x}_n; \hat{G}) &= -2(\text{maximum log-likelihood - estimated bias}) \\ &= -2\log L(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_n) + 2\{\text{estimated bias of G}\}. \end{aligned} \tag{2-22}$$

The next section discusses several different information criteria, and how these methods compute the estimate of the bias.

2-3-3 Examples of Information Criteria and their Properties

There are two important concepts to determine the quality of a model selection technique that need to be defined.

Definition 2.9 (Consistency [19])**.** A model selection technique is consistent when the probability of the selected model being equal to the true model converges to 1 as $n \to \infty$.

Definition 2.10 (Asymptotic Efficiency)**.** A model selection technique is asymptotically efficient when the probability of the selected model being the model that achieves the smallest prediction loss approaches 1 as $n \to \infty$.

Akaike's information criterion (AIC) [6] is considered as one of the first information criteria, formulated by statistician Hirotugu Akaike, created for the purpose of model selection for prediction. It was first published in 1973 and has since been widely used for model selection. The criterion aims to select the model with the best estimate of out-of-sample prediction error on a given set of data. The idea of AIC is to minimize the in-sample loss while also retaining simplicity by penalizing the number of parameters. The objective the AIC criterion minimizes is:

$$\begin{aligned} &\text{AIC} = -2(\text{maximum log-likelihood}) + 2(\text{number of free parameters}), \\ &\text{AIC} = -2\log \hat{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_n) + 2k, \end{aligned} \tag{2-23}$$

where the number of free parameters $k \in \mathbb{N}$ is equal to the dimension of the parameter vector $\hat{\boldsymbol{\theta}}$ plus one. AIC is a very flexible technique as it does not depend on the true probability distribution that generated the data. In the case of an autoregressive AR(p) model, this equation can be written as:

$$\text{AIC} = n\log(\hat{\sigma}^2) + 2p \tag{2-24}$$

where $p \in \mathbb{N}$ is the order of the model, $n \in \mathbb{N}$ the number of data points and $\hat{\sigma}$ is the average in sample loss. Over the recent decades, many corrections and extensions to AIC have been proposed.

One such correction is the corrected version of AIC (AICc). For small samples, the AIC has a high probability of over-fitting [10]. Therefore, the AICc is often considered in such cases. The AICc is formulated as

$$\text{AIC}_c = \text{AIC} + \frac{2k(k+1)}{n-k-1}. \tag{2-25}$$

Bayesian information criterion (BIC) [7] is another popular information criterion for the purpose of model selection. Contrary to AIC, the aim of BIC is to estimate the true dimension of the underlying data generating process. It differs from AIC is the way the number of parameters is penalized. The objective function that defines BIC is:

$$\text{BIC} = -2 \log \hat{L}(\hat{\boldsymbol{\theta}}; \boldsymbol{x}_n) + k \log n. \tag{2-26}$$

In Table 2-1, obtained from Billah et al. [12], the penalty terms of six different popular information criteria are shown.

Criterion	Penalty Function
AIC	k
BIC	$k\log(n)/2$
AICc	$k + (k^2+k)/(n-k-1)$
HQ	$k\log(\log(n))$
GCV	$-n\log(1-k/n)$
FPE	$(n\log(n+k) - n\log(n-k))/2$

Table 2-1: Penalty functions of the Akaike's Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), Akaike's Information Criterion corrected for small sample size (AICc; Hurvich & Tsai, 1989) , Hannan & Quinn's criterion (HQ; Hannan & Quinn, 1979), the Generalized Cross Validation criterion (GCV; Golub et al., 1979) and the Finite Prediction Error criterion (FPE; Akaike, 1970).

AIC and BIC in literature. I present selected examples from literature that discuss the choice of AIC or BIC, and the considerations that come with the choice.

For instance, Ding et al. [19] argues that the choice between AIC and BIC might depend on the framework of the underlying data generating process. A distinction can be made between a parametric framework, in which there exists a model and a set of parameters that exactly describes the true data generating process, and a nonparametric framework, where the true model is excluded from the class of candidate models. BIC is consistent and asymptotically efficient for the parametric framework. Implying that, if the true model exists among considered models, the criterion will consistently select the true model.

On the other hand, in a nonparametric framework where the true model is not among candidate models, the AIC is asymptotically efficient. Many real-life processes are complex and cannot be described in a parametric framework. Therefore, in cases with large samples, the

AIC is often preferred over BIC regarding predictions. However, when one wants to find the true order of a process, the BIC is often suggested as the preferable criterion.

Additionally, Kuha [22] argues that the complications in comparing AIC and BIC lie in their respective definitions of a 'good model'. The AIC is derived by explicitly denying the existence of an identifiable true model, while BIC obtains its motivation from the Bayesian point of view, aiming to identify the model with the highest probability of being the true model.

Besides the theoretical properties of these information criteria, there have been simulation studies analyzing the empirical performance of information criteria. McQuarrie and Tsai [10] conduct extensive simulations considering several information criteria for a wide range of data. In most of their simulations, when the true model was included in the set of candidate models, BIC outperformed other criteria. On the other hand, AIC outperformed BIC in most large sample simulations that excluded the true model.

Further simulation studies on the behaviour of information criteria were performed by Granger and Jeon [23]. They study the performance of AIC and BIC for forecasting macro economic time-series. Their one-step ahead forecasting performance is evaluated by comparing the mean squared forecast error (MSFE) of each. As a baseline, they used an AR(4) model to compare to either method. Results show that for the chosen metric, the AR(4) model outperforms AIC and BIC for every series. They also noted that BIC usually outperforms AIC.

Moreover, adaptations to the classical approaches are also studied. Mantalos [24] presented a comparative study of model selection criteria for forecasting ARMA models. In this paper, a new criterion is introduced named modified divergence information criterion (MDIC). The study first compares its performance in model identification, i.e., it tests the model on different order ARMA processes. Results suggest that MDIC is a superior technique when it comes to order selection. Another study is conducted to compare predictive performance between criteria by evaluating the MSFE for one-step ahead and five-step ahead forecasts. Generally, the MDIC does perform better than AIC and BIC, but not in all cases.

In addition, Billah et al. [12] proposed an empirical information criterion (EIC) where the penalty term is computed empirically. The suggested algorithm uses a grid search to estimate the penalty that achieves the lowest mean absolute percentage error (MAPE), making the method data-adaptive opposed to arbitrary penalization used by popular information criteria – see Table 2-1, the EIC adapts to the particular forecasting task. Billah reports an improved performance of the EIC compared to AIC and BIC.

2-4 Forecast Metrics

There are multiple ways to evaluate the performance of a specific forecast. In this section each method is mentioned and evaluated. Note that the forecast error is always computed by subtracting the forecasted value at time t from the true observed value at time t, i.e., $e_t = Y_t - \hat{Y}_t$.

Scale-dependent errors are some of the most widely used errors in literature. They include the mean absolute error (MAE) and mean square error (MSE). Respectively, they are computed by $\frac{1}{n}\sum_{i=1}^{n}|e_i|$ and $\frac{1}{n}\sum_{i=1}^{n}e_i^2$. These are scale-dependent methods as they depend on the scale of the time series. Therefore, they are accurate measures for assessing the error on a single

series of data, but are not meaningful for assessing errors over multiple series with varying scale [25].

Percentage errors are scale-independent. They can be measured by $p_t = 100e_t/Y_t$. The property of scale-independence can be used to assess forecast performance across series with different scales. The most common metric using the percentage error is the mean absolute percentage error (MAPE). Similarly to MAE, it is computed as $\frac{1}{n}\sum_{i=1}^{n}|p_i|$. Percentage errors can be disadvantageous when values of Y_t are zero. This causes the percentage error to become infinitely large. Hyndman [26] mentions that for values of Y_t close to zero, the distribution will become very skewed. Furthermore, percentage errors have another downside of penalizing positive errors heavier than negative errors.

Relative errors are an alternative to percentage errors involving some kind of normalization. In the case of relative errors the error is divided by the error obtained through some benchmark method of forecasting. The error is then measured as $r_t = e_t/e_t^*$. The naïve forecasting method is commonly used as a benchmark method as it is easy to compute. The forecast value $\hat{Y}_t$ is equal to the value of the last observation [25]. This metric, like percentage metrics, is scale-independent. However, the method fails when the naïve method produces very small error estimates.

A final type of error metric was proposed by Hyndman et al. [26], named mean absolute scaled error (MASE). This error metric is *scale-free* and avoids the problems that the other error measurements have. First, the error is computed by dividing the true error with the in-sample MAE of the naïve method. The scaled error is defined as:

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}. \tag{2-27}$$

The metric computes the performance of a forecast with respect to the performance of the one-step naïve forecast computed in-sample. Thus, if a forecast is better than the naïve forecast, the scaled error will be smaller than one. On the other hand, values of the scaled error larger than one imply a worse forecast than the one-step naïve forecast. MASE is then obtained by computing $\frac{1}{n}\sum_{i=1}^{n}|q_i|$ over the forecast horizon. The only situation where the MASE would be infinite is when all the observations in the series are equal.

Chapter 3

Problem Solution

In this chapter, I present the problem statement and problem solution to my thesis. I outline a typical situation where one is interested in finding a model to forecast. Here, in the first section, I discuss the flaws that established methods bring about, and I examine a possible solution. In the remaining sections I present the relevant theory and design for the proposed method.

3-1 Problem Formulation

Consider an autoregressive process of order p, or simply AR(p), described by [4]:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = W_t, \quad W_t \sim WN(0, \sigma^2), \tag{3-1}$$

where $\{X_t \in \mathbb{R} : t \in \mathbb{N}\}$ is a stochastic process, with parameters $\{\phi_i \in \mathbb{R} : i = 1, \ldots, p\}$. The process noise is an independent and identically distributed (i.i.d.) Gaussian white noise (WN) sequence, $\{W_t \in \mathbb{R} : t \in \mathbb{N}\}$, with variance $\sigma^2 \in \mathbb{R}^+$. The linear prediction X_{n+h}^n denotes the h-step ahead predictor using the last n measurements which is described as

$$X_{n+h}^n = \phi_{n1}^{(h)} X_n + \phi_{n2}^{(h)} X_{n-1} + \ldots + \phi_{nn}^{(h)} X_1 = \phi_n^{\intercal(h)} X = f_h(X), \tag{3-2}$$

where $X = (X_n, X_{n-1}, \ldots, X_1)^\intercal \in \mathbb{R}^n$ and $\phi_n^{(h)} \in \mathbb{R}^n$ for forecast horizon $h \in \mathbb{N}$.

I am interested in finding a model that follows the linear structure in (3-2) that *best* represents the true process as described in (3-1), according to some predefined metric. Selection of models is concerned with finding a parametric model for a specified objective, such as minimum error between forecasted and measured data. In order to find the predictor that obtains minimum mean squared error Theorem 3.1 is required.

Theorem 3.1 ([2])**.** Let X,Y be discrete random variables. For any function $g : \mathbb{R} \to \mathbb{R}$,

$$E\left[(Y - E[Y \mid X])^2\right] \leq E\left[(Y - g(X))^2\right],$$

and we have equality if and only if $g(X) = E[Y \mid X]$.

Ideally, we find the optimal linear predictor by minimizing the mean squared error over a specified prediction horizon, as follows:

$$\begin{aligned} &\min_{f_i} \frac{1}{h} \sum_{i=1}^{h} E[(X_{n+i} - f_i(X))^2 \mid X] \\ =&\min_{f} E[E[Y - f(X))^2 \mid X]], \end{aligned} \tag{3-3}$$

where $Y = (X_{n+1}, X_{n+2}, \ldots, X_{n+h})^\intercal \in \mathbb{R}^h$ and $f(X) = \{f_i(X) : \mathbb{R} \to \mathbb{R},\ i = 1, \ldots, h\}$. Consequently, using Theorem 3.1, it follows that (3-3) is minimized when

$$f(X) = E_Y[Y \mid X]. \tag{3-4}$$

Hence, it is only possible to determine $f(X)$ when the probability distribution of Y is known, which is not the case. Alternatively, if measured values of $y_{n+h} \in \mathbb{R}$ are known, we may use them to construct an empirical distribution. In other words, measured data can be used to fit a model. However, if I consider the objective is to find a model for forecasting, future values are not at my disposal. Thus, instead of using future values to obtain an empirical distribution, past (i.e., available) data must be used to find this function. Model selection for forecasting is concerned with the fitting of models that approximate the probability distribution of the underlying process.

In practice one is often required to work with small to medium sized samples of data. Here, information criteria are useful to reduce over-fitting in model selection. However, in small samples, the asymptotic properties of information criteria no longer hold, therefore, these criteria lack theoretical basis or proof to support their choice on how to penalize model complexity (hence, the large number of different criteria found in literature). Furthermore, extensive empirical studies [10, 11] argue that the predictive performance of information criteria depends largely on the type of data and its characteristics, and not one criteria performs uniformly better than the others. Nonetheless, most existing information criteria have a static penalty term (i.e., independent of data), except for criteria such as the empirical information criterion (EIC) [12].

Research question. Altogether, I believe the conventional information criteria (AIC, AICc, and BIC) leave room for improvement from a theoretical perspective and potentially also from an empirical perspective. The question that I seek to answer is:

Is it possible to formulate a model selection criterion as an optimization problem that uses a data-driven method to quantify forecast performance among models for locally stationary, finite sample time series?

To answer that question, let us first introduce a different situation. Consider the process described in (3-1). Instead of trying to find the optimal linear predictor for the h-step ahead prediction, we want to find the optimal linear predictor for the h-step back prediction.

$$X_{1-h}^{n} = \phi_{n1}^{(h)} X_1 + \phi_{n2}^{(h)} X_2 + \ldots + \phi_{nn}^{(h)} X_n = \phi_{n}^{\intercal(h)} X_B = f_h^B(X_B), \tag{3-5}$$

where $X_B = (X_1, X_2, \dots, X_n)^\intercal \in \mathbb{R}^n$ is a reversed version of X, and $\phi_n^{(h)} \in \mathbb{R}^n$ for backcast horizon $h \in \mathbb{N}$. I present a similar argument to minimize the mean square error, as follows:

$$\begin{aligned} &\min_{f_i} \frac{1}{h} \sum_{i=1}^{h} E[(X_{1-i} - f_i^B(X_B))^2 \mid X_B] \\ =&\min_{f} E[E[Y_B - f^B(X_B))^2 \mid X_B]], \end{aligned} \tag{3-6}$$

where $Y_B = (X_0, X_{-1}, \dots, X_{1-h})^\intercal \in \mathbb{R}^h$ and $f^B(X) = \{f_i^B(X) : \mathbb{R} \to \mathbb{R},\ i = 1, \dots, h\}$. Consequently, using Theorem 3.1, it follows that (3-6) is minimized when

$$f^B(X_B) = E_{Y_B}[Y_B \mid X_B]. \tag{3-7}$$

Again, I can approximate $f^B(X_B)$ if I have an approximation of the distribution of Y_B. However, in contrast to the forward prediction, I can now construct an empirical distribution of Y on the basis of past data. Therefore, I am able to find $f^B(X_B)$. In the next section, I argue that, from a theoretical perspective, $f^B(X) = f(X)$.

3-2 Reversibility of Time Series

A time series is time-reversible when the sequence of random variables $\{X_t, \dots, X_{t+h}\}$ has equal joint probability distribution to the sequence $\{X_{t+h}, \dots, X_t\}$ [27]. Standard AR models driven by Gaussian noise are time-reversible [28]. Therefore, AR models are independent of the direction in which time progresses.

That said, let us consider a case where we want to compare the one-step ahead prediction with the one-step backward prediction making use of n measurements. The linear prediction of the one-step ahead prediction is given by [1]

$$X_{n+1}^n = \varphi_{n1} X_n + \varphi_{n2} X_{n-1} + \ldots + \varphi_{nn} X_1 = \varphi_n^\intercal X, \tag{3-8}$$

with $\varphi_n \in \mathbb{R}^n$. The linear prediction for the one-step backcast is described by

$$X_0^n = \phi_{n1} X_1 + \phi_{n2} X_2 + \ldots + \phi_{nn} X_n = \phi_n^\intercal X_B, \tag{3-9}$$

with $\phi_n \in \mathbb{R}^n$. In this case the vector containing the states is $X = (X_n, X_{n-1}, \dots, X_1)^\intercal$. Here, X_B denotes a time reversed vector of X which is formulated as $X_B = JX$, where $J \in \mathbb{R}^{n \times n}$ is an anti-diagonal identity matrix described as follows:

$$J = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

Next, the Yule-Walker equations (or, *prediction equations*) can be used to determine the parameters for both the predictors as a function of the autocovariance function of the X_t. The Yule-Walker equations are described as follows [1]:

$$\begin{aligned} \Gamma_n \varphi_n &= \gamma_n, \quad \text{for the forward prediction, and} \\ \Gamma_n \phi_n &= \gamma_n, \quad \text{for the backward prediction.} \end{aligned} \tag{3-10}$$

Their corresponding variances are given by

$$\begin{aligned} \sigma_f^2 &= \gamma(0) - \varphi_n^\intercal \gamma_n, \quad \text{for the forward prediction, and} \\ \sigma_b^2 &= \gamma(0) - \phi_n^\intercal \gamma_n, \quad \text{for the backward prediction.} \end{aligned} \tag{3-11}$$

In these equations, $\Gamma_n \in \mathbb{R}^{n \times n}$ is the autocovariance matrix described by

$$\Gamma_n = \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(n-2) & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \dots & \gamma(n-3) & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(n-2) & \gamma(n-3) & \dots & \gamma(0) & \gamma(1) \\ \gamma(n-1) & \gamma(n-2) & \dots & \gamma(1) & \gamma(0) \end{pmatrix},$$

and $\gamma_n = \{\gamma(i) : i = 1, \dots, n\}$ is the vector of autocovariances up to n lags. Both Γ_n and γ_n consist of values from the autocovariance function of $\{X_t\}$, which is symmetric. This means that $\gamma_X(k) = \gamma_{X_B}(k)$. Therefore, the function is independent of the direction of time in which the time series is ordered. Hence, we have that $\varphi_n = \phi_n$.

Example. To further illustrate the property of time-reversibility, consider an AR(1) process described by

$$X_t = \alpha X_{t-1} + W_t, \quad W_t \sim WN(0,1), \quad t = 0, 1, \dots \tag{3-12}$$

If the Yule-Walker equations are applied for the forward prediction as described in (3-8), we obtain

$$\gamma(0)\varphi_{11} = \gamma(1). \tag{3-13}$$

Alternatively, the autocovariance $\gamma(k)$ is found by taking the expectation of the process from (3-12) with a shifted version of itself as follows:

$$\begin{aligned} \gamma(k) &= E[X_t X_{t-k}] \\ &\overset{(3\text{-}12)}{=} \alpha E[X_{t-1} X_{t-k}] + \underbrace{E[\varepsilon_t X_{t-k}]}_{=0} \\ &= \alpha \gamma(k-1). \end{aligned} \tag{3-14}$$

Now, using the result from (3-13) and (3-14), and the fact that the forward and backward parameter vectors are the same, we can conclude that

$$\begin{aligned} \varphi_{11} = \phi_{11} &= \alpha, \text{ and} \\ \sigma_f^2 = \sigma_b^2 &= \gamma(0) - \alpha\gamma(1). \end{aligned} \tag{3-15}$$

This results suggests that the model for the one-step ahead prediction is identical to the model of the one-step backward prediction.

In summary, the above derivation is readily applicable to all orders of an AR(p) model, and also for different prediction horizons. Thus, this means that $f^B(X) = f(X)$. As such, we can leverage previous data to assess the quality of the backcasting, which serves as a proxy to the forecasting capabilities. In the next section, I systematically leverage this insight to introduce a novel information criteria that builds upon these ideas.

3-3 The Backwards Validated Information Criterion

In this section, I introduce a novel information criterion to be referred to as the *backwards validated information criterion (BVIC)* – not to be confused with BIC. The BVIC consists in the integration and unification of three components.

3-3-1 The BVIC Components

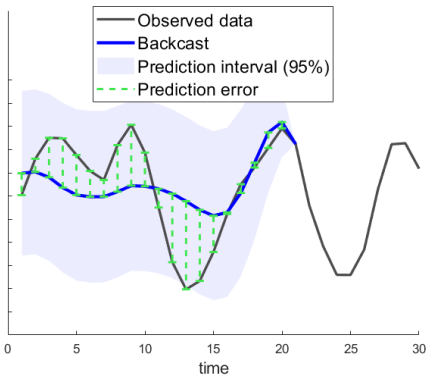
The first component is the *log-likelihood function*. The log-likelihood function can be used to compute the MLE. The MLE is the set of parameters for which the log-likelihood is maximized. For an autoregressive model $j \in \mathcal{M}$ with parameters $\theta_j \in \mathbb{R}^p$, where $p \in \mathbb{N}$, the log-likelihood may be expressed as [5]

$$\ell(\theta_j) = -\frac{n}{2} \log \hat{\sigma}(\theta_j)^2, \tag{3-16}$$

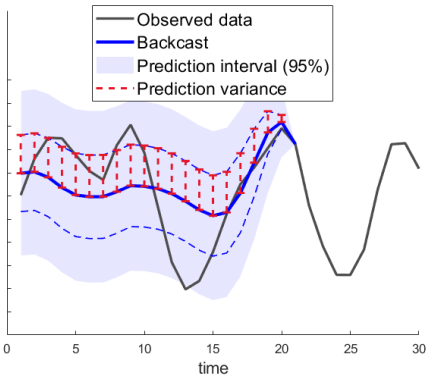
where $n \in \mathbb{N}$ is the number of observations and $\hat{\sigma}(\theta_j)^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - X_i^n(\theta_j))^2$ is the average ℓ_2-loss function for all observations.

Accordingly, an additional component is required that prevents over-fitting inherent to the MLE. Using the results from Section 3-2, where I show that the backwards linear prediction may be used as a proxy for the forward linear prediction, I propose to utilize the backwards prediction to assess the forecast capabilities of candidate models. I consider two metrics to quantify the performance of the backwards prediction: (*i*) the backcast error, and (*ii*) the backcast uncertainty. More specifically, I consider the mean square backcast error (MSBE) and the backcast variance as suitable metrics for evaluation.

Consider an arbitrary autoregressive process where we want to apply the previously described metrics. First, the mean square backcast error may be obtained by taking the mean of the squared error over backcast horizon h, i.e., $\text{err}(\theta_j) = \frac{1}{h} \sum_{t=1}^{h} e_{1-t}^2$, where e_t is depicted in the green dotted lines in Figure 3-1a. Second, the backcast variance represents the variance associated with the backwards prediction for each step. Specifically, I use the average variance over the backcast horizon, illustrated in Figure 3-1b with red dotted lines, which can be described functionally as $\text{var}(\theta_j) = \frac{1}{h} \sum_{t=1}^{h} P_{1-t}^n$.



(a) Sample backcast with the prediction error at each time step indicated with a dotted green line.



(b) Sample backcast with the prediction variance at each time step indicated with a dotted red line.

Figure 3-1: Backcasting metrics. An illustration of the metrics used to quantify the backcasting performance of a specific model.

The last component I consider for the novel criterion is the *model order*. To counteract the risk of over-fitting, classical criteria incorporate the model order p as a penalty variable in their model selection procedure. Consequently, over-fitting is reduced by penalizing higher orders, resulting in increased parsimony of selected models. Arguably, the manner in which model complexity is penalized may differ depending on the application – see Table 2-1. However, classical information criteria operate with a static penalty term, therefore, they are susceptible irregularities and differences between datasets. This results in the absence of a uniformly preferable criterion [10], and leaves it up to the user to determine the best criterion for their data. Considering the data-driven approach to obtaining generalization capabilities that was introduced in the former paragraphs, I decided to exclude the order penalization component from the novel criterion.

Altogether, excluding the model order, we are left with three components. From these three components, a weighted combination is constructed that represents the basic structure of the novel criterion, as follows:

$$-a \times \frac{\ell(\theta_j)}{n} + b \times \mathrm{err}(\theta_j) + c \times \mathrm{var}(\theta_j), \tag{3-17}$$

where a, b, and c are real-valued positive scalars. Technically, θ_j depends on the horizon of the backcast h, therefore, the backcasting components should be $\mathrm{err}(\theta_j^{(1)}, \ldots, \theta_j^{(h)})$ and $\mathrm{var}(\theta_j^{(1)}, \ldots, \theta_j^{(h)})$. However, for simplicity I only use θ_j to indicate the index variable.

The expression in (3-17) has three parameters. To simplify the multi-objective problem I will reduce the number of parameters and normalize the expression. Firstly, consider a simplification step to remove a parameter, a, from the equation in (3-17). For the purpose of model selection, I am only interested in the relative differences between the terms in the criterion. Therefore, it is possible to set α as a constant, while keeping the other two variable. Secondly, each of the terms in the equation is normalized by dividing each expression with a baseline constant value. This value is computed by estimating the parameters θ_{p^*} with a fixed order $p^* = \max_{\{j \in \mathcal{M}\}} j$. Hence, I obtain

$$
\begin{aligned}
a &= \frac{n}{|\ell(\theta_{p^*})|}, \\
b &= \frac{\beta}{\text{err}(\theta_{p^*})}, \text{ and} \\
c &= \frac{\gamma}{\text{var}(\theta_{p^*})}.
\end{aligned} \tag{3-18}
$$

Depending on the scale of the data in the time series, the magnitude of the MLE might be positive or negative. Therefore, to ensure that the normalization does not flip the sign of the objective function, the absolute value of $\ell(\theta_{p^*})$ is used.

Finally, the results from (3-18) can be substituted into the expression in (3-17) to obtain a final criterion. The BVIC is given by

$$
\text{BVIC}(\theta_j, j) = -\frac{\ell(\theta_j)}{|\ell(\theta_{p^*})|} + \beta\frac{\text{err}(\theta_j)}{\text{err}(\theta_{p^*})} + \gamma\frac{\text{var}(\theta_j)}{\text{var}(\theta_{p^*})}, \tag{3-19}
$$

where β, $\gamma \geq 0$ are tunable hyperparameters that can be used to increase the importance of each of the three terms in the expression.

3-3-2 Parameter Estimation

Now that the structure of the BVIC has been decided upon, a method for estimating the parameters must be chosen. I split the estimation of the parameters up into two parts and compute the parameters through separate estimation techniques, detailed next. Hence, the BVIC becomes an index criterion instead of a multi-objective optimization criterion, similar to AIC and BIC. Specifically, an estimate of the indices is computed for different model orders and the one that attains a minimum is considered.

Therefore, for the log-likelihood term I use parameters estimated through MLE to obtain a measure of goodness-of-fit [5]. Here, the argument of (3-16) is maximized as follows:

$$
\hat{\theta}_j = \arg\max_{\theta_j} \ell(\theta_j), \tag{3-20}
$$

where $\hat{\theta}_j \in \mathbb{R}^p$ is the set of parameters for which the log-likelihood function is maximized, i.e., the MLE. As such, the function $\ell(\hat{\theta}_j)$ is essentially a measure of *goodness-of-fit* of the model parameters upon the observed data.

Secondly, as is common in forecasting, I estimate the parameters for the backcasting error and variance using the Yule-Walker equations [1]. The Yule-Walker equations were discussed previously in Section 3-2. This technique is advantageous because it can compute both point estimates and prediction variance. Additionally, this method can be used to estimate parameters for multi-step forecasts. The parameters obtained with Yule-Walker are found by

$$
\theta_j^{YW} = \Gamma_j^{-1}\gamma_j, \tag{3-21}
$$

where $\Gamma_j \in \mathbb{R}^{j\times j}$ is the autocovariance matrix, and $\gamma_j \in \mathbb{R}^j$ is the vector of autocovariances containing j lags.

Altogether, the estimated parameters can be used to formulate a mathematical objective criterion to determine the order of an autoregressive model. This criterion is

$$\text{BVIC}(j) = -\frac{\ell(\hat{\theta}_j)}{|\ell(\hat{\theta}_{p^*})|} + \beta\frac{\text{err}(\theta_j^{YW})}{\text{err}(\theta_{p^*}^{YW})} + \gamma\frac{\text{var}(\theta_j^{YW})}{\text{var}(\theta_{p^*}^{YW})}, \tag{3-22}$$

where $\beta, \gamma \geq 0$. Thus, the model order is obtained through minimization of the BVIC, as follows:

$$j^* = \arg\min_{j \in \mathcal{M}} \text{BVIC}(j). \tag{3-23}$$

Alternatively, I could have performed a multi-objective optimization by consecutively estimating the parameters for each index and computing their respective minimum. The optimization problem would be

$$\theta_j^* = \arg\min_{\theta_j} \text{BVIC}(\theta_j, j), \quad \forall j \in \mathcal{M}, \tag{3-24}$$

followed up by

$$j^* = \arg\min_{j} \text{BVIC}(\theta_j^*, j). \tag{3-25}$$

However, after thorough consideration, I decided that the problem from (3-24) was too computationally expensive and complex to solve. Instead, I decided to take the previously discussed approach and use the BVIC as an index. Nonetheless, I believe there may be merit in exploring the possibility of solving the problem as a multi-objective optimization, but due to time constraints I suggest to explore this in future research.

3-4 Dimensions of the BVIC

Interestingly, the BVIC can be divided into four dimensions: (a) regression, (b) generalization, (c) uncertainty and (d) forecasting. Each dimension illustrates a functionality of the criterion with regards to time series analysis. The four dimensions are depicted in four quadrants in Figure 3-2. Each quadrant contains the functionality of the BVIC depending on the selection of the parameters β and γ. For instance, if $\beta, \gamma = 0$, the BVIC is equal to the MLE. When $\gamma = 0$ and $\beta \gg 1$, the BVIC selects models with the smallest out-of-sample error (on the backcast) that intuitively corresponds to generalization. When both $\beta, \gamma \gg 1$, the focus of the BVIC is to minimize the out-of-sample error along with the theoretical variance (uncertainty) of the backcast, thereby also minimizing these quantities for the forecast.

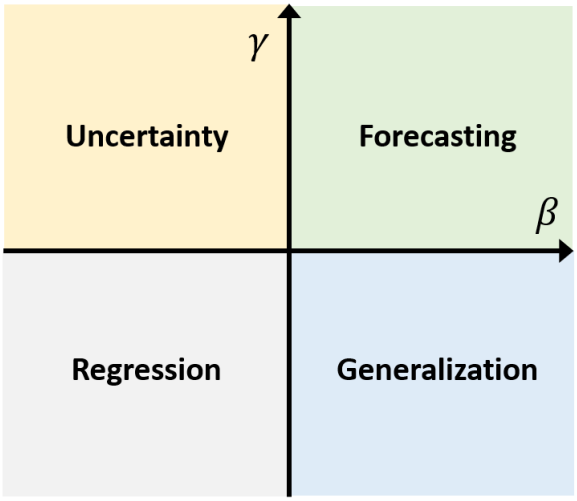


Figure 3-2: Dimensions of the BVIC. The four dimensions of the BVIC divided into four quadrants depending on the size of parameters β and γ. The intersection of the arrows does *not* represent the point at which $\beta, \gamma = 0$.

That said, recall that a forecast consists of both the point and variance estimate, and as such, they both play a key role in forecasting that simply speaking would relate with the precision and accuracy of the predictions. Thus, the different dimensions associated with the parameters of the BVIC give a principle approach to model selection, contrasting with the classical information criteria previously discussed (i.e., AIC, AICc, and BIC). Specifically, this is achieved by replacing the penalization of the order with a regularization term on the backward validation metric and adding parameters that introduce adaptability in the the model selection problem. In other words, the adaptability allows users of the BVIC to select their preferred forecasting goal.

Chapter 4

Simulations

In what follows, I perform Monte Carlo simulations using three experiments with both synthetic and real data. Specifically, the first and second experiments consist in generating data according to specified autoregressive models, and assessing the model order obtained through different information criteria, as well as the goodness-of-fit. Next, I assess the quality of the models when considering intracranial electroencephalographic (i.e., electrocorticographic, or ECoG for short) data from epileptic patients undergoing a seizure.

4-1 Data Description

Firstly, for the synthetic data I generate an autoregressive process of order $p \in \mathbb{N}$ and parameters $\phi \in \mathbb{R}^p$ as described in (3-1). Subsequently, similar to [9], noise is added to the synthetic data. First, the realization is normalized such that the mean and variance are equal to zero and one, respectively. Second, the noise is added as follows:

$$Y_t = X_t + \delta Z_t, \quad \{Z_t\} \sim \mathcal{N}(0,1), \tag{4-1}$$

from which I obtain $Y = \{Y_t : t = 1, 2, \ldots, N\} \in \mathbb{R}^N$ that contains $N \in \mathbb{N}$ measurements. Note that N is used to describe the entire generated sample size, whereas n denotes the effective sample size that may be used for training. Moreover, the sequence $\{Z_t\}$ is i.i.d., and the parameter $\delta \geq 0$ may be used to determine the signal-to-noise ratio (SNR).

Secondly, the ECoG data is obtained from the International Epilepsy Electrophysiology Portal (IEEG Portal) [29]. I look at a range of channels from three different patients from two different locations. The first and second datasets are acquired from two separate patient studies at the Hospital of the University of Pennsylvania, Philadelphia, where the ECoG signals were recorded at a sampling frequency of 512 Hz. The third dataset is recorded at a frequency of 500 Hz, and is from a patient study at the the Mayo Clinic in Rochester, Minnesota.

Seizures are marked by clinical experts [30] and the seizure-onset time and location are defined by the so called *earliest electrographic change (EEC)* and the *uneqiovocal electrographic onset (UEO)* [31], where I consider the period between EEC and UEO to be the pre-ictal phase, i.e., the phase between a normal (interictal) state and a seizing (ictal) state.

I extract univariate time series blocks from channels in which seizures were identified. Each block has been associated to one of three states of the brain, being: (*i*) interictal, (*ii*) pre-ictal, or (*iii*) ictal. Subsequently, two steps of pre-processing were performed on the data. Initially, the common reference was removed from all the recorded data. Hereafter, each recording is filtered through a 60 Hz notch filter to remove line-noise present in the recordings. Both these steps were also performed in [30], where the same database is used.

4-2 Experimental Setup

A specific realization of Y, i.e., $y = \{y_i : i = 1, \dots, N\}$ is referred to as a *window* (of data collected over a period of time). A single window is denoted by $w_j \in \mathbb{R}^{S_w}$, with $j = 1, 2, \dots, N_w$, and $S_w \in \mathbb{N}$ denotes the size of the windows. In each experiment, I generate a collection of $N_w \in \mathbb{N}$ windows. Each window is split into a training set $\mathcal{T}_j \in \mathbb{R}^{S_\mathcal{T}}$ and test set $\mathcal{T}_j^* \in \mathbb{R}^{S_{\mathcal{T}^*}}$ – see Figure 4-1. The respective sizes of the training and validation set depend on the true order (which is known) of the autoregressive process and the forecasting horizon $h \in \mathbb{N}$. Specifically, they can be formulated as $S_\mathcal{T} = 2(p + h)$ and $S_{\mathcal{T}^*} = h$. As a result, when $p = h$, I have training and testing split of $\mathcal{S}_\mathcal{T} = 4h$ (80%) and $\mathcal{S}_{\mathcal{T}^*} = h$ (20%). Subsequently, to ensure that the windows have sufficient samples, the window size is chosen to be $S_w = S_\mathcal{T} + 2S_{\mathcal{T}^*}$. Finally, the windows are always normalized (z-scored) to facilitate a fair evaluation. Lastly, it is important to notice that for the BVIC, the training set consists of a backwards training set $\mathcal{T}_{\mathcal{B},j} \in \mathbb{R}^{S_{\mathcal{T},\mathcal{B}}}$ and a backwards validation set $\mathcal{V}_{\mathcal{B},j} \in \mathbb{R}^{S_{\mathcal{V},\mathcal{B}}}$, where $S_{\mathcal{V},\mathcal{B}} = h$ and, consequently, $S_{\mathcal{T},\mathcal{B}} = S_\mathcal{T} - h$.

4-3 Metrics

The results from the Monte Carlo simulations are evaluated based on three metrics. The first two metrics are based on the L^2-loss function of the forecast (i.e., the mean squared error over the forecast horizon). This metric is computed as follows:

$$L^2_{w,m}(h) = \frac{1}{h} \sum_{i=n+1}^{n+h} (Y_{i,w,m} - Y^n_{i,w,m})^2, \tag{4-2}$$

where $w \in \{1, 2, \dots, W\}$ is the index of the window, and $m \in \{1, 2, \dots, M\}$ is the index of the Monte Carlo simulation. The metrics MSE and VAR are calculated by respectively taking the mean and variance over all the windows and simulations as follows:

$$\begin{aligned} \text{MSE} &= \frac{1}{MW} \sum_{j=1}^{M} \sum_{i=1}^{W} L^2_{i,j}(h), \quad \text{and} \\ \text{VAR} &= \frac{1}{MW} \sum_{j=1}^{M} \sum_{i=1}^{W} (L^2_{i,j}(h) - \text{MSE})^2. \end{aligned} \tag{4-3}$$

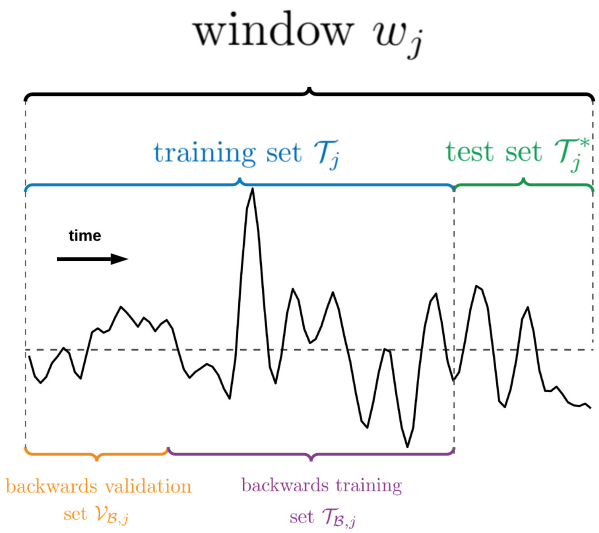


Figure 4-1: Data Splitting. This figure depicts how each data segment (i.e., window) is divided into training and testing data. Special emphasis is drawn to the backward validation scheme required to assess the performance of the BVIC.

Additionally, I consider the prediction uncertainty of the forecast as an evaluation metric by taking the average variance over the forecast horizon, i.e.,

$$P_{w,m} = \gamma_{w,m}(0) - \frac{1}{h}\sum_{i=1}^{h} \gamma'^{(i)}_{n,w,m} \Gamma^{-1}_{n,w,m} \gamma^{(i)}_{n,w,m}. \tag{4-4}$$

Subsequently, the mean over all simulations can be computed as

$$\bar{P}_F = \frac{1}{MW}\sum_{j=1}^{M}\sum_{i=1}^{W} P_{i,j}. \tag{4-5}$$

4-4 Experiments

4-4-1 Experiment 1

In this experiment, I test the ability of the different models determined using the different information criteria to forecast different time series. I evaluate four different synthetic AR(5) models by conducting Monte Carlo simulations for the BVIC and benchmark criteria. Since the autoregressive model is a discrete linear filter, the parameters of the model can be determined when the poles (or, roots) of the system are known [1]. For the autoregressive process to be stationary, the poles of the system need to lie inside the unit circle. The location of the poles affect the frequency behaviour and exponential decay of the time domain signal. For instance, a larger phase angle of a complex conjugate pole set results in higher frequency of the time domain signal. Essentially, the dominant pole(s) (i.e., the poles that lie nearest to the unit circle) of the system determine the majority of this behaviour. Therefore, I define

four sets of dominant poles that display different response behaviour to assess the different information criteria.

- **Case 1.** This set of poles chosen to be similar to the poles of a true ECoG recording, if it was of order 5. These poles are computed using a least squares system identification method detailed in Appendix B. The dominant poles are a positive real pole of $z = 0.9$, and set of complex conjugate poles with positive real part, $z = 0.6 \pm 0.6i$. This results in a frequency of $\omega_0 = 0.79$ rad/s. The magnitude of the poles will lead to slow exponential decay while the phase angle of the complex conjugate will induce intermediate sinusoidal behaviour.
- **Case 2.** The dominant poles are a set of complex conjugate poles with negative real part, $z = -0.6 \pm 0.6i$. This results in a frequency of $\omega_0 = 2.36$ rad/s. The magnitude of the poles will lead to slow exponential decay while the phase angle will induce intermediate sinusoidal behaviour as well as sign switching due to the negative component, resulting in a very high frequency.
- **Case 3.** The dominant poles are a set of complex conjugate poles with small positive real part, $z = 0.1 \pm 0.9i$. This results in a frequency of $\omega_0 = 1.46$ rad/s. The magnitude of the poles will lead to slow exponential decay while the phase angle will induce high frequency sinusoidal behaviour.
- **Case 4.** The dominant poles are a set of complex conjugate poles with positive real part, $z = 0.5 \pm 0.1i$. This results in a frequency of $\omega_0 = 0.20$ rad/s. The magnitude of the poles will lead to (relatively) fast exponential decay while the phase angle will create barely any sinusoidal behaviour.

In Figure 4-2 the four cases are illustrated in a pole-zero map. Along with the pole-zero maps, for each case a time domain sample is displayed from a realization of the process generated using the mentioned poles.

Along with the benchmark criteria, I further consider the BVIC with two different sets of parameters to evaluate the penalization effect each term has on the forecasting performance. The sets are as follows: (*i*) $(\beta, \gamma) = (1, 1)$, and (*ii*) $(\beta, \gamma) = (5, 1)$. Finally, I conduct the Monte Carlo simulations for three different values of the noise parameter, δ. Specifically, I chose the values of $\delta \in \{0, 0.1, 0.316\}$, which corresponds to a signal-to-noise ratio of SNR $= \infty$ dB (no noise), 10 dB, and 5 dB, respectively. In Table 4-1, I summarize the results from Experiment 1.

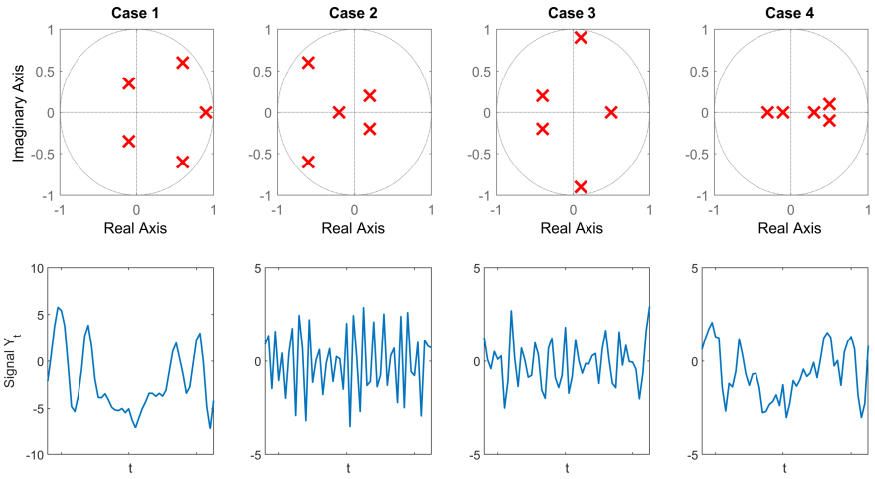


Figure 4-2: Pole-zero maps and observation samples for all cases. The first row contains the pole-zero maps of the four autoregressive processes. Poles are annotated with a $\times$. The second row contains a sample from a realization from each of the cases.

Table 4-1: Summary statistics for the Monte Carlo simulations conducted in Experiment 1.

	$\delta = 0$ (no noise)				$\delta = 0.1$ (10 dB)				$\delta = 0.316$ (5 dB)			
Case 1	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.893	0.49	0.53	4.6	0.902	0.51	0.54	4.8	0.960	0.50	0.61	4.7
$BVIC(\beta = 5, \gamma = 1)$	0.868	0.50	0.56	3.6	0.881	0.49	0.57	3.6	0.945	0.51	0.64	3.5
AIC	0.864	0.51	0.55	3.4	0.878	0.53	0.57	3.1	0.910	0.51	0.66	2.6
BIC	0.864	0.49	0.55	3.1	0.869	0.51	0.58	2.6	0.895	0.49	0.67	2.1
$AICc$	0.857	0.49	0.56	2.5	0.857	0.52	0.58	2.2	0.902	0.50	0.68	1.8
Case 2	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.842	0.49	0.65	4.7	0.874	0.54	0.65	4.7	0.943	0.53	0.67	5.1
$BVIC(\beta = 5, \gamma = 1)$	0.852	0.52	0.68	3.6	0.874	0.51	0.67	3.7	0.935	0.49	0.70	3.9
AIC	0.862	0.51	0.69	2.7	0.884	0.51	0.64	2.6	0.953	0.51	0.73	2.7
BIC	0.860	0.50	0.70	2.3	0.876	0.49	0.70	2.4	0.942	0.49	0.74	2.3
$AICc$	0.847	0.48	0.71	2.0	0.866	0.49	0.71	1.9	0.933	0.50	0.75	2.0
Case 3	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.770	0.36	0.68	5.0	0.781	0.40	0.68	5.0	0.826	0.43	0.70	5.2
$BVIC(\beta = 5, \gamma = 1)$	0.783	0.40	0.70	3.7	0.794	0.42	0.71	3.7	0.844	0.43	0.74	3.6
AIC	0.772	0.41	0.71	3.2	0.783	0.42	0.71	3.1	0.819	0.44	0.75	3.0
BIC	0.777	0.41	0.72	2.7	0.787	0.43	0.73	2.7	0.825	0.42	0.76	2.3
$AICc$	0.762	0.38	0.73	2.2	0.778	0.40	0.74	2.1	0.816	0.42	0.77	2.0
Case 4	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$	MSE	VAR	$\overline{P}_F$	$\bar{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.930	0.44	0.66	4.6	0.942	0.43	0.66	4.6	0.972	0.43	0.69	4.6
$BVIC(\beta = 5, \gamma = 1)$	0.933	0.43	0.68	3.7	0.931	0.41	0.68	3.8	0.967	0.39	0.71	3.6
AIC	0.941	0.42	0.71	2.4	0.951	0.43	0.72	2.4	0.974	0.41	0.75	2.5
BIC	0.927	0.41	0.73	1.8	0.969	0.46	0.74	1.8	0.992	0.44	0.77	1.9
$AICc$	0.920	0.41	0.74	1.5	0.961	0.44	0.75	1.5	0.995	0.43	0.79	1.4

Table notes: the rows of the table display the criteria for each of the four cases. The columns show the metrics that I use to evaluate the criteria. Furthermore, the criteria are evaluated for different values of noise parameter δ.

4-4-2 Experiment 2

The objective is to assess the performance of the BVIC on a range of synthetic autoregressive time series of order $p \in \{10, 20, 30, 40, 50\}$. I conduct a Monte Carlo study in which I generate 100 windows with characteristics similar to those set in Experiment 1. Each window is generated by a set of poles that was generated randomly. For window w_j, with $j = 1, \ldots, N_w$, complex conjugate poles are generated by randomizing a phase angle Φ_j and a magnitude M_j, where $0 \leq \Phi_j \leq \pi$, and $0.5 \leq M_j < 1$. I define a set of complex conjugate poles with real and imaginary part described by $\alpha_j = M_j \cos \Phi_j$, and $\beta_j = M_j \sin \Phi_j$, respectively. To include the possibility of having real-valued poles, there is a 50 % chance that Φ_j is either 0 or π. For a detailed description on how the poles are generated, see Algorithm 1.

That said, regarding the remaining input parameters, I look at two different sets of hyper-parameters for the BVIC: (*i*) $(\beta, \gamma) = (1, 1)$, and (*ii*) $(\beta, \gamma) = (5, 1)$. Simply speaking, the former puts equal weight on uncertainty, regression, forecasting and generalization, and the latter more on forecasting and generalization. Furthermore, the noise parameter δ was set to 0.1 for this experiment.

Algorithm 1: Data generation for Experiment 2.

```
Initialization of variables;
Set order p;
Set measurement noise parameter δ;
Set number of windows N_w;
Set size of windows S_w;
for j := 1 to N_w do
    for k := 1 to p/2 do
        Randomize magnitude M_j between 0.5 and 1;
        Randomize phase angle Φ_j between 0 and π;
        Compute complex conjugate pole set α_j ± β_j i;
    Compute autoregressive parameters using obtained poles;
    Generate a realization {x_t^j}_{t=1,...,N_w} via (3-1);
    Normalize the obtained process (z-score);
    Compute a realization{y_t^j}_{t=1,...,N_w} via (4-1);
```

Assuming that autoregressive models of higher orders are also capable of forecasting over longer horizons, I initially evaluate the performance of each of the criteria over a forecast horizon $h = p$, i.e., $h \in \{10, 20, 30, 40, 50\}$ for each of the previously mentioned orders p, respectively. Additionally, to analyse the effects of the forecasting horizon h, I conducted experiments where instead $h = \text{ceil}(p/4)$, i.e., $h \in \{3, 5, 8, 10, 13\}$. Furthermore, to prevent a possible lack of observations for training of the BVIC, I increased the sample size to include more samples, thereby decreasing the prediction error. The amount of observations used to train models is derived functionally by $S_{\mathcal{T}} = 4.5(p + h)$. Thus, for $p = h$, the size of the training set becomes $S_{\mathcal{T}} = 9h$. A further split of $S_{\mathcal{T}}$ into training and validation for the BVIC gives a training, validation, and test ratio of 0.8, 0.1, and 0.1, respectively. The results of Experiment 2 are summarized in Table 4-2.

Moreover, to give a clearer image of what orders are being selected by the criteria, a graphical representation of the distributions is showed in Figure 4-3. Here, I have plotted the histograms of all the orders that were selected in all simulations by the BVIC with two different sets of hyperparameters, and AICc.

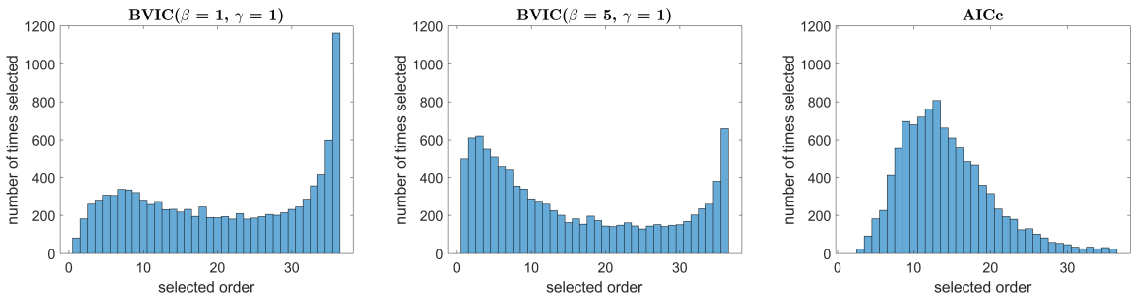


Figure 4-3: Distribution of orders. This figure depicts the distribution of orders selected by the BVIC with $\beta = 1$, the BVIC with $\beta = 5$, and AICc for the case where $p = 30$ in Experiment 2.

Table 4-2: Summary statistics for the Monte Carlo simulations conducted in Experiment 2.

	$h = p$				$h = \text{ceil}(p/4)$			
AR(10) ($p = 10$)	MSE	VAR	$\overline{P}_F$	$\overline{p}$	MSE	VAR	$\overline{P}_F$	$\overline{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.388	0.28	0.34	8.0	0.153	0.09	0.16	7.5
$BVIC(\beta = 5, \gamma = 1)$	0.392	0.28	0.36	6.2	0.161	0.10	0.17	6.0
AIC	0.373	0.26	0.34	7.0	0.150	0.09	0.16	7.2
BIC	0.373	0.26	0.35	4.0	0.147	0.08	0.16	4.1
$AICc$	0.372	0.26	0.34	6.2	0.148	0.09	0.16	6.3
AR(20) ($p = 20$)	MSE	VAR	$\overline{P}_F$	$\overline{p}$	MSE	VAR	$\overline{P}_F$	$\overline{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.352	0.27	0.30	14.4	0.093	0.04	0.10	13.5
$BVIC(\beta = 5, \gamma = 1)$	0.353	0.27	0.31	11.0	0.096	0.04	0.11	10.3
AIC	0.336	0.25	0.30	12.7	0.090	0.03	0.10	12.7
BIC	0.338	0.25	0.31	5.4	0.091	0.03	0.11	5.7
$AICc$	0.334	0.25	0.30	10.9	0.089	0.03	0.10	10.9
AR(30) ($p = 30$)	MSE	VAR	$\overline{P}_F$	$\overline{p}$	MSE	VAR	$\overline{P}_F$	$\overline{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.375	0.28	0.33	20.7	0.085	0.03	0.10	19.2
$BVIC(\beta = 5, \gamma = 1)$	0.377	0.28	0.34	15.5	0.089	0.04	0.10	14.5
AIC	0.358	0.26	0.33	16.8	0.083	0.03	0.10	17.0
BIC	0.359	0.25	0.34	7.1	0.083	0.03	0.10	6.8
$AICc$	0.356	0.25	0.33	14.3	0.082	0.03	0.10	14.4
AR(40) ($p = 40$)	MSE	VAR	$\overline{P}_F$	$\overline{p}$	MSE	VAR	$\overline{P}_F$	$\overline{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.381	0.27	0.33	27.5	0.061	0.01	0.08	24.4
$BVIC(\beta = 5, \gamma = 1)$	0.385	0.28	0.35	20.3	0.063	0.01	0.08	17.8
AIC	0.364	0.26	0.33	21.0	0.058	0.01	0.07	20.3
BIC	0.371	0.26	0.35	8.2	0.060	0.01	0.08	8.1
$AICc$	0.363	0.25	0.34	17.8	0.057	0.01	0.07	17.1
AR(50) ($p = 50$)	MSE	VAR	$\overline{P}_F$	$\overline{p}$	MSE	VAR	$\overline{P}_F$	$\overline{p}$
$BVIC(\beta = 1, \gamma = 1)$	0.362	0.26	0.31	32.1	0.058	0.01	0.07	29.6
$BVIC(\beta = 5, \gamma = 1)$	0.368	0.27	0.33	23.4	0.061	0.01	0.08	21.5
AIC	0.339	0.24	0.31	23.5	0.055	0.01	0.07	22.7
BIC	0.351	0.24	0.33	8.9	0.057	0.01	0.08	9.2
$AICc$	0.339	0.24	0.31	20.1	0.054	0.01	0.07	19.3

Table notes: this table contains summary statistics for each information criterion for higher model orders and larger forecast horizons. I conducted one-way analysis of variance (ANOVA) tests and Kruskal-Wallis (KW) tests to assess if the error distributions found in the experiments were statistically distinguishable. Specifically, I could not find any statistically significant difference based on the one-way ANOVA and KW tests with a 0.05 significance level. These findings are further detailed in Appendix A-1.

4-4-3 Experiment 3

Hereafter, in Experiment 3 I test the predictability of ECoG data during epileptic events. As such, I analyse the ability of the previously discussed criteria to forecast sections of data corresponding to seizures and non-seizures.

Similarly to the previous experiments, sections are extracted from the time series that are subsequently split up into windows. These windows have a total of $N = 1000$ samples such that I can effectively use 800 (80%) samples for training, 100 (10%) for validation, and 100 (10%) samples for testing. The choice of N comes from a sensitivity analysis that I present in detail in Appendix A-2. Next, I perform statistical tests to assess the stationarity of the ECoG recordings used in this experiment. I found that a sample size of $N = 1000$ is a suitable amount that results in sufficient evidence for stationarity in the majority of the considered windows.

Differences in scaling are found in recordings from different patients and between ictal and interictal phases of a single patient. Therefore, to facilitate a fair comparison among the different data, I use the mean absolute scaled error (MASE) [32] as a metric to compare the results from this experiment. The advantage of using the MASE over metrics such as MSE is that the prior is scale-independent. The mean absolute scaled error is formulated as

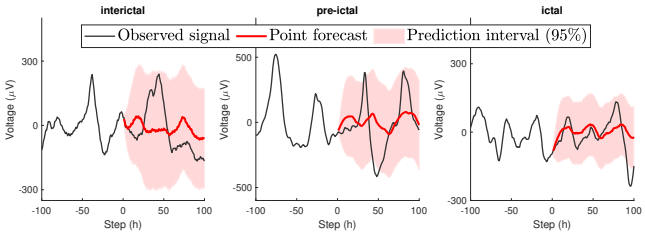
$$\text{MASE} = \frac{\frac{1}{h_2-h_1+1}\sum_{t=n+h_1}^{n+h_2}|Y_t-\hat{Y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t-Y_{t-1}|}, \tag{4-6}$$

where $n \in \mathbb{N}$ is the number of samples in the training set, $h_1, h_2 \in \mathbb{N}$ are variables to indicate the range of forecast steps to include in the metric with $h_2 \geq h_1$. For instance, if $h_1 = h_2$, the MASE is computed for a single horizon forecast. $Y_t \in \mathbb{R}$ and $\hat{Y}_t \in \mathbb{R}$ are the observed and the predicted values of the time series, respectively.

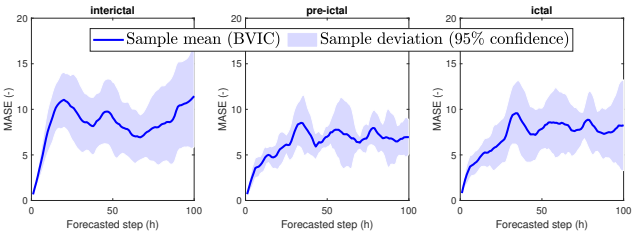
Simply speaking, the MASE is constructed by dividing the mean absolute error (MAE) by the average naïve forecast computed in-sample. Thus, for a single forecast horizon, a MASE of less than one implies that the forecast is better than the average in-sample naïve forecast.

Furthermore, I establish that the performance of the BVIC in comparison to other information criteria is similar, without any statistically significant difference in the majority of the simulations – see details in Appendix A-1. Therefore, for this experiment I consider only the BVIC with $\beta = 1$, and $\gamma = 1$, to assess the ability of the BVIC to forecast electrocorticography data. Also, in Appendix C, the reader can find a comparison in the forecasting error (MASE) obtained the benchmark criteria, relative to the BVIC. Here, one can see that the performance of the criteria is similar.

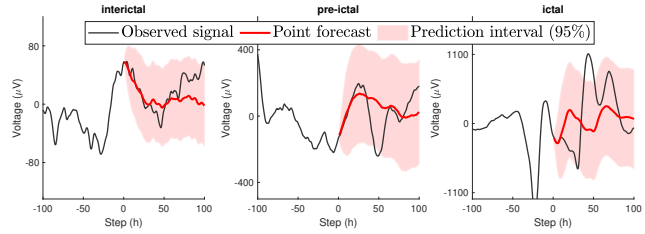
The results of the experiment are plotted in Figure 4-4. Here, I plotted the average MASE over the channels in which a seizure was identified for single forecast horizons ranging from 1 to 100 steps into the future. Figure 4-4a, 4-4c, and 4-4e contain a sample forecast with red color, while Figure 4-4b, 4-4d, and 4-4f depict in blue the average MASE with the variance among different channels. Therefore, it is worth noticing that the red shading in Figure 4-4a, 4-4c, and 4-4e indicates a prediction interval, and shows the estimated interval in which the forecasted observation is within 95% certainty. Whereas, the blue shading in Figure 4-4b, 4-4d, and 4-4f is simply showing the interval in which 95% of the computed values are contained (i.e., $\pm 1.96\sigma$).



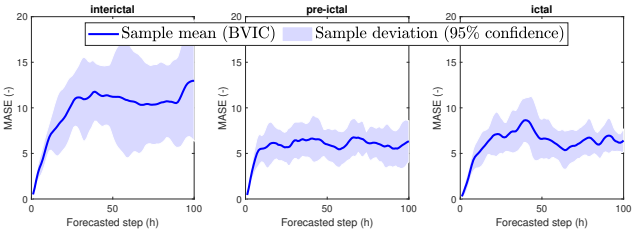
(a) Sample forecast of patient *study 68* from the Hospital of the University of Pennsylvania.



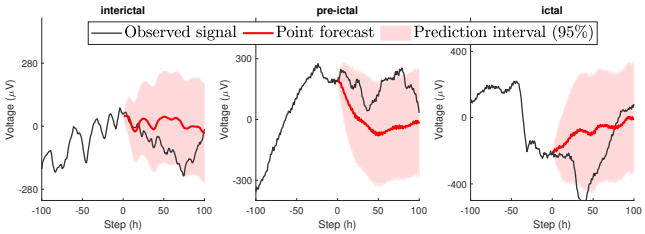
(b) MASE distribution of patient *study 68* from the Hospital of the University of Pennsylvania.



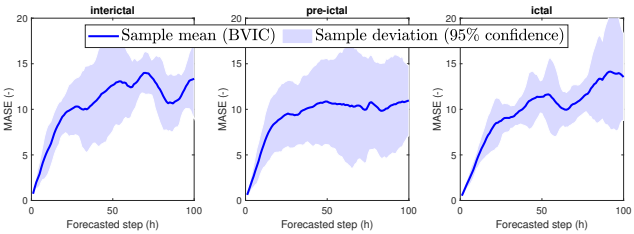
(c) Sample forecast of patient *study 86* from the Hospital of the University of Pennsylvania.



(d) MASE distribution of patient *study 86* from the Hospital of the University of Pennsylvania.



(e) Sample forecast of patient *study 016* from the Mayo Clinic.



(f) MASE distribution of patient *study 016* from the Mayo Clinic.

Figure 4-4: Interictal, pre-ictal, and ictal forecast error. Comparison of mean absolute scaled error (MASE) obtained by the BVIC for 1- to 100-step ahead forecasts. The solid blue line indicates the mean MASE across all channels. The blue shaded areas indicate the range containing the deviations along the considered channels, with a 95% certainty. Additionally, sample forecasts with point forecast and prediction intervals are depicted in (a), (c), and (e).

Chapter 5

Discussion

I introduced a principled analysis of an information criterion that utilizes theoretical principles of time-reversibility and time series to assemble a finite-sample data-driven approach to model selection that eliminates the penalization of the model order and replaces it with a backward validation scheme that can be tuned to trade-off between uncertainty, regression, generalization and forecasting.

Information criteria performance. Experiment 1 explores pedagogical examples to capture the behavior of the different information criteria when different pole locations and signal-to-noise ratios of the time series are considered. It is possible to notice that these impact the performance of the BVIC relative to the other information criteria. For instance, from *Case 1* I notice that for systems with a large real-valued pole ($z_i = 0.9$), the BVIC predicts with larger error compared to the other criteria. On the other hand, when all poles have small absolute values ($|z| < 0.6$), such as in *Case 4*, I observe a relative decrease in prediction error of the BVIC, especially when the signal-to-noise ratio is larger than zero.

Furthermore, considering the two sets of hyperparameters of the BVIC, I also detect a disparity when it comes to the phase angle of the dominant pole(s). Specifically, for systems with high frequency poles, such as in *Case 2* and *Case 3*, I notice that in the case where $\beta = 1$, the BVIC generally obtains lower prediction error as opposed to the case where $\beta = 5$. Contrarily, the opposite seems to be true when the frequency of the poles is small, as can be seen in *Case 1* and *Case 4*.

Moreover, I find that the BVIC selects larger orders than the other criteria, on average. Here, I noticed is that the BVIC shows a certain consistency over all the simulations. For $\beta = 1$, the BVIC finds approximately the true order of 5. Whereas for $\beta = 5$, the average order selected is roughly 3.6. On the contrary, the AIC, BIC, and AICc all have much larger variance in their average selected orders. Thus, in contrast with the three mainstream criteria, the BVIC is more consistent in selecting the order, independent of the location of the poles and the variance of the measurement noise.

In Experiment 2, I provide converging evidence that the different information criteria perform in a similar fashion to the BVIC. Nonetheless, it is important to emphasize that the BVIC provides a principled method that relies on finite samples and offers a trade-off between uncertainty, regression, generalization and forecasting, opposite to information criteria (AIC, BIC, and AICc) where the penalization term is fixed to satisfy asymptotic properties [19]. Specifically, based on statistical tests, namely the one-way analysis of variance and the Kruskal-Wallis test, I found that *none* of the obtained distributions were significantly different between the criteria at a significance level of 0.05.

Model order selection in autoregressive models with the BVIC. It is interesting to notice how the BVIC is able to capture a different range of orders across the different synthetically generated data – see Figure 4-3. Remarkably, I also notice that the BVIC selects orders that are, on average, closer to the true order of the synthetic process – see Table 4-1 and Table 4-2. Thus, providing evidence that the BVIC may be a preferable method when it comes to estimating the true order of an autoregressive model. Lastly, the BVIC is also able to adapt to a changing forecast horizon, where a shorter forecast horizon means that the BVIC selects, on average, lower orders. On the other hand, the selection of the orders by the remaining information criteria is not influenced by the forecast horizon.

Given that the BVIC selects model orders that are, on average, closer to the true order of the model, in Experiment 3, I tested the ability of the BVIC to assess the *memory order* (i.e., the statistical significant dependency or previous realizations of the time series) in the context of seizure prediction [9].

Furthermore, there has been long reported evidence that the memory order increases during the ictal state [33, 34]. Implicitly, an increase in memory would also indicate an increase in the number of steps for which one can forecast ahead. In Experiment 3, I collected converging evidence towards the later points, as the predictability increases during the pre-ictal and ictal state compared with the interictal state.

Nonetheless, it is worth reporting that this is not always the case, as can be seen in Figure 4-4f, where there is no significant decrease in error between the different states. The reason for the irregularity is something that would require more study. However, there are a few possible causes that may attribute to this outcome. First of all, I provided evidence for the stationarity of the three recordings that were used in the experiment – see Appendix A-2. Nevertheless, the recording from patient study 016 showed the weakest evidence for stationarity. Thus, certain amount of non-stationarity in the data could ascribe to the differing results seen in Figure 4-4f. Secondly, following up from the previous argument, I must consider that it might not be possible to predict seizures with a single framework due to the variety of mechanisms that underlie an epileptic seizure. Finally, considering the fact that the average MASE obtained for study 016 is similar for all states, it could be that the electrodes that were identified as seizure electrodes were not actually placed in the location of the seizure, or, the entire recording was incorrectly classified as a seizure [35].

Extensions and future work. Whereas I have focused on the univariate autoregressive models, it would be interesting to extend the BVIC to a multivariate setting. This may reveal to be beneficial when the underlying dynamics captured by a multivariate time series have

spacial dependencies. For instance, this is the case of ECoG recordings explored in Experiment 3, where there is evidence that a seizure propagates through the brain and reveals itself across different channels over time. Additionally, it would be worth to expand the proposed information criteria to handle both moving average and fractional integrative models known to be able to handle long-term memory [1]. Finally, another future research direction would be to establish a method to merge the process of model identification and parameter estimation for the BVIC. The BVIC can be formulated as a multi-objective optimization problem where the both the model order and the model parameters need to be solved simultaneously. Solving this problem could potentially improve the quality of models selected by the BVIC and, therefore, could improve forecasts.

Chapter 6

Conclusion

In my thesis, I assessed the possibility of using the property of time-reversibility to formulate a novel model selection criterion that uses a data-driven method to quantify forecast performance among models for locally stationary, finite sample time series. First, I provided theoretical evidence for the use of a backwards validation scheme to replace the order penalization. Then, I introduce the backwards validation information criterion (BVIC), a data-driven method that accommodates different functionalities with respect to forecasting based on the selected hyperparameters, formulated as a weighted optimization problem. Here, the hyperparameters are the weights that determine whether the criterion more heavily penalizes the regression, generalization (point prediction error), or uncertainty (prediction interval) component of a forecast.

I conducted thorough statistical experiments to examine the performance of the proposed criterion. The results suggest that the BVIC has comparable performance to conventional information criteria. Specifically, in most of the experiments performed, I did not find statistically significant differences between the forecast error of the BVIC under certain parameterizations and that of the different information criteria. As such, the BVIC can be considered as a suitable data-driven criterion for forecasting locally stationary, finite sample, autoregressive time series, eliminating the need for a static penalization of the model order.

Appendix A

Statistical Tests

Two diagnostic tests are conducted. Firstly, I test whether the obtained results for each criterion is obtained from the same distribution, or not. Additionally, for the data used in Experiment 3, I evaluate whether the windows used are stationary.

A-1 Statistical tests for difference among samples

I perform two tests to evaluate whether there is a statistically significant difference in the error distributions obtained in Experiment 2. Specifically, I compare the single step forecast error distributions resulting from the information criteria in the AR(30) process. The tests performed are as follows:

One-way analysis of variance (ANOVA) test. Assesses the null hypthesis that the samples of a set of groups (i.e., information criteria) are drawn from populations with the same mean.

Kruskal-Wallis (KW) test. Assesses the null hypothesis the samples of a set of groups (i.e., information criteria) are drawn from the distributions with the same median.

In Figure A-1, I plotted the results of the ANOVA test and the KW test into a box plot that depicts the distribution of p-values for 100 Monte Carlo simulations, for forecast horizons ranging from 1 to 30. Here, for both ANOVA and KW, the p-values do not drop below the 0.05 significance for any of the forecast horizons. Thus, there is not enough statistical evidence in support of rejecting the null hypotheses.

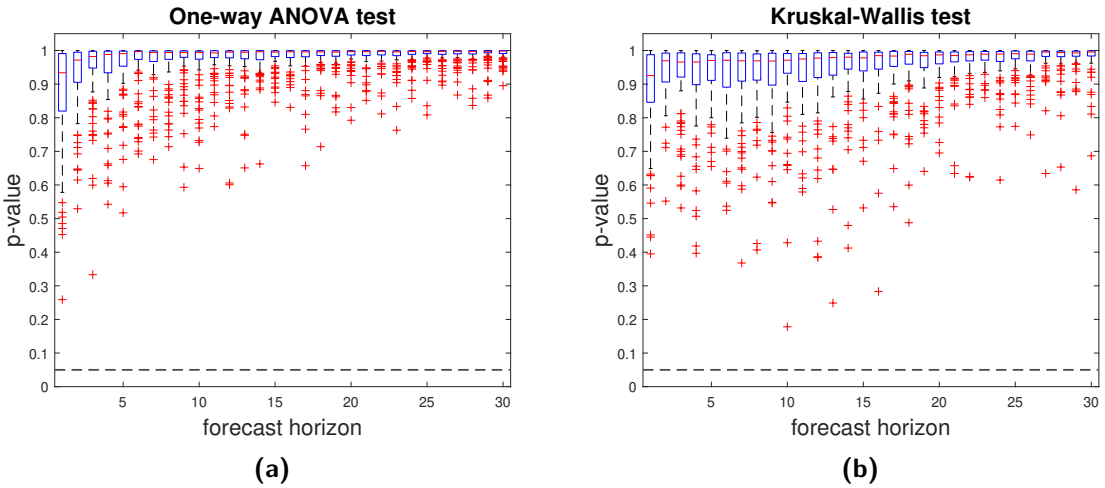


Figure A-1: Statistical tests. Box-plot depicting the spread of p-values obtained by conducting one-way ANOVA and Kruskal-Wallis tests on 100 distributions for all single forecasts horizons.

A-2 Statistical tests for stationarity

An implicit assumption of the BVIC (and all other criteria) is stationarity of the time series. That said, ECoG recordings are said to be approximately stationary over limited periods of time [9]. In this appendix I conduct statistical tests for stationarity to determine what would be a favorable window size. Two hypothesis tests are conducted, where one tests for stationarity while the other tests for non-stationarity. Specifically, the tests conducted are as follows:

Leybourne-McCabe (LMC) test. Assesses the null hypthesis that a time series is a trend stationary AR(p) process. H_0: Y_j is stationary.

Phillips-Perron (PP) test. Assesses the null hypothesis that a unit root is present in a time series. H_0: Y_j is non-stationary.

Ultimately, a failure to reject the null hypothesis of the LMC test and a rejection of the PP test null hypothesis at a 0.05 significance level would be enough evidence to suggest that the time series is stationary. Therefore, I conduct the previously mentioned statistical test on the three different ECoG recordings, where each recording is divided into windows of length N. Over all the windows, I evaluate the percentage of time in which the null hypothesis is rejected, i.e., $r_{H_0} = \frac{\#\text{times } H_0 \text{ is rejected}}{\text{total windows}}$. The results are shown in Figure A-2. Noticeably, the ratio of rejection converge to the preferred outcome. However, it seems that for the patient study 16 from the Mayo Clinic we see less evidence to reject non-stationarity, and more evidence to reject stationarity.

In summary, one may be inclined to choose N as large as possible. However, there is a limit to the length of sections of stationarity that are usable. This is due to shifting states of the brain, from interictal to pre-ictal, and from the latter to ictal. Commonly, the duration of the pre-ictal state constitutes of around 10 to 20 thousands samples. Consequently, I would

like to limit the size of the windows such that I can still have enough windows to perform simulations on. Therefore, I have chosen a window size of $N = 1000$ to be suitable for the data that I am considering, even though there is some evidence for non-stationarity in one of the recordings.

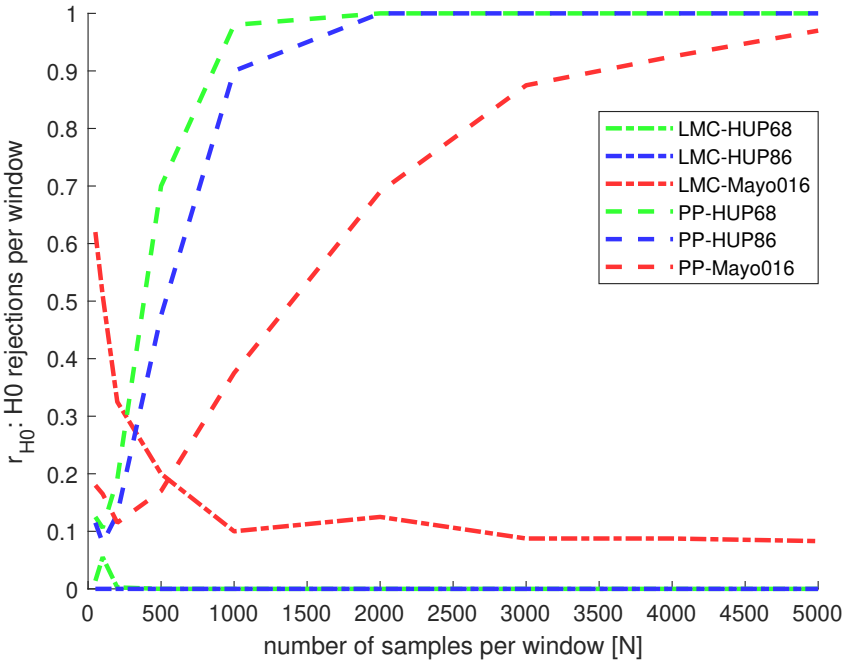


Figure A-2: Stationarity tests. Leybournce-McCabe and Phillips-Perron test outcomes represented as the fraction of rejections over the total amount of tests performed for each of the ECoG recordings evaluated in Experiment 3.

Appendix B

Least-squares Estimation

Let us pose the autoregressive process of order p as a dynamical system, consider the state space form

$$x_{k+1} = Ax_k + \varepsilon_k, \tag{B-1}$$

where $x_k, \varepsilon_k \in \mathbb{R}^p$ are vectors containing lagged values of the state and noise, respectively. The system matrix is denoted by $A \in \mathbb{R}^{p \times p}$. Writing out the equation into matrix form results in

$$\begin{bmatrix} x(k+1) \\ x(k) \\ \vdots \\ x(k-p+1) \end{bmatrix} = \underbrace{\begin{bmatrix} -a_1 & -a_2 & -a_3 & \dots & -a_p \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}}_{A} \begin{bmatrix} x(k) \\ x(k-1) \\ \vdots \\ x(k-p) \end{bmatrix} + \begin{bmatrix} \varepsilon(k+1) \\ \varepsilon(k) \\ \vdots \\ \varepsilon(k-p+1) \end{bmatrix}. \tag{B-2}$$

Note that the A matrix is written in companion form, resulting in the autoregressive parameters to be contained in the top row of the matrix.

To find the A matrix a least-squares problem is formulated that minimizes ε_k as follows:

$$\min ||\varepsilon_k||^2 = \min_A ||x_{k+1} - Ax_k||^2 \tag{B-3}$$

To better capture the dynamical behaviour it is best to increase the amount of data used for the least squares estimator. Therefore, variables x_{k+1} and x_k are used to construct Hankel matrices $H_{x,k+1}$ and $H_{x,k}$. The least squares problem becomes

$$||\underbrace{\begin{bmatrix} x(k+1) & x(k) & \dots & x(p+1) \\ x(k) & x(k-1) & \dots & x(p) \\ \vdots & \vdots & \ddots & \vdots \\ x(k-p+1) & x(k-p) & \dots & x(1) \end{bmatrix}}_{H_{x,k+1}} - A \underbrace{\begin{bmatrix} x(k) & x(k-1) & \dots & x(p) \\ x(k-1) & x(k-2) & \dots & x(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ x(k-p) & x(k-p-1) & \dots & x(0) \end{bmatrix}}_{H_{x,k}} ||_F^2. \tag{B-4}$$

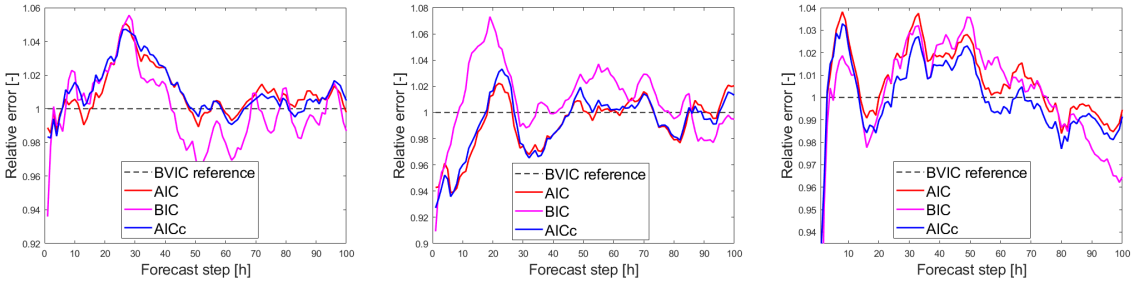
The solution of the least squares problem can be denoted as

$$\hat{A} = H_{x,k+1} H_{x,k}^\top (H_{x,k} H_{x,k}^\top)^{-1}. \tag{B-5}$$

Appendix C

Additional Experimental Results

This appendix shows the relative differences in forecast error of the considered information criteria on ictal data from Experiment 3. Figure C-1 depicts the error over a forecasting horizon of 100 steps, normalized with respect to the error obtained by the BVIC. This is done to more clearly show the differences between the criteria.



(a) Study 68 (Hospital of the University of Pennsylvania)

(b) Study 86 (Hospital of the University of Pennsylvania)

(c) Study 016 (Mayo Clinic, Rochester)

Figure C-1: Relative forecast error. Relative error of each of the benchmark criteria compared to the BVIC when forecasting *ictal* data from different patient studies.

Bibliography

[1] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[2] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*. Springer, 2017.

[3] C. Chatfield, *Time-series forecasting*. CRC press, 2000.

[4] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting*. Springer, 2016.

[5] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

[6] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*, pp. 199–213, Springer, 1998.

[7] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[8] B. E. Hansen, "Multi-step forecast model selection," in *20th Annual Meetings of the Midwest Econometrics Group*, 2010.

[9] Y. Murin, A. Goldsmith, and B. Aazhang, "Estimating the memory order of electrocorticography recordings," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2809–2822, 2019.

[10] A. D. McQuarrie and C.-L. Tsai, *Regression and time series model selection*. World Scientific, 1998.

[11] M. J. Brewer, A. Butler, and S. L. Cooksley, "The relative performance of aic, aicc and bic in the presence of unobserved heterogeneity," *Methods in Ecology and Evolution*, vol. 7, no. 6, pp. 679–692, 2016.

[12] B. Billah, R. J. Hyndman, and A. B. Koehler, "Empirical information criteria for time series forecasting model selection," *Journal of Statistical Computation and Simulation*, vol. 75, no. 10, pp. 831–840, 2005.

[13] M. Verhaegen and V. Verdult, *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

[14] J. Lamperti, *Stochastic processes: a survey of the mathematical theory*, vol. 23. Springer Science & Business Media, 2012.

[15] I. Florescu, *Probability and stochastic processes*. John Wiley & Sons, 2014.

[16] G. S. Grimmet *et al.*, *Probability and random processes*. Oxford University Press, 2020.

[17] K. I. Park and Park, *Fundamentals of Probability and Stochastic Processes with Applications to Communications*. Springer, 2018.

[18] G. E. Box and G. C. Tiao, *Bayesian inference in statistical analysis*, vol. 40. John Wiley & Sons, 2011.

[19] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 6, pp. 16–34, 2018.

[20] S. Vijayakumar, "The bias–variance tradeoff," *University Edinburgh*, 2007.

[21] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[22] J. Kuha, "AIC and BIC: Comparisons of assumptions and performance," *Sociological methods & research*, vol. 33, no. 2, pp. 188–229, 2004.

[23] C. Granger and Y. Jeon, "Forecasting performance of information criteria with many macro series," *Journal of Applied Statistics*, vol. 31, no. 10, pp. 1227–1240, 2004.

[24] P. Mantalos, K. Mattheou, and A. Karagrigoriou, "Forecasting arma models: a comparative study of information criteria focusing on MDIC," *Journal of Statistical Computation and Simulation*, vol. 80, no. 1, pp. 61–73, 2010.

[25] R. J. Hyndman *et al.*, "Another look at forecast-accuracy metrics for intermittent demand," *Foresight: The International Journal of Applied Forecasting*, vol. 4, no. 4, pp. 43–46, 2006.

[26] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[27] A. Lawrance, "Directionality and reversibility in time series," *International Statistical Review/Revue Internationale de Statistique*, pp. 67–79, 1991.

[28] F. J. Breidt and R. A. Davis, "Time-reversibility, identifiability and independence of innovations for stationary time series," *Journal of Time Series Analysis*, vol. 13, no. 5, pp. 377–390, 1992.

[29] J. B. Wagenaar, B. H. Brinkmann, Z. Ives, G. A. Worrell, and B. Litt, "A multimodal platform for cloud-based collaborative research," in *2013 6th international IEEE/EMBS conference on neural engineering (NER)*, pp. 1386–1389, IEEE, 2013.

[30] A. N. Khambhati, K. A. Davis, B. S. Oommen, S. H. Chen, T. H. Lucas, B. Litt, and D. S. Bassett, "Dynamic network drivers of seizure generation, propagation and termination in human neocortical epilepsy," *PLoS computational biology*, vol. 11, no. 12, p. e1004608, 2015.

[31] B. Litt, R. Esteller, J. Echauz, M. D'Alessandro, R. Shor, T. Henry, P. Pennell, C. Epstein, R. Bakay, M. Dichter, *et al.*, "Epileptic seizures may begin hours in advance of clinical onset: a report of five patients," *Neuron*, vol. 30, no. 1, pp. 51–64, 2001.

[32] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The m4 competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020.

[33] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, "Early-warning signals for critical transitions," *Nature*, vol. 461, no. 7260, pp. 53–59, 2009.

[34] V. K. Jirsa, W. C. Stacey, P. P. Quilichini, A. I. Ivanov, and C. Bernard, "On the nature of seizure dynamics," *Brain*, vol. 137, no. 8, pp. 2210–2230, 2014.

[35] A. Ashourvan, S. Pequito, A. N. Khambhati, F. Mikhail, S. N. Baldassano, K. A. Davis, T. H. Lucas, J. M. Vettel, B. Litt, G. J. Pappas, *et al.*, "Model-based design for seizure control by stimulation," *Journal of neural engineering*, vol. 17, no. 2, 2020.

Glossary

List of Acronyms

BVIC	backwards validated information criterion
AIC	Akaike's information criterion
BIC	Bayesian information criterion
EEC	earliest electrographic change
UEO	uneqiovocal electrographic onset
MASE	mean absolute scaled error
AR	autoregressive
ARMA	autoregressive moving-average
SARMA	seasonal autoregressive moving-average
ARIMA	autoregressive integrated moving-average
ARFIMA	autoregressive fractionally integrated moving-average

List of Symbols

β	Penalty weighting on the prediction error in the BVIC
δ	Measurement noise variance
γ	Penalty weighting on the prediction variance in the BVIC
Γ_n	Autocovariance matrix
γ_n	Vector of autocovariances up to n lags
$\mathcal{F}$	Set of possible events
$\mathcal{M}$	Set of candidate models
Ω	Sample space
e_t	Forecast error at time step t

$F(x)$	Probability distribution function of a specified model
$f(x)$	Specified model
$G(x)$	True probability distribution
h	Prediction horizon
k	Number of model parameters
N	Total sample size
n	Sample size available for training
P	Probability measure
p	Model order
P^n_{1-t}	Prediction variance at time step $1-t$
I	K-L information