Model selection and Bayesian nonparametrics for time series and non-standard regression models

by

Gudmund Horn Hermansen

THESIS

Dissertation presented for the degree of

PHILOSOPHIÆ DOCTOR

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 1547

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

To my family.

Contents

Chapter 1

Introduction to the Thesis

This thesis studies certain aspects of model selection and Bayesian nonparametrics in time series models, types of non-standard regression models and in function estimation. In this regard, with the particular models and applications set aside, there are two main themes, those of model selection and Bayesian nonparametrics. It is tempting to view these as opposites. Model selection is essentially frequentistic, data driven and typically belongs to the realm of finite-dimensional models. On the other hand, we have Bayesian nonparametrics, which is concerned with high- and infinite-dimensional objects and by design intended to include subjective elements in the analysis. Moreover, these high-dimensional models are valid under very general conditions, seemingly making model selection irrelevant.

In traditional parametric modelling we usually consider a finite number of models, each with a fixed and low number of parameters. The use of model selection is therefore often very relevant as a guide to determine the appropriate level of complexity that is required. Among other things, we wish to avoid so-called over/under fitting issues which result from a mismatch in the model complexity compared to the underlying truth. In general, such model selection strategies are data-driven with the intention of letting data 'speak for themselves' in order to find the model, among the carefully selected candidates, that 'best' describes the observations.

The idea of making data responsible for selection, makes most model selection strategies reminiscent of classic frequentist ideas. There are some Bayesian approaches, but these are rarely proper Bayesian and usually push proper prior specification out of the discussion by quickly introducing flat, or non-informative, prior distributions wherever it may be needed.

The Bayesian nonparametric label is commonly associated with a Bayesian approach to classical (frequentist) nonparametric modelling; a less restrictive class of models with a high level of flexibility. As a consequence of the Bayesian paradigm,

they are intended to be subjective. Also, if properly specified, they tend to be more transparent and honest, through the process of prior specifications, with regards to the underlying model assumptions. Describing prior distributions on such large objects is 'risky business' and seemingly reasonable constructions may result in a meaningless posterior distributions without any real or practical use. There are some strategies invented to uncover and prevent this from happening in practice, however, e.g. frequentistic justification, such as posterior consistency and convergence rates, and also so-called Bernshteĭn–von Mises theorems. This provides good theoretical large-sample insurance and are among the important topics of this thesis.

The nonparametric modelling approach suggests that by working with such large and flexible objects we may avoid the model selection phase altogether, at least if the sample size is sufficiently large. The origin of this view probably stems from the built-in flexibility and 'model free' construction, which is commonly believed to sufficiently sort out all the necessary details and find an appropriate fit – all by itself.

In small samples, the more traditional and finite-dimensional parametric models become important tools to be able to do anything at all, since large nonparametric structures are not necessarily possible to fit in such cases. There is also a general discomfort in fitting models that are more complex than what is actually needed. The adaptable nonparametric machinery is therefore clearly not always the best answer. The Bayesian approach to nonparametric modelling can provide just the right amount of structure needed for the model to behave reasonably across its domain, even in small samples. At the same time, the nonparametric models are flexible enough to adapt as the number of observations increases and information starts to accumulate.

As a last comment, there is a general concern related to how much one should rely on data in general, with clear implications for both model selection methodology and nonparametric modelling. This is essentially an open question and we do not intend to give the complete solution here. In the work presented in this thesis, there is an underlying assertion that proper structure, through prior or focused modelling, is indeed often needed.

The general introduction now follows with a summary of time series modelling in the frequency domain, Bayesian nonparametrics and some basic model selection methodology in Sections 1–3. This provides the necessary background and will also include results needed in the discussion of the four papers that constitute this thesis. These papers are then presented separately in Sections 4–7, each with its own introduction and general summary. In addition, we will also discuss and point at some potential extensions, applications and future work.

1. Time series analysis in the frequency domain

The aim of the present section is to give a short introduction to time series modelling in the frequency domain and to establish some of the main results necessary for the following sections; more complete introductions can be found in Brillinger (1975), Priestley (1981) and Dzhaparidze (1986). As a motivating illustration, we consider the number of skiing days in a winter season, defined as the number of days with at least 25 cm snow, at the particular location of Bjørnholt in Oslo's skiing and recreation area Nordmarka, see Figure 1.1. Besides being of great interest to skiing enthusiasts, this number is a good indicator of how cold a winter is and is also an indication of the general temperature over a given period of time. To understand the underlying dynamics, the estimated dependency structure is of interest. Moreover, the potential interaction, or joint effect, of trend and dependency (if any) has implications, especially for predictions. Such considerations call out for model selection and assessments tools, since we will be needing proper methodology to decide which model, e.g. with or without a decreasing trend, is 'best' and should be used for making inference; we will return to such questions in Sections 3, 5 and 7 below.

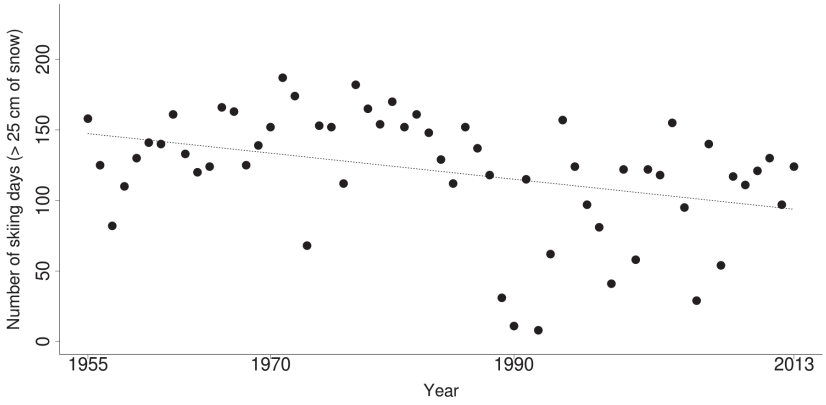

FIGURE 1.1. The number of skiing days for the winter seasons 1954-55 to 2012-13 at Bjørnholt, a location in the countryside just outside Oslo. The global linear trend (dotted line) is seen to be decreasing with estimated slope about -0.9, indicating that the number of skiing days has on average declined by almost one day each year since the mid 1950s.

Let $y_1, \ldots, y_n$ be realisations from a stationary Gaussian time series $\{Y_t\}$ with mean zero. These models are completely defined by their dependency structure, which in the time domain is given by the covariance function $C(h)$, for all lags

$h = 0, 1, 2, \ldots$. The covariance has an alternative and unique representation in the so-called frequency domain by the Fourier transform, which in the current framework of real-valued time series simplifies to

$$C(h) = \int_{-\pi}^{\pi} \cos(\omega h) g(\omega) \, \mathrm{d}\omega, \quad \text{where } g(\omega) = \frac{C(0)}{2\pi} + \frac{1}{\pi} \sum_{h=1}^{\infty} C(h) \cos(\omega h) \tag{1.1}$$

is the spectral density with corresponding cumulative function G referred to as the spectral measure; see among others Brillinger (1975) or Priestley (1981).

We will often prefer to work with the frequency representation. There are various reasons for this, but of most importance here is that it is much easier to construct functions within the frequency domain that results in suitable covariance functions. For a function $C(h)$ to be a proper covariance function it must be positive-semidefinite, which means that for all $n \geq 1$ the following holds: $a^{\mathrm{t}} \Sigma_n a \geq 0$ for all vectors $a \in \mathbb{R}^n$, where Σ_n is the covariance matrix with elements $C(|i - j|)$, for $i, j = 1, \ldots, n$. This condition, which is often difficult to verify directly, is e.g. required to ensure that there will always be non-negative variances. The conditions for positive-semidefiniteness are much easier to verify in the frequency domain, however, which are establish in the following theorem.

Theorem 1. *(Priestley, 1981, Wold's Theorem) A necessary and sufficient condition for $C(h)$, for $h \geq 0$, to be a covariance function for some real valued stationary process $\{Y_t\}$ is that there exists a non-decreasing function G on the interval $(-\pi, \pi)$ such that*

$$C(h) = \int_{-\pi}^{\pi} \cos(\omega h) \, \mathrm{d}G(\omega), \quad \textit{for all } h \geq 0,$$

and where $G(-\pi) = 0$ and $G(\pi) = G_\pi < \infty$.

For this reason we will often write

$$C_g(h) = \int_{-\pi}^{\pi} \cos(\omega h) \, \mathrm{d}G(\omega) = 2 \int_{0}^{\pi} \cos(\omega h) g(\omega) \, \mathrm{d}\omega, \quad \text{for } h \geq 0, \tag{1.2}$$

where the last equality follows under the assumption that G has a spectral density g as its derivative. This is not always true, but if not otherwise stated we will work under the condition that G has a spectral density g, which is further assumed to be at least Lipschitz continuous.

1.1. Maximum likelihood estimation and model misspecification. The purpose of this section is to introduce the basic properties of the maximum likelihood estimator in a misspecified modelling framework, which is especially important for the derivation of Akaike's information criterion (AIC; Akaike (1973)) in Section 7. As

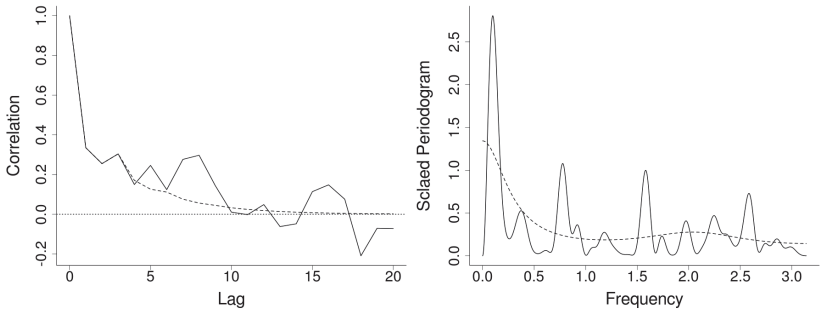

FIGURE 1.2. The estimated correlation function (solid line), with $\widehat{C}(0) = (24.16)^2$, for the Bjørnholt series in Figure 1.1 (left panel) In the right panel, the periodogram (solid line), a nonparametric estimate for the spectral density (see Section 1.2 below), is plotted against the spectral density for the fitted autoregressive process of order three (dashed line), see e.g. Brockwell & Davis (1991) for introduction and conditions. The frequency estimates are scaled to integrate to one, i.e. the analogue of (1.2) for the correlation function.

a gentle start and motivation, we will first discuss estimation under the assumption that the model is correctly specified.

Let f_θ, with $\theta \in \Theta \subset \mathbb{R}^p$ for a finite p, be a parametric spectral density function, the corresponding Gaussian log-likelihood is then

$$\ell_n(f_\theta) = -\tfrac{1}{2}[n\log(2\pi) + \log|\Sigma_n(f_\theta)| + \underline{y}_n^{\mathrm{t}}\Sigma_n(f_\theta)^{-1}\underline{y}_n], \tag{1.3}$$

where $\Sigma_n(f_\theta)$ is the covariance matrix with elements $C_{f_\theta}(|i-j|)$ for $i,j = 1,\ldots,n$, and $\underline{y}_n^{\mathrm{t}} = (y_1,\ldots,y_n)$. The maximum likelihood estimator is then defined as $\widehat{\theta}_n = \arg\max_\theta \ell_n(\theta)$. Suppose the true underlying spectral density f_{θ_0}, for a unique θ_0 in a compact subset of Θ, is bounded away from both zero and infinity and is such that $\sum_{h\le\infty} h|C_{f_{\theta_0}}(h)|^2 < \infty$; a type of short memory condition, see Remark 1 below for some additional comments. Then, as the sample size approaches infinity, the normalised maximum likelihood estimator has the following weak limit

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d U \sim \mathrm{N}_k(0, J(f_{\theta_0})^{-1}), \tag{1.4}$$

in $P_{f_{\theta_0}}$-probability, where

$$J(f_\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \nabla\Psi_{\theta_0}(\omega)\nabla\Psi_{\theta_0}(\omega)^{\mathrm{t}}\,\mathrm{d}\omega, \tag{1.5}$$

with $\Psi_\theta = \log f_\theta$ and $\nabla\Psi_\theta$ as the k-dimensional vector of partial derivatives with respect to θ. Note that (1.4) implies that $\widehat{\theta}_n \to \theta_0$ in $P_{f_{\theta_0}}$-probability; see among others Dzhaparidze (1986) for more details and references.

Remark 1. *By application of an integration by parts argument, it is fairly straightforward to show that the short memory condition* $\sum_{h\leq\infty}|h||C(h)|<\infty$ *holds if the spectral density* g *is both continuous and bounded from above. Moreover, stronger convergence rates of the type* $\sum_{h\leq\infty}|h|^k|C(h)|^\alpha<\infty$*, for finite* $k\geq 1$ *and* $\alpha\geq 1$*, can be shown to follow from the existence of smooth higher order derivatives of* g*, see Carslaw (1921, p. 249) for details.*

Consider the case where $y_1,\ldots,y_n$ are realisations from the stationary time series model with true spectral density g and let f_θ be a parametric spectral density from the class of parametric candidates that do not necessarily span or include the true underlying g, i.e. we are working in a potentially misspecified modelling framework. Then the maximum likelihood estimator $\widehat{\theta}_n$ does not converge to the 'true parameter value', since this does not necessary exist in a misspecified modelling framework. The estimator is instead commonly said to converge to the so-called 'least false parameter value', i.e. $\widehat{\theta}_n\to_{P_g}\theta_0=\arg\min_\theta d(g,f_\theta)$, where

$$d(g,f_\theta)=-\frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log\frac{g(\omega)}{f_\theta(\omega)}+1-\frac{g(\omega)}{f_\theta(\omega)}\right)\mathrm{d}\omega, \tag{1.6}$$

see Dahlhaus & Wefelmeyer (1996) for details. Note that d is positive and fulfills $d(g,f_\theta)=0$ in and only if g is equal to f_θ almost everywhere. The main large-sample properties for $\widehat{\theta}_n$ in the present framework are summarised in the theorem below.

Theorem 2. *(Dahlhaus & Wefelmeyer, 1996, Theorem 3.3) Assume that the true spectral density* g *is Lipschitz-continuous and that* g *and* f_θ *are bounded away from zero and infinity. If the least false parameter value* $\theta_0=\arg\min d(g,f_\theta)$ *is a unique solution in a compact subset* $\Theta\subset\mathbb{R}^p$*, with* p *finite, and* f_θ *is two times differentiable with respect to* θ*, with derivatives that are continuous in both* θ *and* ω *and uniformly bounded in a neighbourhood around* θ_0*. Then*

$$\sqrt{n}(\widehat{\theta}_n-\theta_0)\to_d J(g,f_{\theta_0})^{-1}U,\qquad \text{where } U\sim\mathrm{N}(0,K(g,f_{\theta_0})),$$

with J *and* K *defined by*

$$J(g,f_{\theta_0})=\frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\nabla\Psi_{\theta_0}(\omega)\nabla\Psi_{\theta_0}(\omega)^{\mathrm{t}}g(\omega)+\nabla^2\Psi_{\theta_0}(\omega)(f_{\theta_0}(\omega)-g(\omega))\right]\frac{1}{f_{\theta_0}(\omega)}\,\mathrm{d}\omega$$

and

$$K(g,f_{\theta_0})=\frac{1}{4\pi}\int_{-\pi}^{\pi}\nabla\Psi_{\theta_0}(\omega)\nabla\Psi_{\theta_0}(\omega)^{\mathrm{t}}\left[\frac{g(\omega)}{f_{\theta_0}(\omega)}\right]^2\mathrm{d}\omega,$$

where $\Psi_\theta(\omega)=\log f_\theta(\omega)$ *and* $\nabla\Psi_\theta(\omega)$ *and* $\nabla^2\Psi_\theta(\omega)$ *are the vector and matrix of partial derivatives with respect to* θ*, respectively.*

1.2. The Whittle approximation, periodogram and estimating the spectral measure. The Whittle log-likelihood is an approximation to the full Gaussian log-likelihood (1.3), originally suggested by P. Whittle in a series of works (cf. Whittle (1953)) from the 1950s and is defined by

$$\widetilde{\ell}_n(f) = -\frac{n}{2}\bigg\{ \log 2\pi + \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[2\pi f(\omega)]\,\mathrm{d}\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{I_n(\omega)}{f(\omega)}\,\mathrm{d}\omega \bigg\}, \tag{1.7}$$

where

$$I_n(\omega) = \frac{1}{2\pi n}\bigg| \sum_{t\le n} y_t \exp\{i\omega t\}\bigg|^2 \tag{1.8}$$

is the periodogram. The Whittle approximation is close to the full Gaussian log-likelihood in the sense that

$$\ell_n(f) = \widetilde{\ell}_n(f) + O_{P_g}(1) \tag{1.9}$$

uniformly in f, see Coursol & Dacunha-Castelle (1982) for details; see also Dzhaparidze (1986) for a comprehensive introduction and discussion of the Whittle approximation above and related topics. The approximation motivates an alternative estimation procedure, i.e. the Whittle estimator $\widetilde{\theta}_n = \arg\min_\theta \widetilde{\ell}_n(f_\theta)$. Moreover, the large-sample results in (1.4) above is known to be true with the maximum likelihood estimator replaced by $\widetilde{\theta}_n$ (cf. Dzhaparidze (1986)) and similarly the conclusions of Theorem 2 can also be shown to stay true, see Dahlhaus & Wefelmeyer (1996) for details.

The Whittle approximation is useful in several real applications and is also convenient as a tool in large-sample derivations. The reason is that the spectral density is included more directly in its formulation. In comparison, the full Gaussian log-likelihood (1.3) has the spectral density hidden inside the inverse of the covariance matrix, making estimation and derivation of large-sample properties much more complicated; this is e.g. clearly illustrated in Hermansen & Hjort (2014a,b,d).

The periodogram I_n above, originally introduced to find hidden periodicities (Schuster, 1898), is commonly used as a nonparametric estimator for the underlying spectral density. For stationary time series processes with mean zero, it follows from Brillinger (1975, Theorem 5.2.2) that

$$\mathrm{E}_g\, I_n(\omega) = g(\omega) + O(n^{-1})$$

provided $\sum_{h\le n} |h||C_g(h)| < \infty$.

For this reason, the periodogram is commonly used as a basis for estimating the cumulative spectral measure G and related functionals, e.g. the covariance function

(1.2), however, there are two canonical alternatives

$$\widetilde{G}_n(\omega) = 2\int_0^\omega I_n(u)\,\mathrm{d}\omega \quad \text{or} \quad \widehat{G}_n(\omega) = \frac{4\pi}{n}\sum_{u_j\leq\omega} I_n(u_j), \tag{1.10}$$

where $u_j = 2\pi j/n$, for $j = 0,\ldots,m$ and $m = \lfloor n/2\rfloor$. These are both well studied, see among others Taniguchi (1980) for properties regarding $\widetilde{G}_n$, and $\widehat{G}_n$ are extensively discussed in Brillinger (1975, Ch. 5.10). There are good reasons to why we commonly prefer $\widehat{G}_n$, e.g. under the assumption that $\sum_{h\leq n}|h||C_g(h)| < \infty$

$$I_n(u_j) \approx_d g(u_j)\mathrm{E}_j \quad \text{and} \quad \mathrm{Cov}\{I_n(u_j),\, I_n(u_{j'})\} = O(n^{-1}), \tag{1.11}$$

for $u_j = 2\pi j/n$, $j \neq j'$ and $j, j' = 1,\ldots,\lfloor n/2\rfloor$ and where $\mathrm{E}_j \sim \mathrm{Exp}(1)$; see among others Brillinger (1975, Ch. 5) for details. Moreover, we know the process convergence

$$\sqrt{n}\{\widehat{G}_n(\omega) - G(\omega)\} \to_d W\left(2\pi\int_0^\omega g(u)^2\,\mathrm{d}u\right), \quad \text{for } \omega\in[0,\pi], \tag{1.12}$$

where $W(\cdot)$ is a standard Wiener process (this also holds for $\widetilde{G}_n$), see Ibragimov (1963) for a complete derivation. Note that by application of the continuous mapping theorem (cf. Billingsley (2009)), the weak limit result above automatically induces several large-sample properties (and justifies approximations) for continuous functionals of the spectral distribution, e.g. the nonparametric estimate for the covariance function

$$\widehat{C}(h) = \int_{-\pi}^{\pi}\cos(\omega h)\,\mathrm{d}\widehat{G}(u),$$

for $h \geq 0$. The estimator above and similar constructions will be discussed more thoroughly below, also in the framework of Bayesian nonparametrics.

2. Bayesian nonparametrics

The phrase Bayesian nonparametrics is commonly used for a large and diverse collection of models and methods that extends the classical parametric (finite-dimensional) Bayesian modelling framework; for a complete introduction to parametric modelling from a Bayesian perspective see e.g. Gelman et al. (2013). With respect to the parametric approach, Bayesian nonparametrics typically refers to Bayesian models with a large or infinite number of parameters, i.e. a really large parametric model. In this sense, it can be viewed as a Bayesian take on nonparametric frequentist modelling, e.g. nonparametric regression, density or distribution estimation. The label is also used in relation to parametric models where the number of parameters increases with the size of data, as in Ghosal (2000) and Hermansen

& Hjort (2014a). In general, Bayesian nonparametrics has been successfully implemented and used in a variety of statistical models, e.g. density estimation, nonparametric regression, clustering, hazard rate and survival function estimation and in time series modelling; see Hjort et al. (2010) and references therein for additional examples and applications.

More generally, we can view Bayesian nonparametrics as families of distributions or models that are, or become, dense in a some large space of distributions relevant to the problem at hand. This makes an indirect reference to prior specification, which is as always central to the Bayesian construction. Specifying good priors in standard parametric models is difficult, and for nonparametric problems it becomes even harder and failing to do so properly may cause serious problems, see among others Diaconis & Freedman (1986a,b).

The task of constructing priors that actually represent and model our underlying prior knowledge is difficult and rarely done properly, even in simple and classical parametric models. This task becomes in general even more daunting (but also more important to get right) for nonparametric models, since we now have to specify priors on infinite-dimensional parameter spaces, such as the set of all density functions or all continuous functions on the unit interval, as might be the case in a regression setup.

The aim of the present section is to introduce Bayesian nonparametric modelling in statistics and we will focus on how to build priors on a space of infinite-dimensional objects. The discussion will be built around the classical problem of estimating unknown distribution functions, where we will follow the presentation of Ferguson (1973). It is not intended to be complete and more comprehensive introductions to Bayesian nonparametrics can be found in Ghosh & Ramamoorthi (2003) and Hjort et al. (2010).

2.1. Random probability measures and the Dirichlet process. As already commented on, Bayesian nonparametrics can be seen as a Bayesian approach to nonparametric frequentist methodology, as summarised e.g. in Wasserman (2006). In nonparametric modelling the objects of interest (or parameters) are typically functions indexed by large or infinite-dimensional sets, like regression, density or hazard rate functions. Then, following the Bayesian paradigm we now have to specify a prior on these infinite-dimensional objects/functions in an infinite-dimensional parameter space. The fundamental idea is that this can be achieved by using stochastic processes to make random functions, e.g. random density, regression or spectral density functions; common choices are types of Gaussian processes and independent increment processes (e.g. Lévy processes), see Hjort et al. (2010, Ch. 1).

In this introduction we focus on Dirichlet processes, which is the 'natural' extension of the Dirichlet distributed random variable and was introduced in Ferguson (1973). The Dirichlet process was originally introduced as a prior for the distribution functions, i.e. a method for constructing random distribution functions, see Ferguson (1973, 1974) and Antoniak (1974); this is the motivation we will follow here. The Dirichlet process has since been used extensively as a basis in several Bayesian nonparametric constructions, e.g. infinite mixtures models, hierarchical extensions, clustering and hidden Markov models, see e.g. Hjort et al. (2010, Ch. 2, 3 & 5).

The empirical distribution, a frequentist solution to nonparametric estimation of the distribution function, is well studied and its properties under standard models are well known, see among others van der Vaart (1998). In order to obtain a successful Bayesian analogue, we need a proper prior construction, methods for posterior inference, theoretical justification and properties, potential limitations and restriction.

In order to make the connection to Dirichlet processes, we remember that the multinomial distribution defines a probability measure on the sample space of finitely many integers. In order to motivate the nonparametric construction we therefore start by discussing this parametric relative, i.e. the multinomial model with a Dirichlet prior, which can be viewed as a prior on the sample space of finitely many integers.

We will first consider the simple case with a sample space with two outcomes $\{1, 2\}$, i.e. a Bernoulli experiment, e.g. flipping an unfair coin. The corresponding space of probability distributions can be represented as $\{\pi = (\pi_1, \pi_2) : \pi_1, \pi_2 \geq 0 \text{ and } \pi_1 + \pi_2 = 1\}$. Since $\pi_2 = 1 - \pi_1$ for $0 \leq \pi_1 \leq 1$, any probability measure on $[0, 1]$ defines a prior distribution on this simple set of two outcomes. In other words, a random number on the unit interval provides a suitable random measure. The 'standard' solution is to use a beta distribution where $\pi_1 \sim \text{Beta}(\alpha_1, \alpha_2)$. Let $x_1, \ldots, x_n$ be a sequence of independently distributed random variables according to p. It is then easy to show that the posterior distribution of p is a new beta distribution with parameters $\alpha_1 + \sum_{i \leq n} \delta_{x_i}(1)$ and $\alpha_2 + \sum_{i \leq n} \delta_{x_i}(2)$. This gives valuable insight into the prior specification and posterior by

$$\mathrm{E}\,\pi_1 = \frac{\alpha_1}{\alpha_1 + \alpha_2} \quad \text{and} \quad \mathrm{E}\,\{\pi_1 \mid \text{data}\} = \frac{\alpha_1 + \sum_{i \leq n} \delta_{x_i}(1)}{\alpha_1 + \alpha_2 + n}.$$

The extension to the case with m outcomes, say $\{1, \ldots, m\}$, is straightforward. This can be thought of as throwing an m sided loaded dice where we do not know the probability of a certain face and the task is to make inference about the corresponding unknown probabilities $(\pi_1, \ldots, \pi_m)$. Note that in the classical framework, if we toss the dice n times, then we will estimate $\pi_j = \Pr\{\text{the dice showing } j\}$ by n_j/n, for $j = 1, \ldots, m$. These maximum likelihood estimates are reasonable and

good if we have enough data, but for cases with few observations, it may not work properly (consider the case where we do not observe any θ_k, for some subsets of k). The Bayesian approach with the Dirichlet distribution as a conjugate prior, may still behave reasonably well in such situations, however.

The idea motivating the Dirichlet process as a probability measure on the space of measures, is the construction of a process that works as a finite-dimensional Dirichlet distribution when data are grouped, however, this should be true for any type of grouping mechanism, see Ferguson (1973).

The Dirichlet processes define a prior over probability measures and its sample paths behave almost surly as a discrete distribution function. To motivate this, let H_0 be a probability measure on the real line $\mathbb{R}$ and let α_0 be a positive real number. A Dirichlet process is the distribution of a random probability measure H on $\mathbb{R}$ such that, for any finite partition $(B_1, \ldots, B_k)$ of $\mathbb{R}$, the random vector $(H(B_1), \ldots, H(B_k))$ is distributed as a finite-dimensional Dirichlet distribution, i.e.

$$(H(B_1), \ldots, H(B_k)) \sim \mathrm{Dir}(\alpha_0 H_0(B_1), \ldots, \alpha_0 H_0(B_k))$$

Typically, we write $H \sim \mathrm{DP}(\alpha_0 H_0)$ if H is a random measure distributed according to the Dirichlet process. The probability measure H_0 is often referred to as the base measure and α_0 is called the concentration parameter.

As we obtain samples from the underlying model, say $x_1, \ldots, x_n$, we update the posterior distribution. For a fixed partition, we get a standard Dirichlet update, in the sense that for the cell containing x_1 the exponent is increased by one, all others stays the same. This is true for all cells, which suggests that the posterior is a Dirichlet process with an additional atom at x_1. This is indeed the case and it can further be shown that

$$H \,|\, x_1, \ldots, x_n \sim \mathrm{DP}\bigg(\alpha_0 H_0 + \sum_{i \leq n} \delta_{x_i}\bigg), \tag{2.1}$$

see e.g. Ghosh & Ramamoorthi (2003). This now implies that the posterior mean of H given $x_1, \ldots, x_n$ can be expressed as

$$H_n = \mathrm{E}\left\{H \,|\, x_1, \ldots, x_n\right\} = \frac{\alpha_0}{\alpha_0 + n} H_0 + \frac{n}{\alpha_0 + n} \mathbb{H}_n$$

where $\mathbb{H}_n$ is the empirical distribution; this also gives some intuition on α_0 as the concentration/precision parameter.

2.2. Posterior consistency and Bernshteĭn–von Mises theorems. For standard finite-dimensional models, most reasonable priors will usually be dominated by data as the sample size increases. In this sense the selected prior will eventually be washed out and is therefore not that important, at least if there is

sufficient amount of data. For the nonparametric models, however, proper specification of the prior becomes much more important and reasonably well behaved priors may produce ill-behaved posterior distributions as pointed out above.

Posterior consistency and Bernshteĭn–von Mises theorems are types of frequentistic validation of Bayesian procedures. These are intended to provide large-sample justification for Bayesian procedures, in the sense that as data, and the amount of information increases, the prior beliefs 'disappears' as the information accumulate.

In short, posterior consistency means that the posterior distribution concentrates around the true parameter value (parametric or nonparametric) as the sample size increases. Bernshteĭn–von Mises theorems usually refer to situations where the posterior distribution (suitable normalised) approaches a Gaussian limit distribution. It is also used as a label for models (and prior constructions) where the corresponding frequentist estimator (typically the maximum likelihood estimator) and the posterior distribution share the same type of large-sample properties, with respect to limit distributions and efficiency. In this sense, both posterior consistency and Bernshteĭn–von Mises theorems aim at providing classical frequentist large-sample justifications.

Let $\theta \in \mathbb{R}^p$, where p is finite, and $x_1, \ldots, x_n$ be i.i.d. observations from the model with density function h_{θ_0}. Then for most well-behaved models the maximum likelihood estimator $\widehat{\theta}_n$ is consistent, in the sense that $\widehat{\theta}_n \to_P \theta_0$, as the sample size increases. Moreover, it follows further that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d \mathrm{N}(0, I(\theta_0)^{-1}),$$

where $I(\theta_0)$ is the Fisher information matrix, see e.g. Ferguson (1996); similar results are known to be true in more general models, see e.g. van der Vaart (1998).

Let π be a prior density for θ representing the underlying beliefs about the parameter. Then for most well-behaved and regular priors (typically π is assumed to be positive and continuous in a neighbourhood of the true parameter θ_0) it follows that $\theta \,|\, \text{data} \to_P \theta_0$ and moreover

$$\sqrt{n}(\theta - \widehat{\theta}_n) \,|\, x_1, \ldots, x_n \to_d \mathrm{N}(0, I(\theta_0)^{-1}).$$

Depending on the tradition, such pairs of common weak convergence, as we quite informally have described above, is what we mean when we refer to Bernshteĭn–von Mises theorems; more detailed and general derivations can be found in van der Vaart (1998) or Ghosh & Ramamoorthi (2003). As a final remark we point out that (in both cases) the latter weak convergence is easily seen to imply the consistency result.

In nonparametric estimation all of this becomes more complicated, especially in the Bayesian tradition, where there are few, or essentially no, general theorems that

establish sufficient conditions for establishing types of Bernshteĭn–von Mises theorems. The case of nonparametric posterior consistency is somewhat more successful (cf. Ghosh & Ramamoorthi (2003, Ch. 4)), but both approaches typically require much more caution than with the simpler parametric models; we do not intend to go into details here, however.

For the Dirichlet process prior above, it was shown in Lo (1983) that the posterior of $\sqrt{n}(H - \widehat{H}_n)$, with H as in (2.1), converges weakly to a H_0-Brownian bridge, see e.g. van der Vaart (1998, Ch. 18) for a definition. This now matches up with the well known Donsker theorem that establishes that $\sqrt{n}(\widehat{H}_n - H_0)$ has a H_0-Brownian bridge limit. The derivation uses techniques that rely on the particular structure of the Dirichlet distribution and is therefore not easy to generalise. This is also the case for Hermansen & Hjort (2014a), which is discussed in Section 4. For a somewhat different application of the Dirichlet process, see Hjort & Petrone (2007) for nonparametric inference for quantiles, which also result in types of Bernshteĭn–von Mises theorems.

Posterior consistency and types of Bernshteĭn–von Mises results in Bayesian nonparametrics are both fields of ongoing research, the reader is referred to both Ghosh & Ramamoorthi (2003) and Hjort et al. (2010) for more comments, details and further references.

3. Model selection and focused inference

The task of selecting an appropriate model is an important and integrated part of parametric modelling in statistics. The simple intuition is that in most real life situations we, as statistical model builders, usually have more than one reasonable candidate for modelling the phenomena under study. Model selection has a long history, ranging from visual inspection, to goodness-of-fit tests and the so-called model information criteria; for a general introduction to model selection see Claeskens & Hjort (2008).

The well-known bias–variance trade-off is illustrated in Figure 3.1, where a common expectation is that a 'good' model selection strategy should balance out complexity against simplicity and precision in a reasonable way. The preferred model should be rich enough to capture the essential features, with high enough precision to be useful, and at the same time still be simple enough to be intelligible.

The above considerations focus on model fit and assessment, how good the model is at describing or approximating the observed data. There is another aspect to model selection that is more related to interpretation and is in our view often a bit under-communicated. The idea is easily illustrated with the number of skiing days dataset in Figure 1.1 above. Suppose that we have two candidate models, e.g. (i)

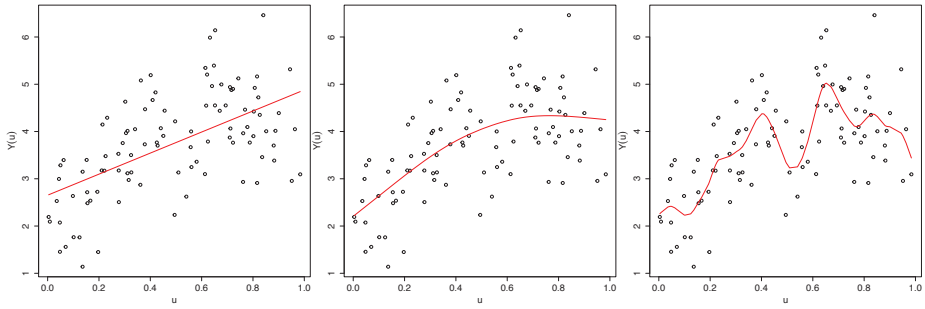

FIGURE 3.1. Bias–variance trade-off: By visual inspection, the model (solid line) in the middle panel is in some sense just right. The other two are either a little to simple (left panel) or to complex (right panel) capturing more noise than signal. A good model selection procedure should hopefully guide the user to make similar conclusions in situations that not as easily inspected as the one above.

$y_t = a + b \times \text{year}_t + \epsilon_t$ or (ii) $y_t = a + \eta_t$, where $\{\epsilon_t\}$ and $\{\eta_t\}$ are stationary time series processes of the same (or similar) structure. If model (i) is judged as better than model (ii), this will have consequences for the inference, especially future predictions, however, we may have learned something new and important about phenomenon we study, i.e. that the number of skiing days at Bjørnholt is actually decreasing.

The popular model information criteria, e.g. Akaike's information criterion (AIC; Akaike (1973)), the Bayesian information criterion (BIC; Schwarz (1978)) and the focused information criterion (FIC; Claeskens & Hjort (2003)), have a considerable appeal, since these are typically simple in both structure and use, resulting in scores which can be used to rank candidate models from best to worst in accordance with some predefined measure of discrepancy. This simplicity has led to widespread and uncritical use of such criteria, especially the AIC, which is commonly used without any concern for the underlying motivation; we will return to this in Section 3.3 below. The AIC and the BIC, which lead to one 'best' model, aiming respectively at the one minimising a certain Kullback–Leibler divergence from the underlying true data generating mechanism to the model in question, and the one maximising the posterior model probability. These are global perspectives, that prefer models that, in some more or less practical sense, capture the main characteristics of the underlying distribution function.

3.1. The Focused information criterion (FIC). The focused information criterion (FIC) was introduced in Claeskens & Hjort (2003) and Hjort & Claeskens (2003) and is based on the comparison of the estimated accuracy of individual model estimators for a chosen focus parameter/function μ. Instead of aiming at a model

that is 'reasonably good at everything', the motivation underlying the FIC is that the intended use of the model and the focus of the investigation should play a central part of the selection procedure. One and the same model is typically not the best for all applications; this is e.g. evident for regression models, where some covariates may be important for some types of questions but of lesser importance for other aspects of what is being studied.

Let $\mu = \mu(\theta, \gamma)$ be the parameter of interest, i.e. the focus parameter or function, e.g. quantiles or certain important regression parameters. For time series models (cf. Hermansen & Hjort (2014d)) the canonical example is h-step ahead predictions, i.e. the task of finding the model the is best for predicting h time steps into the future (cf. Akaike (1969) and Linhart & Göttingen (1985)). Moreover, a wide variety of other focused questions, with more or less of a time series specific nature, are easily motivated, such as estimation of threshold probabilities, determination of confidence bounds, the effect of certain covariates or more direct features like certain covariance lags or properties of the spectral density for frequencies close to zero, to name a few; see also Section 5 for additional comments.

The idea leading to the FIC is to approximate the mean squared error (mse) of $\widehat{\mu}_S$ for each candidate model in a set of nested submodels index by $\{S\}$ and prefer the model that achieves the smallest value. This will be solved in a large-sample framework, where the actual mse will be approximated by estimating the squared bias and variance of $\sqrt{n}(\widehat{\mu}_S - \mu_{\text{true}})$ in the limit experiment, where μ_{true} is the focus function evaluated in the true model. This involves certain technical constructions needed to ensure fruitful approximation formulae, see Claeskens & Hjort (2003) and Claeskens & Hjort (2008, Ch. 5 & 6) for a more complete discussion and additional comments. A more detailed introduction of the FIC, with focus on the stationary time series models, will be given in Section 5 below.

As a last comment, we note that Akaike's final prediction error (FPE; Akaike (1969)) is an example of a criterion that also aims at answering a more precise question. This criterion is derived with the motivation of finding the autoregressive model that will minimise the one-step ahead prediction error in a given dataset; see also Linhart & Göttingen (1985) and Hermansen & Hjort (2014d) for generalisations. Nevertheless, the scores made by the FPE are often viewed as a type of proxy and the criterion must often compete against the global criteria at finding the 'true' model.

3.2. The Akaike information criterion (AIC). Akaike's information criterion (AIC) is among the more popular and important model selection strategies. In

its general form it is defined by

$$\mathrm{AIC}(M) = 2\,\text{log-likelihood}_{\max}(M) - 2\dim(M), \tag{3.1}$$

for each candidate model M, where $\dim(M)$ is the length of the parameter vector. In short, the AIC machinery is to prefer the model that attains the largest value of (3.1) above.

Let $\ell_n(\theta)$ be the log-likelihood function based for some model parametric model represented by the density h_θ from a class of potential candidates. Now, the general AIC formula above becomes $\mathrm{AIC} = 2(\ell_n(\widehat{\theta}_n) - p)$, where $\widehat{\theta}_n = \arg\max \ell_n(\theta)$ is the maximum likelihood estimator. The AIC machinery becomes readily available and a convenient measure for comparing candidate models, which has led to widespread and indiscriminate use of the AIC in statistics and other related fields of research.

In our view, a particular model selection procedure, like the AIC, should not be used, or preferred above others, simply because it is convenient. In order to prefer one criterion above others, a proper rational motivation and understanding is needed, since we need some guarantee that the model preferred, say by the AIC, is the one that actually has the properties that we really care about. Moreover, the rational for using the AIC construction relies on a precise and well motivated chain of large-sample arguments that do not necessarily hold up in general. This means that there is not necessarily a rational motivation for using the AIC as a model selection procedure and ranking the models by their AIC-score, i.e. the attained values of (3.1), become random or irrelevant; see Grønneberg & Hjort (2014) for a case where the general AIC formula is not well motivated and further exploration is required.

Motivated by classical likelihood theory, the structure of the AIC formula seems quite reasonable, since among competing models the one with the largest log-likelihood provides the 'best' fit to data. This strategy, by itself, makes us very vulnerable to overfitting and will have a clear preference for models resulting in a large generalisation errors. For this reason, the second term in (3.1) is commonly interpreted as a penalty term, that penalises models with unnecessarily high complexity. This is partly incorrect (cf. Claeskens & Hjort (2008) or Hermansen & Hjort (2014d)) and for most standard parametric models for i.i.d. observations, the particular structure of the AIC formula has a precise and well motivated large-sample justification, however.

3.3. The AIC for parametric models for i.i.d. data. The aim of this section is to justify the AIC as an coherent extension of the maximum likelihood principle, for estimation across families of parametric models.

Let $x_1, \ldots, x_n$ be i.i.d. realisations from the model with density h° and let h_θ represent a parametric candidate, where $\theta \in \mathbb{R}^p$ for a finite p and h° is not necessarily included or spanned by the set of candidate models, i.e. meaning we are working in a potentially misspecified modelling framework. Then under mild regularity conditions, the maximum likelihood estimator converges $\widehat{\theta}_n \to_{\text{a.s.}} \theta_0 = \arg\min_\theta \mathrm{KL}(h^\circ, h_\theta)$, with

$$\mathrm{KL}(h^\circ, h_\theta) = \int_{\mathbb{R}} \log \frac{h^\circ(x)}{h_\theta(x)} h^\circ(x)\,\mathrm{d}x = \int_{\mathbb{R}} h^\circ(x) \log h^\circ(x)\,\mathrm{d}x - R(\theta) \tag{3.2}$$

being the Kullback–Leibler divergence and where we refer to $R(\theta) = \mathrm{E}_{h^\circ} \log h_\theta(X) = \int h^\circ \log h_\theta\,\mathrm{d}x$ as the model specific part, see e.g. Claeskens & Hjort (2008, Ch. 2) for additional comments.

The maximum likelihood estimator for a particular model h_θ aims at minimising the Kullback–Leibler divergence above. In order to evaluate its performance and compare it with the other competing candidate models, we will study the actually attained Kullback–Leibler divergence

$$\mathrm{KL}(h^\circ, h_{\widehat{\theta}_n}) = \int h^\circ(x) \log h^\circ(x)\,\mathrm{d}x - R(\widehat{\theta}_n) \tag{3.3}$$

which is a random variable. The first term is the same across all models, meaning that it is sufficient to study $R(\widehat{\theta}_n)$, which further suggests that

$$Q_n = \mathrm{E}_{h^\circ} R(\widehat{\theta}_n) = \mathrm{E}_{h^\circ} \int h^\circ(x) \log h_{\widehat{\theta}_n}(x)\,\mathrm{d}x$$

is a reasonable measure for the success. This motivates a model selection strategy by preferring the model that attains the largest value of Q_n. This model is also expected to minimise (3.3) and can therefore be viewed as best at what the maximum likelihood estimator can be interpreted of trying to achieve, i.e. to be close to the true density h° with respect to the expected Kullback–Leibler discrepancy.

In order to implement this strategy in practice we need to calculate Q_n for each candidate model, which depends on the true underlying density h°, which is unknown, meaning that attained values have to be estimated from data. Since we expect $\ell_n(\theta)/n$ to be close to $R(\theta)$ by the law of large numbers, a natural estimator for Q_n is $\widehat{Q}_n = \ell_n(\widehat{\theta}_n)/n$. This motivates in turn the strategy of preferring the model that maximises $\widehat{Q}_n$. This simple log-likelihood based estimator $\widehat{Q}_n$ has a tendency to overshoot its target Q_n and a bias correction is therefore needed, however. In short, the bias correction of $\widehat{Q}_n$ justifies $\mathrm{AIC}(\theta) = 2n\{\widehat{Q}_n - \frac{1}{n}\dim(\theta)\} = 2(\ell_n(\widehat{\theta}_n) - p)$ as an (approximative and asymptotic) first order bias corrected estimator for Q, see also Claeskens & Hjort (2008, Ch. 2) for a more complete derivation and comments.

The penalisation term p has therefore a more substantial meaning as a bias correction term, which is related to the Kullback–Leibler divergence and maximum

likelihood estimation. If this more profound justification were not present, the precise structure of the penalty term becomes essentially arbitrary, since there is no real reason why we should prefer the AIC above any other similar constructions such as $2(\ell_n(\widehat{\theta}_n) - \frac{1}{2}p)$, $2(\ell_n(\widehat{\theta}_n) - \sqrt{p})$ or $2(\ell_n(\widehat{\theta}_n) - \frac{1}{2}(\log n)p)$.

In Akaike (1973) the AIC formula is motivated, by following the so-called extended likelihood principle, which dictates that we should prefer the model that maximises

$$\mathrm{E}_{h^\circ} \log h_{\widehat{\theta}_n}(X) = \mathrm{E}_{h^\circ} \int h^\circ(x) \log h_{\widehat{\theta}_n}(x)\,\mathrm{d}x, \tag{3.4}$$

where $\widehat{\theta}_n$ is the corresponding maximum likelihood estimator and X is a new independent random variable. The expectation to the left is therefore with respect to both the random variable X and the estimator $\widehat{\theta}_n$. The derivation of the AIC formula then follows along a similar line of arguments as presented above.

Note that (3.4) is equivalent to the expected model specific part of the attained Kullback–Leibler divergence above. In the original papers by Akaike, the connection to estimation in a misspecified model framework is not made explicitly. The principle is instead commonly interpreted in relation to a type of predictive performance for a new unobserved point and justified via the connection to Kullback–Leibler divergence, which is further made rational by referring to information theory and entropy, see Akaike (1973, 1974) for more details. The reasoning is unfortunately somewhat vague making the general ideas harder to grasp.

In our view, the best motivation and justification for the AIC, at least for classical parametric models for i.i.d. observations, is from the rational coupling between the AIC formula, via Kullback–Leibler divergence, and the large-sample properties of the maximum likelihood estimator in a misspecified modelling framework. The aim of establishing such connections is the part of the underlying motivation for the work in Hermansen & Hjort (2014d) and Grønneberg et al. (2014, p. 43).

As a final remark, we note that there is a model robust alternative to the AIC formula defined above, often referred to as Takeuchi's information criterion (TIC; Takeuchi (1976)). It is obtained by relaxing a hidden assumption made in the derivation of the general AIC formula. In order to get $\dim(M)$ in (3.1) we have to assume that each candidate model spans the true model, which is an unrealistic and somewhat strange assumption. Let M_p be a p-dimensional candidate model, meaning that $\dim(M_p) = p$ and suppose we aim at a more model robust strategy; in the sense that we do not assume that the candidates contain the true model. In this case, it turns out that the right correction is rather given by $p^* = \mathrm{tr}(J^{-1}K)$, where K and J are analogue to the Fisher information matrices obtained from the variance of the first and the expectation of minus the second derivative of the log-likelihood function, respectively. If the model is correctly specified it follows that $J = K$ and

we obtain the standard formula by $p^* = \mathrm{tr}(J^{-1}K) = \mathrm{tr}(I_p) = p$. In a misspecified model setup, however, the two matrices are not guaranteed to be equal, meaning that $p^* \neq p$. Note that this robust correction term can no longer be interpreted without approximations using data.

4. Bernshteĭn–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach

The paper Hermansen & Hjort (2014a) emerged from unfinished work and ideas of the master thesis Hermansen (2008). The ambitious goal of this master project was to develop Bernshteĭn–von Mises theorems for nonparametric estimation of covariance functions for time series and spatial processes. Moreover, we intended to study processes over both continuous and discrete time domains. In hindsight, a quite demanding project and also somewhat overwhelming, however, at the time I did my best effort and the result was an interesting and very educational exercise, i.e. a good master project. Although I was not able to complete everything, I was still able to establish a good groundwork and sketch out reasonable and heuristic strategies. More importantly, it motivated the study of Bernshteĭn–von Mises theorems for nonparametric function estimation and a family of (quite suitable) prior distribution, the so-called piecewise constant priors, i.e. a prior construction with a growing number of parameters by sample size, which became one of the main building blocks in my first PhD project Hermansen & Hjort (2014a).

4.1. Introduction and summary. In the present section we intend to give a brief overview and motivation for the locally constant prior construction alluded to above. This introduction is needed for the discussion and derivations in Sections 4.2 and 4.3 below, where we intend to sketch some actual solutions to some of the unfinished goals of Hermansen (2008) and Hermansen & Hjort (2014a). The overall idea and construction is easiest motivated by discussing the prototype example of Hermansen & Hjort (2014a, Section 3).

Consider the model

$$Y_i = f^\circ(i/n) + \varepsilon_i, \quad \text{where } \varepsilon_i \sim \mathrm{N}(0, {\sigma^\circ}^2) \tag{4.1}$$

for $i = 1, \ldots, n$, and where the signal f° is an unknown smooth and bounded function on the unit interval. The purpose is to make inference for the cumulative function $F^\circ(t) = \int_0^t f^\circ(u)\,\mathrm{d}u$, which is almost the same as $F_n^\circ(t) = \frac{1}{n}\sum_{i/n\leq t} f^\circ(i/n)$ for $0 \leq t \leq 1$. A canonical estimator for F° is the cumulative average process

$$F_n^*(t) = \frac{1}{n}\sum_{i/n\leq t} Y_i \quad \text{for } 0 \leq t \leq 1. \tag{4.2}$$

In this simple model it is easily shown that $F_n^*(t)$ is uniformly strongly consistent and that there is process convergence

$$\sqrt{n}\{F_n^*(t) - F_n^\circ(t)\} \to_d W(\sigma_0^2 t) \quad \text{as } n \to \infty, \tag{4.3}$$

where $W(\cdot)$ is a standard Wiener process (i.e. Brownian motion). This holds also without the Gaussian assumption of (4.1), see Hermansen & Hjort (2014a) for additional details and comments.

To simplify the current presentation we assume that σ° is known. Let f_π represents our prior guess about f and σ_π our precision, a simple conjugate approach is to use $f(i/n) \sim \mathrm{N}(f_\pi(i/n), \sigma_\pi^2)$, for $i = 1, \ldots, n$. From familiar conjugacy properties of normal-normal Bayesian models it follows that our object of interest, the cumulative function $F_n(t) = n^{-1} \sum_{i/n \le t} f(i/n)$, has a Gaussian posterior distribution. Moreover, it is straightforward to derive exact formulae for the mean and variance. This simple approach turns out to be too naive and possesses some undesirable features, however, e.g. it is easy to verify that we are not able to ensure posterior consistency with this prior, see Hermansen & Hjort (2014a, Section 3). The initial prior construction is in a sense too informative, with a separate prior for each of n parameters, not quite leaving the information in the n data points the chance to accumulate and wash out the prior, as typically seen in lower-dimensional models.

To reduce the influence of the prior we shall instead work with a class of priors for which f is taken as piecewise constant on a set of subintervals, or windows, and where the number $m = m_n$ of windows will be allowed to increase with sample size n. The windows are for the current presentation assumed to be of equal size and hence catching essentially the same number of data points. Writing k_j for the number of i/n points inside window $W_j = ((j-1)/m, j/m]$ we have $k_j \doteq n/m$. We will explore the dynamics between the number of windows and the number of data points in each window, with the main task being to derive conditions required to arrive at the appropriate Bernshteĭn–von Mises results.

Let $\bar{Y}_j$ be the average of the observations in window W_j, then the frequentist equivalent to the piecewise constant modelling above, is to estimate the cumulative F by

$$\widehat{F}_n(t) = \frac{1}{m} \sum_{j/m \le t} \bar{Y}_j \quad \text{for } t \text{ of type } \ell/m, \tag{4.4}$$

with linear interpolation between these points, i.e. $\widehat{F}_n(t) = \widehat{F}_n((j-1)/m) + \{t - (j-1)/m\}\bar{Y}_j$ for t in window W_j. Moreover, in view of (4.3), it now follows that

$$\sqrt{n}\{\widehat{F}_n(t) - F_n^\circ(t)\} \to_d W(\sigma_0^2 t) \quad \text{as } n \to \infty, \tag{4.5}$$

provided $\sqrt{n}/m^2 \to 0$, see Hermansen & Hjort (2014a, Section 9.2).

Let f_j be the level of the function inside window W_j and if f_π, a bounded function on the unit interval, that represents our prior belief, then we take the prior distribution to be given by

$$f_j \sim \mathrm{N}(f_\pi(w_j), \sigma_\pi^2) \quad \text{for } j = 1, \ldots, m, \tag{4.6}$$

independently across windows, with w_j denoting the midpoint of window W_j. Provided $m \to \infty$ and $m/\sqrt{n} \to 0$, it now follows that

$$\sqrt{n}\{F_n(t) - \widehat{F}_n(t)\} \mid \text{data} \to_d W(\sigma_0^2 t) \quad \text{as } n \to \infty. \tag{4.7}$$

This may be seen as half of a Bernshteĭn–von Mises result, in partial parallel with result (4.3), which involves F_n^* of (4.2) rather than $\widehat{F}_n$ of (4.4), as here. Summarising the above, we have shown that the conditions

$$m/\sqrt{n} \to 0 \quad \text{and} \quad \sqrt{n}/m^2 \to 0, \tag{4.8}$$

which translate to $m^2/n \to 0$ and $m^4/n \to \infty$ as $m \to \infty$, secure the Bernshteĭn–von Mises theorem (4.5)–(4.7). Note that if $m = cn^\alpha$, for example, then we need $\alpha \in (\frac{1}{4}, \frac{1}{2})$. In general, there are additional technical conditions needed to extend the result to general models beyond the simple Gaussian with a conjugate prior.

4.2. Extending to processes over a d-dimensional window. In Hermansen & Hjort (2014a) all functions f° considered were assumed to be defined on some finite one-dimensional interval. There are examples where piecewise constant modelling over higher-dimensional domain will be natural, however, e.g. as with Poisson maps, where rate estimates are presented county by county. It is, however, not entirely straightforward to extend the general large-sample results (this includes the frequentistic and Bayesian extensions) to include such cases.

In the techniques used to prove the Bernshteĭn–von Mises theorems, the true cumulative function and the corresponding estimators do, in a sense, not 'know' the shape or size of the window partition, making it easy to extend the results to processes over d-dimensional domains, with $d \geq 2$. The main difficulty with the extension is related to approximation accuracy of the piecewise constant model in relation to the Riemann sum representation of the integral.

To indicate the problem, let f° be a function defined on the unit square, which is furthermore divided into m^2 equal squares, or windows. Then, given n equidistant observations, the number of observations in each window is $k_j \asymp n/m^2$, for all $j = 1, \ldots, m^2$. This suggests that in order to ensure posterior consistency, we have to limit the number of windows m^2 such that $\sqrt{n}/\max_{j \leq m^2}\{k_j\} \asymp m^2/\sqrt{n} \to 0$ is satisfied; this is analogue to the one-dimensional case.

Secondly, we need to make sure that the difference in mean, between the frequentistic piecewise constant model $\widehat{F}_n$ and the cumulative average process F_n^*, is negligible. By a similar argument to that of Hermansen & Hjort (2014a, Lemma 9.1) it follows that the largest contribution by one window, to the total difference, is of order m^{-3} and there are at most $2m$ such small contributions. Therefore, to ensure that the total is negligible, at the claimed level of precision, it is required that $\sqrt{n}/m^2 \to 0$, which violates the condition needed for posterior consistency.

This suggests that some additional care is needed to generalise the results to processes defined over high-dimensional domains. As an alternative, we could reconsider what we see as our natural frequentist target for the Bernshteĭn–von Mises theorems, since the problem is related to the accuracy of the piecewise constant representation of a function over such domains.

4.3. Nonparametric Bayesian estimation of the covariance function for stationary time series. In general, there seems to be little work on Bayesian nonparametric methods for either covariance or spectral density/measure estimation in the statistical literature. In Choudhuri et al. (2004a) a Bernstein polynomial prior (developed in Petrone (1999a,b) to construct a nonparametric prior for probability densities on the unit interval) is used to describe prior distributions on the space of spectral densities. It is really the only other proper attempt we know of today. It is an interesting paper and parts of the work presented below will use techniques and results presented there; in general there is essentially no real overlap with our work, however.

As a motivating, we consider two failed attempts for making random covariance functions. Let

$$C(h) = \exp\{-G(h, \alpha, \beta)\}, \tag{4.9}$$

where G is a for example a Gamma process for suitable choices of α and β, i.e. to concentrate the prior around a reasonable prior guess, say $C(h) = \rho^{|h|}$. As already pointed out, this will not work in practice. The reason is that the construction above may produce invalid covariance functions, i.e. functions that do not necessarily result in positive-semidefinite covariance matrices. The failure of the construction is easiest checked through simulations.

The solution to this problem is to work in the frequency domain, see (Wold's) Theorem 1 above, since essentially all positive bounded functions on $[0, \pi]$ will result in a valid construction. Let

$$C(h) = \int_{-\pi}^{\pi} \cos(\omega h)\, \mathrm{d}B(\omega\,; \alpha, \beta) \tag{4.10}$$

where B is a for example a Beta process defined on $[-\pi, \pi]$. This construction will indeed produce the right type of functions, i.e. covariance functions, and can

therefore be viewed as an prior on the space/subspace of covariance functions. The prior support of this construction is still an open question and we do not intend to solve this here, however. Analogously to the discussion of the prototype example discussed in Section 4.1 above, it will become impossible to ensure good large-sample properties such as Bernshteĭn–von Mises or posterior consistency with this prior. This conclusion will be independently verified in Section 4.6 below.

The above considerations motivate a construction, or strategy, where belief about the covariance structure are translated to the frequency domain and represented by an analogue prior on the spectral measures. This prior is then updated, before the posterior distribution is translated back to the time domain, where (Wold's) Theorem 1 is used to preserve the validity of the different mappings (the proper 'covarianceness' of the functions) and properties are 'carried along' by the continuous mapping theorem (cf. Billingsley (2009)). The workflow is illustrated below:

$$\begin{array}{ccc} \pi\{C_G\} & & \pi\{C_G \,|\, \text{data}\} \\ {\scriptstyle \text{(i)}}\downarrow & & \uparrow{\scriptstyle \text{(iii)}} \\ \pi\{G\} & \xrightarrow[\text{(ii)}]{} & \pi\{G \,|\, \text{data}\} \end{array}$$

From Section 1.2 we also already know that $\sqrt{n}\{\widehat{G}_n(\omega) - G(\omega)\}$, with $\widehat{G}_n$ as defined in (1.10), converges weakly to a Wiener process, as shown in (1.12). Moreover, even without the different mappings, deriving types of Bernshteĭn–von Mises theorems within this frequency domain is indeed interesting in itself, see Brillinger (1975), Priestley (1981) and Gray (2006) for several applications.

Now, remember that the sequence of periodogram ordinates $\{I_n(u_j)\}_{j\leq m}$, where $u_j = 2\pi j/n$ for $j \leq m = \lfloor n/2 \rfloor$, behaves (according to (1.11) above) almost as a sequence of independent exponentials. Therefore, in view of the discussion in Section 4.1, it seems reasonable to follow the general approach of Hermansen & Hjort (2014a), using the locally constant prior construction. This introduces the view of g as constant over an increasing number of small windows given by a sample size dependent partition of $0 = v_0 < v_1 < \cdots < v_m = \pi$, where $m = cn^\alpha$, for $\frac{1}{4} < \alpha < \frac{1}{2}$ and suitable finite constant c. The function, or parameter, of interest is now

$$G_n(\omega) = \sum_{v_j \leq \omega} g_j \Delta_j, \text{ for } \omega \in \{v_j\}_{j\leq m} \tag{4.11}$$

with linear interpolation between these points and where $\Delta_j = v_{j+1} - v_j$.

Let $B_n(\omega) = \sqrt{n}\{G_n(\omega) - \widehat{G}_n(\omega)\}$, then the second half of the Bernshteĭn–von Mises theorems follows provided

$$B_n(\omega) \,|\, y_1, \ldots, y_n \to_d W\left(2\pi \int_0^\omega g(u)^2 \,\mathrm{d}u\right), \quad \text{for } \omega \in [0, \pi], \tag{4.12}$$

in P_g-probability, where $W(\cdot)$ is a Wiener process.

Remark 2. *The 'in probability' statement used in (4.12) should be considered to mean that for all $\epsilon > 0$ and large n, the set of observations $y_1, \ldots, y_n$ that violate the convergence in (4.12) has a probability less than ϵ.*

This 'direct' attack above makes the analytical derivations and the update of the prior complicated, since our target, the spectral measure, is hidden inside the inverse of the covariance matrix. The results in (1.11) suggests, however, that we will have a more direct method for inference if we change the perspective and take a more pseudo-Bayes approach where we aim at establishing

$$B_n(\omega) \,|\, I_n(u_1), \ldots, I_n(u_m) \to_d W\left(2\pi \int_0^\omega g(u)^2 \,\mathrm{d}u\right), \quad \text{for } \omega \in [0, \pi], \tag{4.13}$$

in P_g-probability.

4.4. Contiguity of the Whittle measure for a Gaussian time series. The goal is now to apply the general result of Hermansen & Hjort (2014a) to establish (4.13), these are not directly applicable by the assumption of independent observations, however. It turns out that we are able to bypass this difficulty by application of a contiguity result from Choudhuri et al. (2004b); for a general introduction to the concept of contiguity see Roussas (1972) or van der Vaart (1998, Ch. 6).

Definition 1. *Let P_n and Q_n be two sequences of probability measures defined on the same measurable space $(\Omega_n, \mathcal{A}_n)$ for $n \geq 1$. Then P_n and Q_n are said to be mutually contiguous if, for every sequence of sets $A_n \in \mathcal{A}_n$, $P_n(A_n) \to 0$ if and only if $Q_n(A_n) \to 0$.*

Corollary 1. *(Choudhuri et al., 2004b, Corollary 1) Let $\{Y_t\}$ be a stationary Gaussian time series with spectral density g that is bounded away from zero and has absolute summable covariance function in the sense that $\sum_h |h|^\alpha C_g(h) < \infty$, for $\alpha > 1$. Then the actual joint distribution of the periodogram ordinates $\{I_n(u_j)\}_{j \leq m}$ and the joint distribution of independent exponential random variables with means $g(u_j)$ are mutually contiguous.*

The result of Corollary 1 establishes a link or connection between the periodogram ordinates and sequences of independent exponentials. It implies that in practice, for all problems regarding convergence in probability, we can work with the sequence of periodogram ordinates $I_n(\omega_1), \ldots, I_n(\omega_m)$ as if they actually were the sequence $g(u_1)\mathrm{E}_1, \ldots, g(u_m)\mathrm{E}_m$, where $\mathrm{E}_1, \ldots, \mathrm{E}_m$ is a sequence of i.i.d. exponential where $\mathrm{E}_j \sim \mathrm{Exp}(1)$. Then, under the conditions of Corollary 1 it follows from Hermansen & Hjort (2014a, Illustration 5.3) that (4.13) is indeed true.

4.5. Bridging the gaps. A link or bridge from the sequences of periodogram ordinates to independent exponentially distributed variables is created by the contiguity results of Choudhuri et al. (2004b). The Bernshteĭn–von Mises results derived by updating the prior with the sequence $I_n(\omega_1), \ldots, I_n(\omega_m)$ are indeed valid, interesting and creates a nice symmetry with the empirical spectral measure estimated using the same sequence.

The hope is to take this a step further and also establish (4.12). To do this we need a second 'bridge' indicated by the double arrow (i) below

$$\{Y_t\}_{t\le n} \quad \overset{\text{(i)}}{\longleftrightarrow} \quad \{I_n(u_j)\}_{j\le m} \quad \overset{\text{(ii)}}{\longleftrightarrow} \quad \{g(u_j)\mathrm{E}_j\}_{j\le m}.$$

The strategy is to appropriately replace the full Gaussian log-likelihood ℓ_n, as defined in (1.3), with the numerically and analytically simpler $\widetilde{\ell}_n$ of (1.7). This will automatically introduce the periodogram ordinates into the derivations and we may then apply Corollary 1 and the general results of Hermansen & Hjort (2014a) to obtain (4.12).

First of all, note that it is not possible to simply replace ℓ_n with $\widetilde{\ell}_n$ in the derivations. This can be seen since $\ell_n(f) = \widetilde{\ell}_n(f) + \Delta_n$, uniformly over f, where Δ_n is merely $O_{P_g}(1)$ and not $o_{P_g}(1)$, see (1.9). Moreover, it follows further that if f_θ, with $\theta \in \Theta \subset \mathbb{R}^k$, is k times uniformly differentiable in θ and Θ is compact, then $\nabla^k \ell_n(f_\theta) = \nabla^k \widetilde{\ell}_n(f_\theta) + \Delta_n'$, uniformly over f_θ, where Δ_n' is again $O_{P_g}(1)$, making a standard Taylor expansion not necessarily directly applicable either.

Remark 3. *If the spectral measure G_m is a step function with $m = \lfloor n/2 \rfloor$ is the number of steps. It is then possible, with carefully designed set of the jump/step locations, to obtain $\ell_n(G_m) = \widetilde{\ell}_n(G_m)$. This result indicates that we can hope for a a better approximation rate than $O_{P_g}(1)$ as in (1.7) for the locally constant $g_1, \ldots, g_m$ of (4.11); this will require additional exploration.*

The uniform bounds above are in a sense too general for our purpose. To illustrate this and indicate what is actually needed, let $y_1, \ldots, y_n$ be realisations form a stationary Gaussian time series process with true spectral density f_{θ_0}, where $\theta_0 \subset \mathbb{R}$; the p-dimensional parametric spectral densities can be solved similarly. Define $H_n(t) = \ell_n(\widehat{\theta}_n + t/\sqrt{n}) - \ell_n(\widehat{\theta}_n)$ and let π be the prior density representing our beliefs regarding θ. Then under conditions (i)–(v) of Ghosh & Ramamoorthi (2003, Theorem 1.4.2) asymptotic normality of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is obtained provided

$$\int_{\mathbb{R}} \left| \pi(\widehat{\theta}_n + t/\sqrt{n}) \exp\{H_n(t)\} - \pi(\widehat{\theta}_n) \exp\{-t^2 J_n/2\} \right| \mathrm{d}t \to_{P_{\theta_0}} 0 \tag{4.14}$$

where $J_n = -\nabla^2 \ell_n(\widehat{\theta}_n)/n$, which converges in probability to $J(\theta_0)$ as defined in (1.5).

The proof of Ghosh & Ramamoorthi (2003, Theorem 1.4.2), which is similar to the strategy used in Hermansen & Hjort (2014a), is structured such that (4.14) is shown to be small on three separate regions of the real line, namely$A_1 = \{t : |t| < c \log \sqrt{n}\}$, $A_2 = \{t : c \log \sqrt{n} < t < \delta\sqrt{n}\}$ and $A_3 = \{t : |t| > \delta\sqrt{n}\}$. In the present discussion we will not worry about A_3, since what we intended to change does not affect this part of the argument.

To see how we can introduce the Whittle approximation, observe that

$$\begin{aligned} H_n(t) &= \ell_n(f_{\widehat{\theta}_n + t/\sqrt{n}}) - \ell_n(f_{\widehat{\theta}_n}) \\ &= \frac{t^2}{2n}\nabla^2\ell_n(f_{\widehat{\theta}_n}) + \frac{1}{6}\left(\frac{t}{\sqrt{n}}\right)^3 \nabla^3\ell_n(f_{\bar{\theta}_n}) = \frac{t^2}{2n}\nabla^2\ell_n(f_{\widehat{\theta}_n}) + R_n(t), \end{aligned}$$

where $|\bar{\theta}_n - \widehat{\theta}_n| < |\theta_0 - \widehat{\theta}_n|$ and by the analogue of condition (ii) in Ghosh & Ramamoorthi (2003) it follows that $R_n(t) = O_{P_0}(t^3/\sqrt{n})$ for large n. A similar expansion is easily obtained from the Whittle approximation, where $\widetilde{H}_n(t) = t^2/(2n)\nabla^2\widetilde{\ell}_n(\widetilde{\theta}_n) + \widetilde{R}_n(t)$, with $\widetilde{R}_n(t) = O_{P_0}(t^3/\sqrt{n})$ under essentially the same conditions to those introduced above. Then, since

$$H_n(t) = \widetilde{H}_n(t) + r_n(t) + [R_n(t) + \widetilde{R}_n(t)],$$

we are 'allowed' to replace the full Gaussian log-likelihood the Whittle approximation, provided $r_n(t) = O_{P_0}(t^3/\sqrt{n})$ on $A_1 \cup A_2$ or otherwise become negligible.

From Coursol & Dacunha-Castelle (1982) it now follows that

$$|r_n(t)| = \frac{t^2}{2n}|\nabla^2\ell_n(\widehat{\theta}_n) - \nabla^2\widetilde{\ell}_n(\widetilde{\theta}_n)| = \begin{cases} o_{P_{\theta_0}}(1), & \text{for } t \in A_1 \\ O_{P_{\theta_0}}(\delta), & \text{for } t \in A_2 \end{cases},$$

which is easily seen to be sufficiently small to allow the change of log-likelihood without breaking the validity of the general argument.

This means that with some additional work we are able to successfully obtain Bernshteĭn–von Mises theorems for parametric spectral densities, via sequences of independent and non-identical exponentially random variables, with the result that

$$\sqrt{n}(\theta - \widehat{\theta}_n) \,|\, y_1, \ldots, y_n \to_d U \sim \mathrm{N}(0, J(f_{\theta_0})^{-1}), \quad \text{in } P_{\theta_0}\text{-probability} \tag{4.15}$$

where J is as defined in (1.5).

Finally, extending the above arguments to the nonparametric framework with a locally constant prior should be fairly straightforward since $r_n(t)$ can in general be shown to be uniformly smaller than the remainder $R_n(t)$ above; additional care and work is need to make it into a rigorous proof and we do not intend to solve this here, however.

4.6. Convergence of experiments and some related ideas. As illustrated, working directly with the full Gaussian log-likelihood is often impractical with respect to the underlying spectral density, since it is hidden inside the inverse of the covariance matrix. The introduction of the Whittle approximation in Section 1.2 and the corresponding connection, by contiguity, with a sequence of independent exponentials in Section 4.4, made it possible to derive types of large-sample results more easily. In this section we attempt to motivate an alternative approach, which also creates a bridge to a simpler, but different, modelling framework, using the notion of convergence, or equivalence, of statistical experiments/models (cf. Blackwell et al. (1951)); see among others Le Cam & Yang (2000) and van der Vaart (1998, Ch. 6, 7 & 9) and also Florens et al. (1990) for a more Bayesian related discussion of this topic.

First of all, a collection of probability measures $\mathcal{E} = \{P_h : h \in H\}$ defined on a σ-algebra $\mathcal{A}$ of subsets of $\mathcal{X}$ is a collection of statistical models and as a mathematical construction this is what we refer to by the phrase statistical experiment.

As an initial illustration, suppose that instead of experiment $\mathcal{E}$ above we have the opportunity to carry out a different experiment $\mathcal{F}_n = \{Q_h : h \in H\}$ on $(\mathcal{Y}, \mathcal{B})$, but for some reason we are not able to perform both. The question is then which one should we prefer. This motivates the need to compare statistical experiments in a relevant fashion and to answer when an experiment $\mathcal{E}$ is more informative than experiment $\mathcal{F}$; see Le Cam (1996) and Le Cam & Yang (2000) for a complete discussion.

In what follows we will give a quick summary to introduce the notation and the general idea of convergence of sequences of statistical experiments, see Golubev et al. (2010) and references therein for a more complete discussion. Consider two sequences of experiments, or families of measures, $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ and $\mathcal{F}_n = (\mathcal{Y}_n, \mathcal{B}_n, Q_{n,h} : h \in H)$, then we will say that $\mathcal{E}_n$ and $\mathcal{F}_n$ are asymptotically equivalent, and write $\mathcal{E}_n \approx \mathcal{F}_n$, if $\Delta(\mathcal{F}_n, \mathcal{E}_n) \to 0$, where $\Delta(\mathcal{F}, \mathcal{E}) = \max(\delta(\mathcal{F}, \mathcal{E}), \delta(\mathcal{E}, \mathcal{F}))$ and $\delta(\mathcal{E}, \mathcal{F}) = \inf_K \sup_h \|P_h - KQ_h\|$ is the total variation distance between the two measures and K is a Markov kernel such that KQ is a measure on the same space as P.

The following results shows that the statistical experiment with a stationary Gaussian time series processes is asymptotically equivalent to the simpler experiment consisting of a sequences of independent normal variables.

Theorem 3. *(Golubev et al., 2010, Theorem 1.1) For symmetric spectral densities $f \in L_2(-\pi, \pi)$ assume there exist an $M > 0$ and $\alpha > 1/2$ such that $C_f(0)^2 + \sum_h |h|^{2\alpha} C_f(h)^2 \leq M$ and f is bounded below by $1/M$, for all f. Then*

the experiment given by $\{y_t\}_{t\le n}$*, a stationary Gaussian sequence with mean zero and spectral density* f*, and the one given by* $\{z_i\}_{i\le n}$*, where* z_i *are independent* $\mathrm{N}(0, J_{i,n}(f)$*,*

$$J_{i,n}(f) = n\int_{\omega_{i-1}}^{\omega_i} f(u)\,\mathrm{d}u$$

and $\omega_i = 2\pi i/n - \pi$*, are asymptotically equivalent.*

The result of Theorem 3 suggests that we should be able to obtain the result of (4.15), at least, from the simpler asymptotically equivalent experiment of independent and non-identical sequences of Gaussian random variables. This should follow by similar arguments as used above, which exploits the implied nearness of likelihoods by the converging experiments; a rigorous proof is needed, however. This also motivates a potential extension to nonparametric models, using the piecewise constant prior construction; we do not intend to do this here, however.

As an additional remark, Theorem 3 provides an independent verification and motivation for the piecewise constant prior construction for the time series processes. The reason is that we can not expect to have more information, or better performances, than in the asymptotically equivalent experiment. This means that in order for the data to accumulate and to ensure interesting large-sample results in this modelling framework (such as Bernshteĭn–von Mises theorems), it is necessary to apply the piecewise constant strategy of Hermansen & Hjort (2014a), or another similar construction, which is easily seen from the asymptotically equivalent sequences of independent normals.

5. Focused information criteria for time series

The idea for deriving a variation of the focused information criterion (FIC) for stationary time series processes was initially intended as an independent section in the Hermansen & Hjort (2014d), which at that point was a broad draft discussing several model selection related issues for stationary time series processes. As the projects evolved, however, we realised that both would prosper more as separate papers. Compared to the remarkably difficult-to-write Hermansen & Hjort (2014d), the parametric framework of the FIC made it easier to develop the results needed. The extension of the FIC to time series processes did indeed require new methodology, but by building on the existing framework of Hjort & Claeskens (2003) and Claeskens & Hjort (2003), and also by modifying methodology and results in Davis (1973) and Dzhaparidze (1986), we had a rapid progression for quite some time.

We therefore decided that the project should be extended to include stationary models with trend. From the familiarity with the work of R. Dahlhaus we decided to develop this within the general family of locally stationary processes, see

e.g. Dahlhaus (1997). This general class of models, which also includes time series models with smooth trend functions, resulted in FIC methodology for a large family of models. Moreover, developing the methodology for these general models provided a unified framework for interpretation, theory, and focused model selection.

After the introduction and extension to locally stationary processes, we realised more fully the added complexity of developing such focused model selection methodology for time series processes. The time aspect and the introduction of dependency meant that several interesting focused questions, like predictions or estimation of future threshold probabilities, are most naturally formulated as conditional on past observations. This type of dependency required a new and extended modelling framework, which in turn led to a proper generalisations and also motivated a new conditional focused information criterion (cFIC).

5.1. Introduction and summary. The aim of this section is to provide the basic introduction and motivation needed to derive the FIC machinery for stationary Gaussian time series processes. The motivation, which eventually will result in the FIC, is to obtain suitable large-sample approximations that can be used to estimate the mean squared error of the model estimates for the focus parameters μ, see Section 3.1 above for an informal introduction.

In order to do this properly, we will be needing a series of technical conditions and assumptions. This is an important part of the general construction and is needed for the discussion below. Let $y_1, \ldots, y_n$ be realisations from the model with true spectral density

$$f_{\text{true}} = f_{\theta_0, \gamma_0 + \delta/\sqrt{n}}, \tag{5.1}$$

where (θ_0, γ_0) is an inner point in a compact parameter space $\Theta \times \Gamma \subset \mathbb{R}^{p+q}$. The idea behind studying the behaviour of estimators and model selectors in this local large-sample framework is to ensure that the variances and squared biases are all of size $O(1/n)$, leading to fruitful approximation formulae, see Claeskens & Hjort (2003) and Claeskens & Hjort (2008, Ch. 5 & 6) for a more complete discussion and additional comments.

The candidate models are nested between a smallest model $f_\theta = f_{\theta,\gamma_0}$, i.e. the baseline model included in all submodels, and the largest $f_{\theta,\gamma}$ that includes all modelling parameters. Then, via all possible inclusion/exclusion arrangements of the elements in γ, the result is a set of 2^q potential submodels, one for each subset S of $\{1 \ldots, q\}$, ranging from the narrow model, with $S = \emptyset$, to the full wide model where $S = \{1 \ldots, q\}$. In practice, we include only the submodels we regard as sufficiently interesting or plausible.

The FIC methodology will allow for a wide variety of focused questions, but for the present purpose we shall assume that $\mu = \mu(\theta, \gamma)$ depends the underlying model only through the parameters. The different submodel estimates for the focus function are then given by

$$\widehat{\mu}_S = \mu(\widehat{\theta}_S, \widehat{\gamma}_S, \gamma_{0,S^c}), \tag{5.2}$$

where $S \subseteq \{1 \ldots, q\}$ and S^c is the complement of S, i.e. the indexes of the parameters we do not estimate for that model.

The FIC now follows from the derivation of a large-sample approximation for the mean squared error of $\sqrt{n}(\widehat{\mu}_S - \mu_{\text{true}})$ for each submodel S, where $\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ is the focus parameter evaluated in the true model. A key element in this derivation is the large-sample properties of the scaled score functions

$$Z_n = \sqrt{n}(U_n, V_n) = \frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0, \gamma_0), \tag{5.3}$$

evaluated in the narrow null-model. Before we derive the asymptotic distribution of Z_n, observe that

$$J_{\text{wide}} = \lim_{n\to\infty} \text{Var}_{P_0} Z_n = \lim_{n\to\infty} \frac{1}{2n} \text{tr}\{(\Sigma_0^{-1}[\nabla\Sigma_0])^2\} = \frac{1}{4\pi} \int_{-\pi}^{\pi} \nabla\Psi_0(\omega) \nabla\Psi_0(\omega)^{\text{t}} \, \text{d}\omega, \tag{5.4}$$

where $\Sigma_0 = \Sigma_n(f_0)$, $\Psi_0(\omega) = \log f_0(\omega)$, $f_0 = f_{\theta_0,\gamma_0}$ and where $P_0 = P_{0,n}$ represents the distribution associated with a Gaussian vector of length n from the model with spectral density f_0. We will also be needing the following block-representation

$$J_{\text{wide}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}, \quad \text{with inverse } J_{\text{wide}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}, \tag{5.5}$$

where J_{00} is the upper $p \times p$-matrix of J_{wide} and the other block matrices are defined accordingly.

Proposition 1. *(Hermansen & Hjort, 2014b, Proposition 3.2) Let $y_1, \ldots, y_n$ be realisations from the model (5.1) and let $P_\delta = P_{\delta,n}$ be the associated probability measure. Furthermore, suppose the spectral density $f_{\theta,\gamma}$ is continuous and bounded away from zero and infinity and that $f_{\theta,\gamma}$ is also two times differentiable in (θ, γ), with derivatives that are differentiable in ω with uniformly continuous derivatives, in a neighbourhood of (θ_0, γ_0). Then*

$$Z_n = \sqrt{n} \begin{pmatrix} U_n \\ V_n \end{pmatrix} \to_d \begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix} + Z, \quad \text{where } Z \sim \text{N}_{p+q}(0, J_{\text{wide}}), \tag{5.6}$$

and J_{01} and J_{11} are block elements of J_{wide} as defined in (5.5). Moreover, $\widehat{J}_n(\bar{\theta}_n, \bar{\gamma}_n) = -\nabla^2 \ell_n(\bar{\theta}_n, \bar{\gamma}_n)/n \to_{P_\delta} J_{\text{wide}}$, provided $(\bar{\theta}_n, \bar{\gamma}_n) \to_{P_\delta} (\theta_0, \gamma_0)$.

With the above results in place, we have almost all the elements needed to obtain the limiting mean squared error for the normalised submodels estimators from (5.2). A bit more notation is needed first, however. Define $Q = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$ and for each submodel $S \subset \{1, \ldots, q\}$ define the projection matrix π_S that maps the vector $v = (v_1, \ldots, v_q)^{\mathrm{t}}$ to the subvector $\pi_S v = v_S$ with components v_j for $j \in S$. Let $Q_S = J^{11,S} = (\pi_S Q^{-1} \pi_S^{\mathrm{t}})^{-1}$, $G_S = \pi_S^{\mathrm{t}} Q_S \pi_S Q^{-1}$ and finally $\nu = J_{10}J_{00}^{-1}\nabla_\theta \mu(\theta_0, \gamma_0) - \nabla_\gamma \mu(\theta_0, \gamma_0)$ and $\tau_0^2 = \nabla_\theta \mu(\theta_0, \gamma_0)^{\mathrm{t}} J_{00}^{-1} \nabla_\theta \mu(\theta_0, \gamma_0)$.

From Hermansen & Hjort (2014b, Section 3.2) it now follows that

$$\sqrt{n}(\widehat{\mu}_S - \mu_{\mathrm{true}}) \xrightarrow{d} \Lambda_S = \Lambda_0 + \nu^{\mathrm{t}}(\delta - G_S D), \tag{5.7}$$

where $D \sim \mathrm{N}_q(\delta, Q)$ and $\Lambda_0 \sim \mathrm{N}(0, \tau_0^2)$ are independent random variables. The mean squared error of (5.7) is then

$$\mathrm{mse}(S, \delta) = \tau_0^2 + \nu^{\mathrm{t}} G_S Q G_S^{\mathrm{t}} \nu + \nu^{\mathrm{t}}(I_q - G_S)\delta\delta^{\mathrm{t}}(I_q - G_S)^{\mathrm{t}}\nu. \tag{5.8}$$

In the limit experiment, the quantities appearing in (5.8) are known, apart from δ, for which the statistical information is $D \sim \mathrm{N}_q(\delta, Q)$, with known variance matrix Q. Since DD^{t} has mean $\delta\delta^{\mathrm{t}} + Q$, the unbiased estimator of $\delta\delta^{\mathrm{t}}$ is $DD^{\mathrm{t}} - Q$. Thus an unbiased risk estimate here is $r(S) = \tau_0^2 + \nu^{\mathrm{t}} G_S Q G_S^{\mathrm{t}} \nu + \nu^{\mathrm{t}}(I_q - G_S)(DD^{\mathrm{t}} - Q)(I_q - G_S)^{\mathrm{t}}\nu$. Estimating the required quantities from the data at hand we arrive at

$$\mathrm{FIC}(S) = \widehat{\tau}_0^2 + \widehat{\nu}^{\mathrm{t}} G_S \widehat{Q} G_S^{\mathrm{t}} \widehat{\nu} + \widehat{\nu}^{\mathrm{t}}(I_q - \widehat{G}_S)(D_n D_n^{\mathrm{t}} - \widehat{Q})(I_q - \widehat{G}_S)\widehat{\nu}, \tag{5.9}$$

seen as an estimate of the limiting risk $\mathrm{mse}(S, \delta)$ of (5.8). The model with smallest $\mathrm{FIC}(S)$ is then preferred; for more details and discussion see Hermansen & Hjort (2014b, Section 3) and Claeskens & Hjort (2008, Ch. 5 & 6).

In Hermansen & Hjort (2014b, Sections 5 & 6) the FIC methodology for foci of the type $\mu = \mu(\theta, \gamma)$ is extended along with the modelling framework to handle wider range of focus functions that include predictions and functions that may also depend on the sample size n and/or parts of the already observed time series. These sections are particularly important, since they extended the original work to allow for foci that are more relevant in a time series framework, e.g. h-step ahead predictions but also $\mu(\theta, \gamma, y_1, \ldots, y_m) = \Pr\{Y_{n+1} > \alpha, Y_{n+2} > \alpha \,|\, y_1, \ldots, y_m\}$ for a suitable choice of α. In addition, it motivates a new and more general data-dependent versions of the FIC as well as the so-called conditional focused information criterion (cFIC); the argument is somewhat involved and is therefore omitted from this presentation.

5.2. Local asymptotic normality and an alternative proof for Proposition 1. The purpose of this section is to elaborate on the original argument used to prove Proposition 1 above, which was discarded from Hermansen & Hjort (2014b) after the introduction and inclusion of the locally stationary processes of Dahlhaus

(1997), which essentially made it obsolete. The original proof is shown to be true under weaker conditions, however, and we also believe that it shows interesting use of techniques.

In order to show how, we need to introduce some ideas from the theory of so-called local asymptotic normal sequence of statistical models/experiments (LAN), see Le Cam & Yang (1990), Le Cam & Yang (2000) or van der Vaart (1998, Ch. 7). The goal is to show that under standard LAN conditions, the main results needed to establish the FIC machinery, i.e. the analogue of Proposition 1 above, follows quite easily and without too much additional work.

A parametric family of distributions $\{P_{\theta,n} : \theta \in \Theta \subset \mathbb{R}^p\}$, where p is finite, is said to be LAN at a fixed point θ_0 if $\{P_{\theta_0+\delta/h_n,n}\}$ and $\{P_{\theta_0,n}\}$ are contiguous (see Section 4.4 above for a definition) and there exist a sequence of random vectors $\Delta_{\theta_0,n} \to_d \Delta_{\theta_0} \sim \mathrm{N}(0, I_{\theta_0})$ and a positive definite $p \times p$-matrix I_{θ_0} such that $\Lambda_n(\theta_0 + \delta/h_n, \theta_0) = \log\{\mathrm{d}P_{\theta_0+\delta/h_n}/\mathrm{d}P_{\theta_0}\}$ satisfies

$$\Lambda_n(\theta_0 + \delta/h_n, \theta_0) - \left[\delta^{\mathrm{t}}\Delta_{\theta_0,n} + \tfrac{1}{2}\delta^{\mathrm{t}} I_{\theta_0}\delta\right] = o_{P_{\theta_0,n}}(1),$$

for all p-dimensional vectors δ; see also Davis (1973) and Dzhaparidze (1986) for a time series oriented discussion.

In particular, we are interested in what is commonly known as Le Cam's third lemma, which states (with a slight abuse of notation) that if

$$(Z_n,\, \log\{\mathrm{d}P_\delta/\mathrm{d}P_0\}) \to_d N_{p+q+1}\left(\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \rho \\ \rho^{\mathrm{t}} & \sigma^2 \end{pmatrix}\right), \tag{5.10}$$

is true under P_0, it follows that $Z_n \to_d N_{p+q}(\mu + \rho, \Sigma)$ under P_δ, see e.g. van der Vaart (1998, Example 6.7).

If Z_n is the score functions of (5.3), Le Cam's third lemma provides a general method for obtaining the large-sample behaviour of Z_n, in the local large-sample framework, from properties of the simpler null-model. This implies that if (5.10) is known in advance for a particular class of models, the results needed to justify the FIC will follow by application of standard techniques. Moreover, in several LAN experiments it is straightforward to establish (5.10), as we intend to illustrate for the time series processes with the new/old proof of Proposition 1 below.

Proposition 2. *Let $y_1, \ldots, y_n$ be realisations from the model (5.1) and let $P_\delta = P_{\delta,n}$ be the associated probability measure. Suppose that in a neighbourhood of (θ_0, γ_0) the spectral density $f_{\theta,\gamma}$ is continuous and uniformly bounded away from both zero and infinity. Moreover, suppose that $f_{\theta,\gamma}$ is also two times differentiable, with respect to (θ, γ), with derivatives that are continuous and uniformly bounded*

in both ω and θ in a neighbourhood of (θ_0, γ_0). Then

$$Z_n = \sqrt{n}\begin{pmatrix} U_n \\ V_n \end{pmatrix} \to_d \begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix} + Z, \quad \text{where } Z \sim \mathrm{N}_{p+q}(0, J_{\mathrm{wide}}),$$

with J_{wide} as defined in (5.4). Moreover, $\widehat{J}_n(\bar\theta_n, \bar\gamma_n) = -\nabla^2 \ell_n(\bar\theta_n, \bar\gamma_n)/n \to_{P_\delta} J_{\mathrm{wide}}$, provided $(\bar\theta_n, \bar\gamma_n) \to_{P_\delta} (\theta_0, \gamma_0)$.

Proof. To see how we can obtain (5.10) in the current framework, we remind ourselves that the large-sample behaviour of

$$Z_n = (4n)^{-1/2}\{\mathrm{tr}(\Sigma_0^{-1}[\nabla\Sigma_0]) + \underline{y}_n^{\mathrm{t}}\Sigma_0^{-1}[\nabla\Sigma_0]\Sigma_0^{-1}\underline{y}_n$$

is already well studied and that $Z_n \to_d \mathrm{N}(0, J_{\mathrm{wide}})$, under the narrow model, i.e. where $\underline{y}_n$ is generated by P_0, see among others Dzhaparidze (1986, Ch. I and II); see also Lemma 1 and Remark 5 below for some additional comments. This implies that $\mu = 0$ and $\Sigma = J_{\mathrm{wide}}$ in (5.10). Then, by Lemma 11.1 in Hermansen & Hjort (2014b) it is clear that

$$\begin{aligned}
\rho &= \lim_{n\to\infty} \mathrm{Cov}_{P_0}(Z_n, \log\{\mathrm{d}P_\delta/\mathrm{d}P_0\}) \\
&= \lim_{n\to\infty} -\frac{1}{2\sqrt{n}}\,\mathrm{tr}\{[\nabla\Sigma_0](\Sigma_\delta^{-1} - \Sigma_0^{-1})\} = \lim_{n\to\infty} -\frac{1}{2\sqrt{n}}\,\mathrm{tr}([\nabla\Sigma_0]\Sigma_\delta^{-1}[\Sigma_0 - \Sigma_\delta]\Sigma_0^{-1}) \\
&= \lim_{n\to\infty} \frac{1}{2n}\,\mathrm{tr}([\nabla\,\Sigma_0]\Sigma_\delta^{-1}\Sigma_n(\delta^t\nabla_\gamma\, f_{\theta_0,\bar\gamma_n})\Sigma_0^{-1}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \nabla_\theta\,\Psi_0(\omega)[\delta^t\nabla_\gamma\,\Psi_0(\omega)]\,\mathrm{d}\omega,
\end{aligned}$$

where $|\bar\gamma_n - \gamma_0| \le \delta/\sqrt{n}$. This means that we have the claimed result provided the joint limit of (5.10) can be established. From Dzhaparidze (1986, Section 2.2) we have

$$\log\{\mathrm{d}P_\delta/\mathrm{d}P_0\} = -2^{-1}(\log|\Sigma_0|/|\Sigma_\delta| + \underline{y}_n^{\mathrm{t}}(\Sigma_\delta^{-1} - \Sigma_0^{-1})\underline{y}_n) \xrightarrow{d} N(-\frac{1}{2}\sigma^2, \sigma^2),$$

where $\sigma^2 = \delta^t J_{\mathrm{wide}}\delta$, which means that all that is needed is a fairly standard Cramér–Wold type of argument, which is omitted here. The second part of the proposition is included for completeness and we do not intend to prove this again here; the results are also extensively discussed in Hermansen & Hjort (2014b). □

Remark 4. *It is possible to derive the results of Proposition 2 using more standard techniques, such as those used in Dzhaparidze (1986) and Dahlhaus & Wefelmeyer (1996). The local large-sample framework, however, makes the derivations quite cumbersome and the use of Le Cam's third lemma simplifies the argument considerably.*

Lemma 1. *Let $X_n \sim N_n(\mu, \Sigma_n)$ and $Y_n = X_n^{\mathrm{t}} W_n X_n$, then*

$$U_n = \frac{Y_n - \mathrm{E}\,Y_n}{\sqrt{\mathrm{Var}(Y_n)}} \to_d N(0,1),$$

if and only if $\max_{i\leq n}\lambda_i/(\sum_{i\leq n}\lambda_i^2)^{1/2}\to 0$, *where* $\lambda_1,\ldots,\lambda_n$ *is the eigenvalues of* $\Sigma_n W_n$.

PROOF. See the technical report Hermansen & Hjort (2014c). □

Remark 5. *From Lemma 1 it is fairly straightforward to apply a Cramér–Wold type of argument to show that (5.10) is indeed satisfied under the narrow-null model. The main observations needed is the existence of suitable bounds on the eigenvalues, which can be shown to follow provided that there exist positive and finite numbers* m *and* M *such that* $0<m\leq f_0,f_\delta\leq M<\infty$ *and* $\|\nabla f_0\|,\|\nabla f_\delta\|\leq M$.

5.3. Parametric or nonparemtric. In the work presented above, and also in Hermansen & Hjort (2014b), all models worked with are assumed to be nested parametric models in a so-called local large-sample framework, see (5.1) above. The use of nonparametric estimation is a well studied and a common practice in time series modelling and the restriction to only nested parametric models may not always be appropriate, however. The aim of the present section is to motivate an extension that will justify comparison and selection among nonparametric and parametric candidate models. The derivation follows a similar reasoning as in Jullum & Hjort (2014), which discusses focused inference and model selection among parametric and nonparametric models for i.i.d. observations.

By including a nonparametric candidate among the parametric models, we will be able to detect if our parametric models are completely off-target. In this sense, the parametric vs nonparametric selection can behave as an insurance against poorly specified parametric candidates. Furthermore, we usually achieve higher precision with the parametric models when these are sufficient.

Let $y_1,\ldots,y_n$ be realisations from a stationary Gaussian time series model with zero mean and true spectral density F. Note that we do not work in a local large-sample framework. Let $\mu=\mu(F)$ be a focus function, i.e. a functional mapping of the spectral measure F to a scalar value. Suppose that we have a collection of parametric candidate models, represented by F_θ, which do not have to be nested or include the true F. The question is then which model should we use – parametric or nonparametric – for estimating the focus μ.

Let $\widehat{\mu}_{\rm np}=\mu(\widehat{F}_n)$ be the nonparametric estimate for the focus function and suppose

$$\sqrt{n}(\widehat{\mu}_{\rm np}-\mu_{\rm true})\to_d \mathrm{N}(0,v_{\rm np}),$$

where $\mu_{\rm true}=\mu(F)$ is the focus function for the true model F. In addition, assume that our fitted parametric candidates $\widehat{\mu}_{\rm pm}=\mu(F_{\widehat{\theta}_n})$ possess a similar large-sample

property, i.e.

$$\sqrt{n}(\widehat{\mu}_{\rm np} - \mu_0) \to_d {\rm N}(0, v_{\rm pm}),$$

where $\mu_0 = \mu(F_{\theta_0})$ is the focus function evaluated under the least false model F_{θ_0}; all of this will be discussed in more details below. Then, still without going into details, the large-sample results above motivate the following first-order approximations for the mean squared error for the estimated focus parameters:

$$\text{mse}_{\rm np} = 0^2 + v_{\rm np}/n = v_{\rm np}/n \quad \text{and} \quad \text{mse}_{\rm pm} = b^2 + v_{\rm pm}/n \tag{5.11}$$

where $b = \mu_0 - \mu_{\rm true} = \mu(F_{\theta_0}) - \mu(F)$. The remainder of the section will be used to motivate and obtain good estimators for the mean squared errors in (5.11).

To make the general derivation more transparent and also to arrive at more precise answers, we simplify the general framework above and will only study focus functions of the type

$$\mu = \mu(F, h) = \int_{-\pi}^{\pi} h(\omega) f(\omega)\, {\rm d}\omega,$$

where h is a continuous and bounded function on $[-\pi, \pi]$, with potentially a finite number of jump discontinuities. This is a quite general class that includes the covariance/correlation function and also certain smooth functions these. Moreover, we may also use this construction to study specific parts of the spectral density by using indicator functions, see also Gray (2006) for some additional applications.

In the derivation below, the parametric candidates F_θ will be fitted using the Whittle estimator $\widetilde{\theta}_n$ as defined in (1.7) and we will also use the periodogram I_n as a basis for estimating F by $\widetilde{F}_n$ as defined in (1.10); see Brillinger (1975) for alternative and smoothed versions. The Whittle estimates and the nonparametric spectral measure estimator based on the periodogram gives a convenient symmetry which will simplify the derivations below; the use of full maximum likelihood estimation or smoothed periodograms should become a straightforward extension.

The nonparametric and parametric estimator are now given by

$$\widetilde{\mu}_{\rm np} = \int_{-\pi}^{\pi} h(\omega) I_n(\omega)\, {\rm d}\omega = \frac{1}{n} \underline{y}_n^{\rm t} \Sigma_n(h) \underline{y}_n = X_n \quad \text{and} \quad \widetilde{\mu}_{\rm pm} = \int_{-\pi}^{\pi} h(\omega) f_{\widetilde{\theta}_n}(\omega)\, {\rm d}\omega.$$

The following lemma establish the joint limit distribution for estimators above (suitable normalised), which in turn will be used to obtain good approximations for their respective mean squared error estimates.

Lemma 2. *Let $y_1, \ldots, y_n$ be realisations from a stationary Gaussian time series model with spectral density g that is uniformly bounded away from both zero and infinity. If f_θ is two times differentiable with respect to θ and if f_θ, ∇f_θ and $\nabla^2 f_\theta$*

are continuous and uniformly bounded in both ω and θ in a neighbourhood of the least false parameter value $\theta_0 = \arg\min R(\theta)$, where R is as defined in (1.6). Then

$$\begin{pmatrix} \sqrt{n}(\widetilde{\mu}_{\text{np}} - \mu_{\text{true}}) \\ \sqrt{n}(\widetilde{\mu}_{\text{pm}} - \mu_0) \end{pmatrix} \to_d \begin{pmatrix} X \\ c^{\text{t}} J(g, f_{\theta_0})^{-1} U \end{pmatrix} \sim \text{N}_2 \left(0, \begin{pmatrix} v_{\text{np}} & v_c \\ v_c & v_{\text{pm}} \end{pmatrix} \right), \tag{5.12}$$

where

$$v_{\text{np}} = (2\pi)^3 \int_{-\pi}^{\pi} \{h(\omega)g(\omega)\}^2 \, \text{d}\omega \quad \text{and} \quad v_{\text{pm}} = c^{\text{t}} J(g, f_{\theta_0})^{-1} K(g, f_{\theta_0}) J(g, f_{\theta_0})^{-1} c,$$

with J and K as defined in Theorem 2, and $v_c = c^{\text{t}} J(g, f_{\theta_0})^{-1} d$, where $c = \nabla\mu(f_{\theta_0})$ and

$$d = \text{Cov}(X, U) = (2\pi)^3 \int_{-\pi}^{\pi} \frac{\nabla f_{\theta_0}(\omega) h(\omega) g(\omega)^2}{f(\omega)^2} \, \text{d}\omega.$$

Proof. It follows from the results in Dzhaparidze (1986) that

$$\widetilde{\theta}_n - \theta_0 = J(g, \theta_0)^{-1} U_n + o_P(1/\sqrt{n})$$

where

$$\begin{aligned} U_n = \nabla \widetilde{\ell}_n(f_{\theta_0}) &= -\frac{n}{2} \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \nabla \Psi_{\theta_0}(\omega) \, \text{d}\omega - \frac{1}{2\pi} \int_{-\pi}^{\pi} \nabla \Psi_{\theta_0}(\omega) \frac{I_n(\omega)}{f_{\theta_0}(\omega)} \, \text{d}\omega \right\} \\ &= -\frac{1}{2} \{ \text{tr}(\Sigma_n(\nabla \Psi_{\theta_0})) - \underline{y}_n^{\text{t}} \Sigma_n(\nabla \Psi_{\theta_0} / f_\theta) \underline{y}_n \} \end{aligned}$$

and $\Psi_{\theta_0} = \log f_{\theta_0}$ and $\nabla \Psi_{\theta_0}$ is the vector of parietal derivatives. This means that the marginal distribution and the respective mean and variance are easily obtainable by the application of the standard delta method. Since $X_n = \underline{y}_n^{\text{t}} \Sigma_n(h) \underline{y}_n / n$, the joint distribution is readily obtainable by a Cramér–Wold type of argument; we will not go into details on this here. The final piece needed to complete the argument is the limiting covariance

$$\begin{aligned} \text{Cov}(X_n, U_n) = \frac{2}{n} \, \text{tr}\{\Sigma_n(h) \Sigma_n(g) \Sigma_n(\nabla \Psi_\theta / f_\theta) \Sigma_n(g)\} \\ \to (2\pi)^3 \int_{-\pi}^{\pi} \frac{\nabla f_{\theta_0}(\omega) h(\omega) g(\omega)^2}{f(\omega)^2} \, \text{d}\omega, \end{aligned}$$

which can be seen to follow from results in Dzhaparidze (1986) or from Dahlhaus & Wefelmeyer (1996, Lemma A.5). □

The nonparametric estimator is by construction unbiased in the limit, however, to obtain estimates for the mean squared error, we need to derive the squared bias of the parametric model. Following Jullum & Hjort (2014) we start with $b = \mu_0 - \mu_{\text{true}}$, which is estimated with $\widetilde{b} = \widetilde{\mu}_{\text{pm}} - \widetilde{\mu}_{\text{np}}$ and by (5.12) above, we have that

$$\sqrt{n}(\widetilde{b} - b) \to_d c^t J^{-1} U - X \sim \text{N}(0, \kappa),$$

where $\kappa = v_{\rm pm} + v_{\rm np} - 2v_c$. It can further be shown that $\mathrm{E}\,\widetilde{b}^2 \approx b^2 + \kappa/n + o(1/n)$, which leads to the estimators

$$\begin{aligned}\mathrm{FIC}_{\rm np} &= \widetilde{\mathrm{mse}}_{\rm np} = \widetilde{v}_{\rm np}/n \quad \text{and} \\ \mathrm{FIC}_{\rm pm} &= \widetilde{\mathrm{mse}}_{\rm pm} = \widetilde{\mathrm{bsq}} + \widetilde{v}_{\rm pm}/n = \max(0, \widetilde{b}^2 - \widetilde{\kappa}/n) + \widetilde{v}_{\rm pm}/n.\end{aligned}$$

Remark 6. *In the discussion above we have only considered one parametric alternative, the extension to several parametric candidates is straightforward, see Jullum & Hjort (2014) for additional comments.*

Remark 7. *Another class of focus functions that should also be fairly straightforward to work with is given by*

$$\mu(H) = \int_{-\pi}^{\pi} H(f(\omega))\,\mathrm{d}\omega,$$

where H is continuous on the range of f say $[m_f, M_f]$ and where f is the true underlying spectral density; see Grenander & Szegő (1958), Gray (2006) and Taniguchi (1980) and von Sachs (1994) for estimation.

In general, the goal is to extend the above to essentially all reasonable and well behaved functionals of the spectral measure F. For a general functional $T(\cdot)$, let $\mu = T(F)$ and again $\widetilde{F}$ be the spectral measure estimated from the periodogram and $\theta_0 = \arg\min_\theta R(\theta)$, where $R(\theta)$ is the model specific part of (1.6). Similarly, define $\widetilde{\theta}_n = \theta_0(\widetilde{F}_n) = \arg\max_\theta \widetilde{R}_n(\theta)$, where $\widetilde{R}_n(\theta)$ is the R above with the periodogram I_n estimating the unknown spectral density. Then under fairly general model assumptions it follows by similar techniques as used in van der Vaart (1998, Ch. 20) that

$$\begin{aligned}\sqrt{n}(\widetilde{\mu}_{\rm np} - \mu_{\rm true}) &= \sqrt{n}\{T(\widetilde{F}_n) - T(F)\} = \dot{T}(X_n) + o_P(1) \\ \sqrt{n}(\widetilde{\theta}_n - \theta_0) &= \sqrt{n}(\theta_0(\widetilde{F}_n) - \theta_0) = \sqrt{n} J(g,\theta_0)^{-1} U_n + o_P(1)\end{aligned} \tag{5.13}$$

for a suitable functional derivative $\dot{T}$, where $X_n(\omega) = \sqrt{n}\{\widetilde{F}_n(\omega) - F(\omega)\} \to_d X(\omega)$ and $X(\omega)$ is the Wiener process with variance as in (1.12). Then, with the above construction and large-sample results in (5.13), it should be possible to establish a general version of Lemma 2 above. That in turn can be used for derivation and justification of $\mathrm{FIC}_{\rm np}$ and $\mathrm{FIC}_{\rm pm}$ formulas for more general classes of foci.

As a last remark, we note that similar an additional layer of complexity is introduced by studying time series models, since several interesting foci are naturally related to predictions, or formulated conditional on past observations. The dependency on previous data requires a new and extended modelling framework, which in Hermansen & Hjort (2014b, Sections 5 & 6) lead to generalisations and also

motivated the conditional focused information criterion (cFIC). These considerations need to be taken properly into account in a complete extension of the FIC methodology for time series in the framework of selecting among parametric and nonparametric models.

Remark 8. *An alternative approach is to retain the local large-sample framework from the parametric FIC construction and work with spectral densities of the type* $f_r(\omega) = f_{\theta_0}(\omega) + r(\omega)/\sqrt{n}$*, where* f_{θ_0} *is a standard type of parametric model. Such structures have already been worked with in Davis (1973) and Dzhaparidze (1986), making the extension potentially less cumbersome, but this will not be dealt with here, however.*

6. Estimation, inference and model selection for jump regression models

The project was initially based on a short technical report by N. L. Hjort on estimation in regression models with jump discontinuities, or change points; see among others Frick et al. (2014) for a comprehensive discussion and review of the literature. On this basis, we intended to derive an AIC-like 'jump information criterion', for selecting the appropriate number of break points, motivated along the same line of rational arguments used to justify the AIC in Section 3.2. The motivation for doing this is again the common temptation to use $\text{AIC} = 2(\ell_{n,\max} - p)$ as a general model selector, without having the appropriate rational justification. In this sense, the project is in the same explanatory spirit as Grønneberg & Hjort (2014) and Hermansen & Hjort (2014d) which study the rational for using AIC-like criteria for copulas and time series models respectively. The project started out as an investigation of large-sample properties of regression models with jumps, which because of discontinuities at the jump locations, is not possible to solve using standard techniques. The first half of the project, which was included in the PhD thesis of S. Grønneberg, was mainly concerned with establishing the large-sample properties of the model estimators in a potentially misspecified modelling framework. The jump information criterion (AJIC) alluded to above was developed later and is now included in the current version of Grønneberg et al. (2014). There still remain some technical details to complete the project, but we believe this is within reach, however.

6.1. Introduction and summary.

Consider pairs of observations (x_i, y_i), for $i \leq n$, from the regression model

$$y_i = m(x_i, \theta) + \varepsilon_i \quad \text{for } i = 1, \ldots, n, \tag{6.1}$$

where ε_i are zero-mean i.i.d. errors with standard deviation level σ. For convenience and without essential loss of generality we take the x range to be the unit interval, and study the model where

$$m(x,\theta) = a_j \quad \text{for } \gamma_{j-1} \le x < \gamma_j, \tag{6.2}$$

for windows $[\gamma_{j-1},\gamma_j]$, with $j = 1,\ldots,d$ and $\gamma_0 = 0, \gamma_d = 1$. The unknown parameters to estimate from data are the $d-1$ break point positions and the d levels, along with the spread parameter σ.

The case where m is a smooth function is discussed in Grønneberg et al. (2014, Section 2), where proper large-sample motivation is discussed for the model robust AIC formal, along with a justification of the BIC. As already commented on above, the goal is to obtain a jump information criterion derived along the same large-sample arguments that motivates the general AIC formula for i.i.d. observations as in Section 3.2. In order to do this properly, the large-sample properties of the model above are needed outside the model, i.e. in a so-called misspecified modelling framework. To do this requires a fair amount of technical work and extensions of the sargmax principle (cf. Seijo et al. (2011) and Kosorok (2008)) and we do not intend to go into the details on this here, however; see Grønneberg et al. (2014, Section 3).

From this, a well motivated bias correction of $\ell_{n,\max}$ an estimator for the attained expected model specific part of the Kullback–Leibler divergence emerges. The expected bias is

$$b = 1 + d\frac{\sigma^2_{\text{true}}}{\sigma_0^2} + \frac{1}{\sigma_0^2}\sum_{j=1}^{d-1}\kappa_j,$$

for large n, where σ_{true} is related to the variance in the true underlying model and

$$\kappa_j = \sigma_{\text{true}}|a_{0,j+1} - a_{0,j}|\,\mathrm{E}\,W_j^*(\lambda,\widehat{s}_j),$$

and $\widehat{s}_j = \text{argmax}(M_j)$, M_j and W^* are related to certain two-sided compound Poisson processes, see Grønneberg et al. (2014, Sections 3–6) for the complete derivation and additional discussion. This now results in an AIC type of model selection scheme by preferring the model that attains the largest value of

$$\text{AJIC}^* = 2\ell_{n,\max} - 2\widehat{b} = 2(-n\log\widehat{\sigma}_0 - \tfrac{1}{2}n) - 2\Big(1 + d\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2} + \frac{1}{\widehat{\sigma}_0^2}\sum_{j=1}^{d-1}\widehat{\kappa}_j\Big).$$

6.2. Applications and related ideas. In this section we discuss some possible applications with extensions and some related ideas. In general, there are several real life phenomena where the change point model above is reasonable, i.e. where the underlying model has actual discontinuities; examples can be found in biology, geology, medicine, marine biology and oceanography, some of which are discussed in

Frick et al. (2014). In what follows we will discuss an application related to geology.

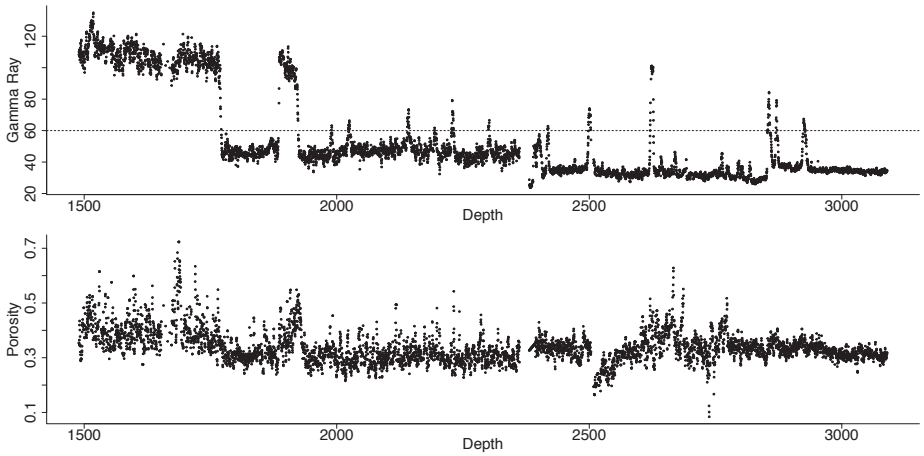

FIGURE 6.1. The gamma ray log (top panel) and the porosity log (bottom panel) from one well from the Sleipner field in the North Sea. The horizontal line (dashed line) is of particular importance, since it is commonly interpreted as the boundary between shale (above) and sand (below).

Figure 6.1 shows well-log observations from one well at the Sleipner field in the North Sea. This shows the observed gamma ray log (top panel), which measures the radioactivity in the well and is commonly used as an indicator for facies changes, i.e. the transition between distinctive rock units/types. Detecting such underlying change points is important in reservoir modelling and flow simulations, since e.g. thin layers of shale can completely block the flow of oil and gas in the reservoir; see e.g. Caers (2005) and references therein. A similar example is discussed in Fearnhead (2006) as an application for a BIC criterion developed; the general approach is quite different with essentially no overlap to the work in Grønneberg et al. (2014), however.

This well-log illustration motivates some potential extensions of the work in Grønneberg et al. (2014). First of all, note that the gamma ray log in Figure 6.1 (top panel) indicates that the variance is not constant across facies. Extending the original work to include one σ_d-parameter for each window should be a straightforward extension. In addition, in Figure 6.1 (bottom panel) we have included the porosity log for the same well. The porosity is strongly associated with flow rates and for obvious reasons this also carries information about the underlying rock type, i.e. facies. Note that several other types measurements are usually also collected from the wells, e.g. permeability, electrical resistivity and density, to name a few. This

suggests that an extension of AJIC to multivariate series of data, to increase the statistical strength, will in some situations be of great interest.

7. A new approach to Akaike's information criterion and model selection issues in stationary time series

This project turned out to be a remarkably hard paper to write. The main reason for this is that it is difficult to give one precise and unified motivation. In the paper we explore the rational for using Akaike's information criterion (AIC) and other related criteria for time series processes. In addition, we try to answer various model selection related questions in time series modelling. The project was motivated from a growing interest in statistical model selection and a general feeling that some information criteria, like the AIC and FPE, do not have the required rational motivation in general classes of time series processes. This was also supported by a remark in Dahlhaus (1996a, p. 184), claiming that the AIC and the model robust version alternative AIC* were not satisfactorily solved for time series models. In order to provide the proper motivation for the AIC as a rational extension of the maximum likelihood principle, for estimation across families of parametric models, the large-sample properties of the maximum likelihood estimator outside the model are needed. The original motivation for the project was to develop the methodology needed and derive the related large-sample properties for the maximum likelihood in a misspecified time series framework. The idea was to extend methodology in Dzhaparidze (1986) and Davis (1973) to obtain the large-sample results needed. After approximately six months of work, however, I discovered Dahlhaus & Wefelmeyer (1996) who had already solved this. As a result, we began developing the framework needed to obtain a proper rational motivation for using the AIC to select among time series models, which eventually resulted in Hermansen & Hjort (2014d); a technical report that still requires some extra work.

7.1. Introduction and summary. The Akaike's information criterion (AIC) is one of the more important and probably also the most widely used model information criterion in statistics and related fields of scientific research. For convenience we repeat its general definition

$$\mathrm{AIC}(M) = 2\,\text{log-likelihood}_{\max}(M) - 2\dim(M),$$

for each candidate model M in a collection of competing models. In practice, the AIC strategy is to prefer the model that attains the largest value of the formula above; for a more complete introduction and comments see Section 3.2 above or Claeskens & Hjort (2008, Ch. 2).

The two main goals of Hermansen & Hjort (2014d) are to investigate the underlying rational motivation for using the AIC and similarly structured model information criteria, like Akaike's final prediction error (FPE), but we will also argue for more use of the model robust version AIC*, which will be introduced in equation (7.4) below. The reason is that there is a common criticism of the AIC, which claims that the criterion has a tendency to prefer unnecessarily complex models when used with classical time series models, see e.g. McQuarrie & Tsai (1998). For this reason, several modifications and adjustments have been suggested to enhance its performance and applicability, e.g. in Hurvich & Tsai (1989) where the original AIC formula is corrected to improve its performance in small samples. In Hermansen & Hjort (2014d) we argue that the failure of the AIC in such cases is more related to the method of estimation than the general structure of the criterion. Moreover, we also demonstrate that this bias towards large models can be further avoided by using the model robust AIC*.

The general AIC formula above has a natural motivation and justification in most classical models for i.i.d. observations and can be justified as a canonical extension of the maximum likelihood principle, for estimation across families of parametric models. There is essentially no reason to expect this relationship to be true in general. In Hermansen & Hjort (2014d) we show that such a connection and motivation exist for time series processes and we will now give a short review that will work as a stepping stone towards the discussion in Section 7.2 below.

Let $\{Y_t\}$ be a stationary Gaussian time series with true spectral density g and let f_θ represent a parametric candidate from a suitable set of candidate models that do not necessarily include or span the true model g. Then, observe that there is a potential ambiguity, or non-uniqueness, in the definition of the Kullback–Leibler for non-i.i.d. data, in that we may define

$$\mathrm{KL}_n(g, f_\theta) = d_n(g, f_\theta) = \frac{1}{n}\mathrm{E}_g\left\{\ell_n(g) - \ell_n(f_\theta)\right\}, \tag{7.1}$$

where ℓ_n is the Gaussian log-likelihood function, with limit

$$\mathrm{KL}_\infty(g, f_\theta) = \lim_{n\to\infty} \mathrm{KL}_n(g, f_\theta) = -\frac{1}{4\pi}\int_{-\pi}^{\pi} (\log g(\omega) + 1)\,\mathrm{d}\omega - R(\theta), \tag{7.2}$$

and where we refer to

$$R(\theta) = -\frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log f_\theta(\omega) + \frac{g(\omega)}{f_\theta(\omega)}\right)\mathrm{d}\omega$$

as the model specific part of the asymptotic Kullback–Leibler divergence (7.2). From Section 1.1 above it now follows that $\widehat{\theta}_n \to_{P_g} \theta_0 = \arg\min_\theta d(g, f_\theta)$, which in turn (by similar arguments as used in Section 3.3) can be shown to motivate an AIC-like

model information criterion

$$\mathrm{AIC}^*_\infty = 2(\widetilde{\ell}_{n,\max} - \widetilde{p}^* - \widetilde{q}^*), \tag{7.3}$$

where $\widetilde{p}^* = \mathrm{tr}(\widetilde{J}^{-1}\widetilde{K})$ and and where $\widetilde{J} = J(I_n, f_{\widetilde{\theta}_n})$ and $\widetilde{K} = K(I_n/\sqrt{2}, f_{\widetilde{\theta}_n})$ are as defined in Theorem 2 of Section 1.1 and $\widetilde{q}^*$ is an additional bias correction term introduced by the first order bias of the periodogram as an estimator for the underlying spectral density g, see Hermansen & Hjort (2014d, Section 3.2) for details. Finally, it is easily checked that if f_{θ_0} is equal to g (a.e.) then $J = K$ and $p^* = p$.

Let $\{\theta_{0,n}\}_{n\geq 1}$, where $\theta_{0,n} = \arg\min_\theta d_n(g, f_\theta)$ and d_n as defined in (7.1), be a sequence of least false parameter values. This sequence can now be interpreted as the target for the maximum likelihood estimator in the sense that $\|\widehat{\theta}_n - \theta_{0,n}\| \to_{P_g} 0$. This enables a motivate for a more classical AIC type of criterion by

$$\mathrm{AIC}^*_n = 2(\ell_{n,\max} - \widehat{p}^*) \tag{7.4}$$

where $\widehat{p}^* = \mathrm{tr}(\widehat{J}^{-1}\widehat{K})$, where $\widehat{J} = J(I_n, f_{\widehat{\theta}_n})$ and $\widehat{K} = K(I_n/\sqrt{2}, f_{\widehat{\theta}_n})$ are as defined below in (7.3).

The second criterion AIC^*_∞, which is related to $\mathrm{KL}_\infty(g, f_\theta)$ above, is also naturally connected to the Whittle approximation in (1.7). It is important to note, however, that AIC^*_∞ is not an approximation of AIC^*_n; each criterion has its own internal and independent rational justification. Moreover, in Hermansen & Hjort (2014d) we show that it will not make sense to use $\widetilde{\mathrm{AIC}}^* = 2(\widetilde{\ell}_n(f_{\widetilde{\theta}}) - \widetilde{p}^*)$ as an approximation to AIC_n unless all candidate models span the true model model, which would be a rather strange and unnatural assumption.

As a final remark, we point out that the two discrepancy measures KL_n vs KL_∞ are perhaps best viewed as discrepancy measures concerned with different parts of the model. The KL_n measures a type of average performance in a new sample, of the same size, while the limit Kullback–Leibler divergence KL_∞ is concerned with the performance of the entire process. From this perspective, the two AIC formulations AIC_∞ and AIC_n are simply two different criteria based on different discrepancy measures aiming at answer different questions.

The main concern of Hermansen & Hjort (2014d) is to motivate and redevelop the AIC for stationary time series. In addition, we use the developed methodology to improve on a generalisation of Akaike's final prediction error (FPE; Akaike (1969)) introduced in Linhart & Göttingen (1985), where we derive a more correct bias correction. Moreover, we use the general methodology to motivate a class of model information criteria built for the frequency domain representation of time series models, see Hermansen & Hjort (2014d, Section 6) for details and additional discussion.

7.2. The AIC for stationary time series processes with smooth trend. A natural extension of the work in Hermansen & Hjort (2014d) is to develop similar methodology and rational large-sample justification for using the AIC in stationary time series models that also includes a smooth or regression type of trend. A sketch of the derivation needed for this extension is now presented. The derivation is not complete, in the sense that this requires consistent nonparametric estimates of both trend function and spectral density, which is hard to do in practice.

Consider the model

$$Y_t = Y_{n,t} = \mu(t/n) + \epsilon_t, \quad \text{for } t = 1, \ldots, n, \tag{7.5}$$

where μ is a smooth trend function on the unit interval and $\{\epsilon_t\}$ is a stationary Gaussian time series model with zero mean and true spectral density g. The construction with triangular array may seem artificial, which it is, however, it is an abstract construction needed to make sure that we will have a meaningful asymptotic theory, e.g. it guarantees that more observations will provide more information about this underlying signal, see Dahlhaus (1997, Section 2) for additional comments and discussion.

For a particular parametric candidate, represented by the trend $m(\cdot, \beta)$ and spectral density f_θ, for $(\beta, \theta) \in \mathbb{R}^{p+q}$, the full Gaussian log-likelihood is given by

$$\ell_n(\beta, \theta) = -\frac{n}{2}\left\{\log(2\pi) + \log|\Sigma_n(f_\theta)| + \frac{1}{n}(\underline{y}_n - \underline{m}_\beta)^{\mathrm{t}}\Sigma_n(f_\theta)^{-1}(\underline{y}_n - \underline{m}_\beta)\right\}, \tag{7.6}$$

where $\underline{m}_\beta = (m(1/n, \beta), \ldots, m(1, \beta))^{\mathrm{t}}$ and where we do not necessarily assume that our set of candidate models span the true model. Then, under the conditions of Proposition 3 it now follows from Dahlhaus (1996b) that the corresponding maximum likelihood estimators $(\widehat{\beta}_n, \widehat{\theta}_n) = \arg\max_{(\beta,\theta)} \ell_n(\beta, \theta)$ converges in probability to value $(\beta_0, \theta_0) = \arg\min_{(\beta,\theta)} d(\beta, \theta)$, where

$$d(\beta, \theta) = \lim_{n\to\infty} \frac{1}{n}\mathrm{E}\left\{\ell_n(\mu, g) - \ell_n(\beta, \theta)\right\} = -\frac{1}{4\pi}\int_{-\pi}^{\pi}(\log g(\omega) + 1)\,\mathrm{d}\omega - R(\beta, \theta), \tag{7.7}$$

and where

$$\begin{aligned} R(\beta, \theta) &= \lim_{n\to\infty} R_n(\beta, \theta) \\ &= \lim_{n\to\infty} \frac{1}{n}\mathrm{E}\,\ell_n(\beta, \theta) \\ &= -\frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\log f_\theta(\omega) + \frac{g(\omega)}{f_\theta(\omega)}\right]\mathrm{d}\omega \\ &\quad - \frac{1}{2\pi f_\theta(0)}\int_0^1 (\mu(u) - m(u, \beta))^2\,du - \log(2\pi), \end{aligned}$$

will be referred to as the model specific part of (7.7). The following proposition summarises the main large-sample properties of the maximum likelihood estimator above.

Proposition 3. *(Dahlhaus, 1996a, Theorem 2.4) Suppose the spectral density f_θ and g are continuous and bounded away from zero and infinity. In addition, suppose f_θ and g and all components of ∇f_θ and $\nabla^2 f_\theta$ are differentiable in ω with uniformly continuous derivatives. Then if the components of $\mu(u)$, $m(u,\beta)$, $\nabla m(u,\beta)$ and $\nabla^2 m(u,\beta)$ are differentiable in u with uniformly continuous derivatives, it follows that*

$$\begin{pmatrix} \sqrt{n}(\widehat{\beta}_n - \beta_0) \\ \sqrt{n}(\widehat{\theta}_n - \theta_0) \end{pmatrix} \to_d J(\beta_0,\theta_0)^{-1} U, \qquad \text{with } U \sim \mathrm{N}_{p+q}(0, K(\beta_0,\theta_0))$$

where $J(\beta_0,\theta_0)$ and $K(\beta_0,\theta_0)$ are block diagonal

$$\begin{aligned} J(\beta_0,\theta_0) &= \begin{bmatrix} J_{01}(\beta_0,\theta_0) & 0 \\ 0 & J_{11}(\beta_0,\theta_0) \end{bmatrix} \quad \text{and} \\ K(\beta_0,\theta_0) &= \begin{bmatrix} K_{01}(\beta_0,\theta_0) & 0 \\ 0 & K_{11}(\beta_0,\theta_0) \end{bmatrix} \end{aligned} \tag{7.8}$$

and the corresponding block matrices are given by

$$J_{01}(\beta_0,\theta_0) = J(\theta_0) + \frac{\nabla^2\Psi_{\theta_0}(0) - \nabla\Psi_{\theta_0}(0)\nabla\Psi_{\theta_0}(0)^{\mathrm{t}}}{2\pi f_{\theta_0}(0)} \int_0^1 [\mu(u) - m(u,\beta_0)]^2 \,\mathrm{d}u,$$

$$J_{11}(\beta_0,\theta_0) = \frac{1}{2\pi f_{\theta_0}(0)} \int_0^1 \{\nabla m(u,\beta_0)\nabla m(u,\beta_0)^{\mathrm{t}} - \nabla^2 m(u,\beta_0)[\mu(u) - m(u,\beta_0)]\} \,\mathrm{d}u,$$

and

$$K_{01}(\beta_0,\theta_0) = \frac{g(0)}{2\pi f_{\theta_0}(0)^2} \int_0^1 \nabla m(u,\beta_0)\nabla m(u,\beta_0)^{\mathrm{t}} \,\mathrm{d}u,$$

$$K_{11}(\beta_0,\theta_0) = K(\theta_0) + \frac{\nabla\Psi_{\theta_0}(0)\nabla\Psi_{\theta_0}(0)^{\mathrm{t}}}{2\pi f_{\theta_0}(0)^2} g(0) \int_0^1 [\mu(u) - m(u,\beta_0)]^2 \,\mathrm{d}u,$$

with $J(\theta) = J(g, f_\theta)$ and $K(\theta) = K(g, f_\theta)$ as defined in Theorem 2 in Section 1.1.

Proof. The result is essentially a special case of Dahlhaus (1996a, Theorem 2.4) and can therefore be seen to follow automatically as a corollary. □

Next, we define $Q = \mathrm{E}\, R(\widehat{\beta}_n, \widehat{\theta}_n)$. By following the by now familiar recipe of Section 3.2 (and Section 7.1 above), a model selection strategy is easily motivated by aiming at the candidate model that maximises Q, for which we need a suitable estimator. Previously, the canonical starting point is to work with a nonparametric version of $R(\widehat{\beta}_n, \widehat{\theta}_n)$, however, this becomes unpractical since this requires joint nonparametric estimation of both trend and spectral density, which are not readily

available. For this reason, we decide to change our focus here and will instead aim at deriving an unbiased estimator for $Q_n = \mathrm{E}\, R_n(\widehat{\beta}_n, \widehat{\theta}_n)$ given by

$$R_n(\beta,\theta) = \frac{1}{n}\{\log|\Sigma_n(f_\theta)| + \mathrm{tr}\{\Sigma_n(g)\Sigma_n(f_\theta)^{-1}\} + (\underline{\mu}-\underline{m}_\beta)^{\mathrm{t}}\Sigma_n(f_\theta)^{-1}(\underline{\mu}-\underline{m}_\beta)\} \quad (7.9)$$

where $\underline{\mu}^{\mathrm{t}} = (\mu(1/n),\ldots,\mu(1))$ and $\underline{m}_\beta^{\mathrm{t}} = (m(1/n,\beta),\ldots,m(1,\beta))$. This change indirectly introduces a new and alternative framework for interpretation, which either requires the extended likelihood principle of Akaike (1973) or the construction of a sequence of least false parameter values $\{(\beta_{0,n},\theta_{0,n})\}_{n\geq 1}$, defined by $(\beta_{0,n},\theta_{0,n}) = \arg\min_{(\beta,\theta)} d_n(\beta,\theta)$ with

$$d_n(\beta,\theta) = \mathrm{KL}_n(\beta,\theta) = \frac{1}{n}\mathrm{E}\,\ell_n(\mu,g) - R_n(\beta,\theta), \quad (7.10)$$

see Section 3.2 above or Hermansen & Hjort (2014d, Sections 3 & 4) for a more complete discussion and comments. Now, observe that

$$\widehat{R}_n(\widehat{\theta}_n,\widehat{\beta}_n) - R_n(\widehat{\theta}_n,\widehat{\beta}_n) = (\Delta_n - \delta_n) + \epsilon_n, \quad (7.11)$$

where $\Delta_n = \widehat{R}_n(\widehat{\beta}_n,\widehat{\theta}_n) - \widehat{R}_n(\beta_{0,n},\theta_{0,n})$, $\delta_n = R_n(\widehat{\beta}_n,\widehat{\theta}_n) - R_n(\beta_{0,n},\theta_{0,n})$ and $\epsilon_n = \widehat{R}_n(\theta_0) - R_n(\theta_0)$, where by similar arguments as used in Hermansen & Hjort (2014d, Section 4.1) it is easily seen that $\mathrm{E}\,\epsilon_n = 0$. The following lemma is the final argument needed to obtain a proper rational motivation and the model robust version of the AIC in the current framework.

Lemma 3. *Under the conditions of Proposition 3 we have*

$$\Delta_n - \delta_n = \frac{1}{n}\{W_n + o_p(1)\},$$

where $W_n \to_d W$, *as* $n\to\infty$, *with* $\mathrm{E}\,W = \mathrm{tr}\{J(\beta_0,\theta_0)^{-1}K(\beta_0,\theta_0)\}$ *and where* J *and* K *are as defined in (7.8).*

Proof. The proof is left as an exercise to the reader. □

Summarising the chain of large-sample arguments motivate

$$\mathrm{AIC}^* = 2\{\ell_{n,\max} - (\widehat{p}^* + \widehat{q}^*)\}, \quad \text{where } \widehat{p}_1^* = \mathrm{tr}(\widehat{J}_{01}^{-1}\widehat{K}_{01}) \quad \text{and} \quad \widehat{q}^* = \mathrm{tr}(\widehat{J}_{11}^{-1}\widehat{K}_{11}),$$

as a rational model selection strategy. In order to use the above formula in practice, we need consistent estimates for the block elements of J and K, which presumably does not have a simple solution, since we will be needing consistent estimates for the true g and μ. Some alternatives are: to use the widest model (if the models are nested), bootstrap or work under the assumption that $J = K$, which can be seen to imply that $p^* = \dim(\beta)$ and $q^* = \dim(\theta)$ with the result that $\mathrm{AIC} = 2\{\ell_{n,\max} - (p+q)\}$. As a final comment, note that there is a nice symmetry in that if $m_{\beta_0} = \mu$, then the AIC^* formula above is easily seen to reduce to the the formula obtained in Section 7.1 for the standard stationary models with zero mean.

Bibliography

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics* **21**, 243–247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*. pp. 267–281.

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

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

Billingsley, P. (2009). *Convergence of Probability Measures*. John Wiley & Sons.

Blackwell, D. et al. (1951). Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, vol. 1.

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston.

Brockwell, P. J. & Davis, R. (1991). *Time Series: Theory and Methods*. Springer.

Caers, J. (2005). *Petroleum Geostatistics*. Society of Petroleum Engineers Richardson.

Carslaw, H. S. (1921). *Introduction to the Theory of Fourier's Series & Integrals*. Glasgow university press.

Choudhuri, N., Ghosal, S. & Roy, A. (2004a). Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association* **99**, 1050–1059.

Choudhuri, N., Ghosal, S. & Roy, A. (2004b). Contiguity of the Whittle measure for a Gaussian time series. *Biometrika* **91**, 211–218.

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

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Coursol, J. & Dacunha-Castelle, D. (1982). Remarks on the approximation of the likelihood function of a stationary Gaussian process. *Theory of Probab. Appl.* **27**, 162–167.

Dahlhaus, R. (1996a). Maximum likelihood estimation and model selection for locally stationary processes. *Journal of Nonparametric Statistics* **6**, 171–191.

Dahlhaus, R. (1996b). On the Kullback–Leibler information divergence of locally stationary processes. *Stochastic Processes and their Applications* **62**, 139–168.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics* **15**, 1–37.

Dahlhaus, R. & Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Annals of Statistics* **24**, 952–973.

Davis, R. B. (1973). Asymptotic inference in stationary Gaussian time-series. *Advances in Applied Probability* **5**, 469–497.

Diaconis, P. & Freedman, D. (1986a). On inconsistent Bayes estimates of location. *Annals of Statistics* , 68–87.

Diaconis, P. & Freedman, D. (1986b). On the consistency of Bayes estimates (with discussion). *Annals of Statistics* **14**, 1–26.

Dzhaparidze, K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple change-point problems. *Statistics and computing* **16**, 203–213.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* , 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* , 615–629.

Ferguson, T. S. (1996). *A course in Large Sample Theory*. Chapman & Hall London.

Florens, J.-P., Mouchart, M., Rolin, J.-M. et al. (1990). *Elements of Bayesian statistics*. Marcel Dekker.

Frick, K., Munk, A. & Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 495–580.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate*

Analysis **74**, 49–68.

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

Golubev, G. K., Nussbaum, M., Zhou, H. H. et al. (2010). Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Annals of Statistics* **38**, 181–214.

Gray, R. M. (2006). *Toeplitz and Circulant Matrices: A Review*. now publishers Inc.

Grenander, U. & Szegő, G. (1958). *Toeplitz Forms and Their Applications*. University of California Press.

Grønneberg, S., Hermansen, G. H. & Hjort, N. L. (2014). Estimation, inference and model selection for jump regression models. Tech. rep., Norwegian Business School and University of Oslo and Norwegian Computing Centre.

Grønneberg, S. & Hjort, N. L. (2014). The copula information criteria. *Scandinavian Journal of Statistics* **41**, 436–459.

Hermansen, G. H. (2008). *Bayesian nonparametric modelling of covariance functions, with application to time series and spatial statistics*. Master's thesis, Department of Mathematics, University of Oslo.

Hermansen, G. H. & Hjort, N. L. (2014a). Bernshteĭn–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach. Tech. rep., University of Oslo and Norwegian Computing Centre.

Hermansen, G. H. & Hjort, N. L. (2014b). Focused information criteria for time series. Tech. rep., University of Oslo and Norwegian Computing Centre.

Hermansen, G. H. & Hjort, N. L. (2014c). Limiting normality of quadratic forms, with applications to time series analysis. Tech. rep., University of Oslo and Norwegian Computing Centre.

Hermansen, G. H. & Hjort, N. L. (2014d). A new approach to Akaike's information criterion and model selection issues in stationary Gaussian time series. Tech. rep., University of Oslo and Norwegian Computing Centre.

Hjort, N., Holmes, C., Müller, P. & Walker, S., eds. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

Hjort, N. L. & Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. *Advances in statistical modeling and inference* **3**, 463–492.

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

Ibragimov, I. (1963). On estimation of the spectral function of a stationary Gaussian process. *Theory of Probability & Its Applications* **8**, 366–401.

Jullum, M. & Hjort, N. L. (2014). Parametric or nonparametric: The FIC approach. Tech. rep., University of Oslo and Norwegian Computing Centre.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

Le Cam, L. (1996). Comparison of experiments: A short review. *Lecture Notes-Monograph Series* **30**, 127–138.

Le Cam, L. & Yang, G. L. (1990). *Locally Asymptotically Normal Families*. Springer.

Le Cam, L. & Yang, G. L. (2000). *Asymptotics in Statistics: Some basic Concepts*. Springer.

Linhart, H. & Göttingen, P. V. (1985). On a criterion for selection of models for stationary time series. *Metrika* **32**, 181–196.

Lo, A. Y. (1983). Weak convergence for Dirichlet processes. *Sankhyā: The Indian Journal of Statistics, Series A* , 105–111.

McQuarrie, A. D. R. & Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics* **27**, 105–126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scandinavian Journal of Statistics* **26**, 373–393.

Priestley, M. B. (1981). Spectral analysis and time series .

Roussas, G. G. (1972). *Contiguity of Probability Measures: Some Applications in Statistics*. Cambridge University Press.

Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism* **3**, 13–41.

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

Seijo, E., Sen, B. et al. (2011). A continuous mapping theorem for the smallest argmax functional. *Electronic Journal of Statistics* **5**, 421–439.

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

Taniguchi, M. (1980). On estimation of the integrals of certain functions of spectral density. *Journal of Applied Probability* **17**, 73–80.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

VON SACHS, R. (1994). Estimating non-linear functions of the spectral density, using a data-taper. *Annals of the Institute of Statistical Mathematics* **46**, 453–474.

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer.

WHITTLE, P. (1953). The analysis of multiple stationary time series. *Journal of the Royal Statistical Society. Series B (Methodological)* , 125–139.

Chapter 2

Paper 1: Bernshteĭn–von Mises theorems for nonparametric function analysis via locally constant modelling: A unified approach

Chapter 3

Paper 2: Focused information criteria for time series

Chapter 4

Technical report 1: Estimation, inference and model selection for jump regression models

Chapter 5

Technical report 2: A new approach to Akaike's information criterion and model selection issues in stationary time series

A NEW APPROACH TO AKAIKE'S INFORMATION CRITERION AND MODEL SELECTION ISSUES IN STATIONARY TIME SERIES

GUDMUND HORN HERMANSEN[1,2] AND NILS LID HJORT[2]

Abstract. For various classical models, the Akaike's information criterion (AIC) is best motivated as a coherent extension of the maximum likelihood principle, for estimation across families of parametric models. In a framework involving potentially misspecified candidate models the AIC or close relatives may be justified from a precise and well motivated chain of large-sample approximations. This rational line of reasoning does not necessarily hold up in general, however. The main purpose of this paper is to revisit the underlying justification for using the general AIC formula as a model selection strategy for stationary time series processes. The Kullback–Leibler divergence from the true model to its parametric approximation is an important component of this argument. For non-i.i.d. observations, there is a potential non-uniqueness in the definition and interpretation of the Kullback–Leibler divergence. This is often unproblematic, depending on the underlying purpose. For time series processes this ambiguity does matter, however, which motivates two natural variations of the AIC formula. Thus the general AIC line of strategy is not as seemingly uniquely and simply defined as it is often presented as being. In addition, we derive certain model robust versions of the AIC. Finally, using the developed methodology, we are able to improve on a generalisation of Akaike's final prediction error criterion (FPE), and also derive a general class of model information criteria built for the frequency domain representation of time series models.

1. Introduction and summary

The task of selecting an appropriate model is an important part of any statistical analysis. This is especially true for the classical time series models like the autoregressive (AR), moving average (MA) and the mixture (ARMA), where an appropriate model order is required before the models can be put into real use. This makes both model selection and assessment an integrated part of the model building. The so-called model information criteria have therefore a considerable appeal, since these are typically simple in both structure and use, resulting in scores which can be used to rank candidate models from best to worst, in accordance with some predefined measure of discrepancy. Moreover, the time series models referred to above are also very adaptable and all can be shown to approximate essentially any type of stationary process, to any degree of accuracy, by increasing the model complexity, see Brockwell & Davis (1991, Ch. 4.4). Therefore, if coupled with an appropriate

Key words and phrases. AIC, autoregressive processes, FPE, Kullback–Leibler divergence, model selection, model misspecification, stationary time series.

[1]Norwegian Computing Centre.

[2]Department of Mathematics, University of Oslo.

model information criterion, we essentially have a mechanised method for both model fitting and assessment. This indicates the potential role of model selection procedures and also underlines the importance and need for proper understanding and rational justification.

There is a substantial literature on model selection for time series processes, each method derived and justified from different points of departure. Among the more popular are Akaike's information criterion (AIC; Akaike (1973)), Bayesian information criterion, (BIC; Schwarz (1978)), Hannan-Quinn information criterion (HQ; Hannan & Quinn (1979)), Akaike's final prediction error (FPE; Akaike (1969)) and the focused information criterion (FIC; Hermansen & Hjort (2014)); see also Akaeĕl (1982) and the final chapters of Linhart & Zucchini (1986) for a general overview. A comprehensive survey focusing on consistency in autoregressive processes, i.e. the ability to discover the correct, or underlying true, model order, can be found in McQuarrie & Tsai (1998). Finally, for a broad introduction to model selection in statistics, see among others Claeskens & Hjort (2008).

In this paper we focus on the underlying motivation for using the AIC for stationary time series processes. This criterion is one of the more important and probably also the most widely used. In its general form the AIC is defined by

$$\mathrm{AIC}(M_\theta) = 2\,\text{log-likelihood}_{\max}(M_\theta) - 2\dim(\theta) \tag{1.1}$$

for each candidate model M_θ in a collection of competing models, where $\dim(\theta)$ is the length of its parameter vector. The AIC strategy is then to prefer the model that attains the largest AIC-score, i.e. the model that has the largest value of (1.1).

In several classical models, like parametric models for i.i.d. observations and standard regression with independent errors, the AIC is a well motivated model selection strategy. In short, the AIC can be seen as a rational extension of the maximum likelihood principle to estimation across families of parametric models in a potentially misspecified modelling framework, i.e. where the true model is not necessarily included, or spanned, by any of the candidates, see Claeskens & Hjort (2008, Ch. 2). This provides a more sophisticated interpretation of the penalty term $p = \dim(\theta)$ in (1.1), which is actually a bias correction term needed to make the AIC an unbiased estimate for a specific large-sample quantity; this will be discussed in more detail in Sections 2–4 below.

The main purpose of this paper is to obtain good answers to whether a similar – or any – rational justification exist for using the AIC machinery for selecting among time series models. In particular, we are interested in seeing how much of the rational motivation alluded to above that carries over from classical models. This turns out to be quite complicated, however. To indicate what becomes problematic, let $X_1, \ldots, X_n$ be i.i.d. variables from a model with density function h° and let h_θ be a parametric model from a set of candidates. The Kullback–Leibler divergence (cf. Kullback & Leibler

(1951)), an asymmetric measure between probability distributions, is defined as

$$\mathrm{KL}_n(h^\circ, h_\theta) = \frac{1}{n}\mathrm{E}_{h^\circ}\left\{\sum_{i=1}^{n}\log\frac{h^\circ(X_i)}{h_\theta(X_i)}\right\} = \int_{\mathbb{R}}\log\frac{h^\circ(x)}{h_\theta(x)}h^\circ(x)\,\mathrm{d}x = \mathrm{KL}_1(h^\circ, h_\theta) \tag{1.2}$$

where the subscript h° indicates that the expectation is with respect to the true model. From the internal structure it is clear that for i.i.d. observations the Kullback–Leibler divergence, as defined above, is independent of the size of the observed sample. This property does not necessary hold for non-i.i.d. observations and this is partly why the AIC argument becomes more ambiguous for the time series processes.

Let Y_t, for $t \geq 1$, be a stationary Gaussian time series with true spectral density g and let f_θ, with $\theta \in \mathbb{R}^p$ and p finite, be a parametric candidate from a set of competing models that do not necessarily include the true g. Moreover, let $\ell_n(g)$ and $\ell_n(f_\theta)$ be the corresponding Gaussian log-likelihood functions, then the Kullback–Leibler divergence (1.2) can now be written as

$$\mathrm{KL}_n(g, f_\theta) = \frac{1}{n}\mathrm{E}_g\left\{\ell_n(g) - \ell_n(f_\theta)\right\}, \quad \text{with limit } \mathrm{KL}_\infty(g, f_\theta) = \lim_{n\to\infty}\mathrm{KL}_n(g, f_\theta) \tag{1.3}$$

provided that it exists.

In Sections 3 and 4 we show that both KL_n and KL_∞ will emerge naturally from different attempts to extend the maximum likelihood principle to estimation across families of parametric models, resulting in two versions of the AIC formula with different underlying motivations. The derived criteria will not always give the same answers and are given by

$$\mathrm{AIC}_n = 2(\ell_{n,\max} - p) \quad \text{and} \quad \mathrm{AIC}_\infty = 2(\widetilde{\ell}_{n,\max} - p - \widetilde{q}), \tag{1.4}$$

where $\widetilde{\ell}_n$ is the Whittle log-likelihood, an approximation to the full Gaussian log-likelihood introduced in Whittle (1953), and $\widetilde{q}$ is an additional correction term that has to be estimated from data. It is important to note that AIC_∞ is not an approximation to AIC_n, however, since each criterion has its own internal and independent rational justification.

The typical behaviour of the formulae in (1.4) is illustrated in the simulated example in Figure 1.1 and Table 1.

k	0	1	2	3	4	5	6	7	8	9
AIC_n	-89.2	-91.2	-90.9	-89.8	-72.6	-69.4	-68.7*	-70.7	-72.6	-70.1
AIC_w	-89.2	-91.2	-90.3	-87.6	-67.6**	-67.6*	-69.1	-71.1	-71.7	-69.3
p	1	2	3	4	5	6	7	8	9	10
$\widetilde{q}$	0.0	0.0	-0.3	-0.9	-6.00	-6.1	-6.2	-6.2	-6.7	-8.3

TABLE 1. The AIC scores from Figure 1.1, note the jump in q as the estimated models begin to provide a reasonable fit of the data.

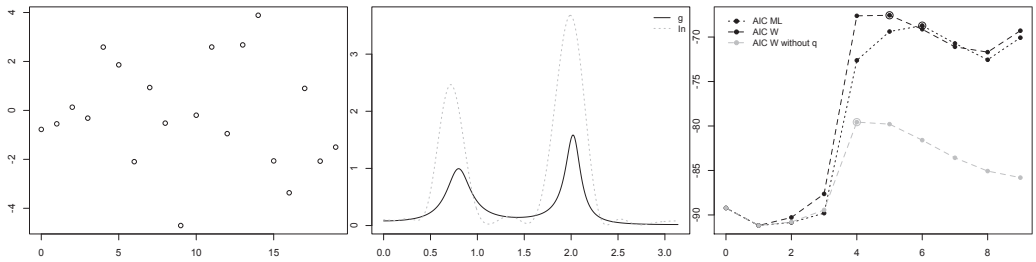

FIGURE 1.1. Left: Simulated sample of size $n = 20$ from an autoregressive models with parameters $\sigma = 1$ and $\rho = (1, 0.4, -0.6, 0.4, -0.6)$. Middle: The spectral density of the autoregressive model (solid) and estimated periodogram (dotted). Right: AIC scores, see also Table 1.

From a practical point of view, obtaining maximum likelihood estimates is sometimes challenging in time series models and can also be computationally time consuming and numerically unstable. The Whittle approximation alluded to above solves most of these problems. In Dahlhaus & Wefelmeyer (1996) the maximum likelihood estimator and the analogously defined Whittle estimator $\widetilde{\theta}_n = \arg\max_\theta \widetilde{\ell}_n(\theta)$ are shown to have the same large-sample properties in terms of limit distributions and efficiency, see also Dzhaparidze (1986). This suggests that the Whittle estimator can be used as an alternative to full maximum likelihood estimation. Moreover, under mild regularity conditions it follows from Coursol & Dacunha-Castelle (1982) that $\ell_n(f) = \widetilde{\ell}_n(f) + O_{P_g}(1)$, uniformly in f, it is therefore tempting, since the AIC is essentially motivated by large-sample arguments, to question whether $\widetilde{\text{AIC}}_n = 2(\widetilde{\ell}_{n,\max} - p)$ can be used as an approximation for the general AIC_n formula without changing the underlying interpretation. In Section 3 we show that this is usually not the case, and that it only works under certain strong and quite restrictive modelling assumptions. The conclusion is that the approximation $\widetilde{\text{AIC}}_n$ is generally only meaningful if we include $\widetilde{q}$, making it equal to AIC_∞ and therefore also changing the underlying interpretation.

In Section 2 we will discuss the case of parametric density estimation for i.i.d. observations and show how to motivate the AIC as a rational extension of the maximum likelihood principle. This motivate the derivations in Section 3 where we aim at a similar line of reasoning for the stationary time series processes; this results in the AIC_∞ and an independent motivation for the Whittle approximation and related estimates. In addition, proper model robust versions of the AIC will also be derived, these are commonly referred to as the Takeuchi's information criterion (TIC; Takeuchi (1976)). In Section 4 we show how to motivate and obtain the more classical AIC formula (1.1) for the time series processes. This is in a sense more in line with the original work in Akaike (1973, 1974), however, we generally see the motivation in Section 3 as more coherent. Two additional approaches for motivating AIC-like criteria are explored in Section 5. In Section 6 we use the developed methodology to revisit and give a proper correction of a criterion derived in Linhart & Göttingen (1985) as a generalisation of the final prediction error (FPE); a criterion that approximates the mean squared error of the one-step-ahead

predictions originally suggested in Akaike (1969, 1970). In Section 7 we review a common criticism of the AIC that claims that the criterion has a tendency to prefer too complex models in small samples, see e.g. Hurvich & Tsai (1989) or McQuarrie & Tsai (1998). We will argue that this is more related to the method of estimation, which is typically based on simplifications of full maximum likelihood estimation, see e.g. McQuarrie & Tsai (1998, Ch. 3) for illustrations. Finally, some concluding remarks are offered in Section 8.

2. The AIC for parametric models for i.i.d. data.

The aim of this section is to introduce the standard derivation that motivates the general AIC formula (1.1) as an extension of the maximum likelihood principle to estimation across families of parametric models. This will be discussed in the framework of parametric density estimation for i.i.d. observations.

The following derivation of the AIC relies heavily on the large-sample properties of the maximum likelihood estimator in a potentially misspecified modelling framework. In order to see how, let the true model be represented by the density function h° and let h_θ be a parametric candidate from a set of candidate models, where $\theta \in \Theta \subset \mathbb{R}^p$ and p is finite and h° is not equal to h_θ almost everywhere for any θ. Let $X_1, \ldots, X_n$ be i.i.d. observations from the true model h° and let ℓ_n be the corresponding log-likelihood function. Then the maximum likelihood estimator $\widehat{\theta}_n = \arg\max_\theta \ell_n(\theta)$ converges, under mild regularity conditions, almost surely to the point

$$\theta_0 = \arg\min_{\theta\in\Theta} \mathrm{KL}_1(h^\circ, h_\theta) = \arg\max_{\theta\in\Theta} R(\theta), \tag{2.1}$$

where

$$\mathrm{KL}_1(h^\circ, h_\theta) = \int \log\frac{h^\circ(x)}{h_\theta(x)} h^\circ(x)\,\mathrm{d}x = \int h^\circ(x)\log h^\circ(x)\,\mathrm{d}x - \mathrm{E}_{h^\circ}\log h_\theta(X) \tag{2.2}$$

is the Kullback–Leibler divergence and where we refer to

$$R(\theta) = \mathrm{E}_{h^\circ}\log h_\theta(X) = \int h^\circ(x)\log h_\theta(x)\,\mathrm{d}x$$

as the model specific part, see Claeskens & Hjort (2008) for additional comments. The subscript h° above is there to indicate that the expectation is with respect to the true model. Note that the limit point θ_0 in (2.1) does not have the interpretation as the true parameter value, as this does not necessarily exist in a misspecified modelling framework. Instead, the maximum likelihood estimator is said to converge to the least false parameter value.

The maximum likelihood estimator for a particular model h_θ aims at minimising the Kullback–Leibler divergence above. Therefore, in order to evaluate its performance and compare it with the competing candidates, we will study the actually attained Kullback–Leibler divergence

$$\mathrm{KL}_1(h^\circ, h_{\widehat{\theta}_n}) = \int h^\circ(x)\log h^\circ(x)\,\mathrm{d}x - R(\widehat{\theta}_n), \tag{2.3}$$

which is a random variable. The first term is the same across all models, meaning that it is sufficient to study $R(\widehat{\theta}_n)$, which suggests that

$$Q_n = \mathrm{E}_{h^\circ} R(\widehat{\theta}_n) = \mathrm{E}_{h^\circ} \int h^\circ(x) \log h_{\widehat{\theta}_n}(x) \, \mathrm{d}x$$

is a reasonable measure for the success of each candidate. In turn, this motivates a model selection strategy by preferring the model that attains the largest value of Q_n. This model will also be expected to minimise (2.3) and can therefore be interpreted as best at what the maximum likelihood estimator is trying to achieve, i.e. to be close to the true density h° with respect to the expected Kullback–Leibler discrepancy.

In order to implement this in practice we need to calculate Q_n for each candidate, these depend on the true underlying density h°, which is unknown, meaning that Q_n must to be estimated from data. Since we expect $\ell_n(\theta)/n$ to be close to $R(\theta)$ by the law of large numbers, a natural start estimator for Q_n is

$$\widehat{Q}_n = \frac{1}{n} \ell_n(\widehat{\theta}_n). \tag{2.4}$$

This simple log-likelihood based estimator has a tendency to overshoot its target Q_n and a correction is therefore needed. In short, the bias correction of $\widehat{Q}_n$ can be shown to justify $\mathrm{AIC}(\theta) = 2n\{\widehat{Q}_n - \frac{1}{n}\dim(\theta)\} = 2(\ell_n(\widehat{\theta}_n) - p)$ as an (approximative and asymptotic) first order bias corrected estimator for Q; see also Claeskens & Hjort (2008, Ch. 2) for a more complete derivation and additional comments.

As an additional remark we note that the second term in the AIC formula, which is often referred to as a penalty term that penalises models for unnecessary high model complexity. Actually has a much deeper and more profound meaning as a bias correction with connection to the Kullback–Leibler divergence and maximum likelihood estimation. Without this interpretation, the precise structure of the penalty term becomes essentially arbitrary, since there is no real reason we should prefer the current form above any other similar constructions, e.g. $\frac{1}{2}p$, $4p$ or even $(\log n)p$.

3. The AIC_∞ for stationary time series models

The aim of this section is to explore the justification for using AIC-like criteria for time series processes and the main question is whether a similar – or any – rational justification for using the AIC can be carried over from the i.i.d. case to the time series framework. The AIC is already well established as a model selection strategy for time series processes, see among others Akaike (1976), Ogata (1980) and Hurvich & Tsai (1989). The general structure and underlying motivation of the present paper is quite different and we believe that our detailed derivation will provide new and valuable insights.

3.1. Maximum likelihood estimation in misspecification in time series models.

Let Y_t, $t \geq 1$ be a stationary Gaussian time series process with zero mean. The dependency structure, which

determines the entire model, is completely specified by the covariance function $C(h)$ for all lags $h = 0, 1, 2, \ldots$ and since we will only work with real-valued series, we have

$$C_f(h) = \mathrm{Cov}(Y_{t+h}, Y_t) = \int_{-\pi}^{\pi} \cos(\omega h) f(\omega)\, \mathrm{d}\omega, \qquad \text{for } h \geq 0, \tag{3.1}$$

where f is the so-called spectral density; for detailed introductions to time series modelling in the frequency domain see among others Brillinger (1975), Priestley (1981) or Dzhaparidze (1986). The 'true' spectral density, i.e. the spectral density of the underlying observed process, will be denoted by g, and f_θ, with $\theta \in \Theta \subset \mathbb{R}^p$ for finite p and a compact subspace Θ, represents a parametric model from a collection of candidates that do not necessarily include g.

Assumption 3.1. The spectral density $g(\omega)$ is Lipschitz-continuous and bounded away from zero and infinity. The spectral densities $f_\theta(\omega)$ are all bounded away from zero and infinity and are two times differentiable with respect to θ, with bounded derivatives that are continuous in both θ and ω.

Remark 3.2. The conditions listed in Assumption 3.1 are more or less equivalent to Assumptions 1.1, 1.2 and 1.3 in Dahlhaus & Wefelmeyer (1996). There is a difference in that in Assumption 1.2 it is assumed that the derivative of the spectral density of the candidate models is also bounded away from zero. This is presumably a clerical error since it is not needed in the proofs, making it an unnecessarily restrictive assumption, since it would only allow strictly monotonic spectral densities.

If the conditions of Assumption 3.1 are all satisfied it follows from Dahlhaus & Wefelmeyer (1996) that the maximum likelihood estimator converges,

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} \ell_n(f_\theta) \to_{P_g} \theta_0 = \arg\min_{\theta\in\Theta} d(g, f_\theta) = \arg\max_{\theta\in\Theta} R(\theta), \tag{3.2}$$

provided the least false parameter value θ_0 exists uniquely inside the compact Θ, where

$$d(g, f_\theta) = \mathrm{KL}_\infty(g, f_\theta) = -\frac{1}{4\pi}\int_{-\pi}^{\pi} (\log g(\omega) + 1)\, \mathrm{d}\omega - R(\theta) \tag{3.3}$$

is referred to as the asymptotic Kullback–Leibler divergence and where

$$R(\theta) = -\frac{1}{4\pi}\int_{-\pi}^{\pi} \left(\log f_\theta(\omega) + \frac{g(\omega)}{f_\theta(\omega)} \right) \mathrm{d}\omega \tag{3.4}$$

is the corresponding model specific part. The discrepancy measure (3.3) is easily motivated as the limit of the (scaled) Kullback–Leibler divergence

$$\begin{aligned} d_n(g, f_\theta) = \frac{1}{n}\,\mathrm{KL}_n(g, f_\theta) &= \frac{1}{n}\mathrm{E}_g\left\{\ell_n(g) - \ell_n(f_\theta)\right\} \\ &= -\frac{1}{2n}\bigg(\log|\Sigma_n(g)| + n - \log|\Sigma_n(f_\theta)| - \mathrm{tr}\{\Sigma_n(g)\Sigma_n(f_\theta)^{-1}\} \bigg), \end{aligned} \tag{3.5}$$

where $\Sigma_n(g)_{i,j} = C_g(|i-j|)$, as n approaches infinity; see e.g. Gray (2006) for details.

Following the above motivation, the general strategy is to prefer the model that maximises $\mathrm{E}_g\, R(\widehat{\theta}_n)$. This is equivalent to searching for the model that minimises the attained value of $\mathrm{E}_g\, d(g, f_{\widehat{\theta}_n})$,

i.e. the model that is best at what the maximum likelihood estimator is trying to achieve in the limit; to be closest to the truth with respect to the expected asymptotic Kullback–Leibler divergence.

Since the expected model specific part $\mathrm{E}_g\, R(\widehat{\theta}_n)$ depends on the true spectral density, which is unknown, it must be estimated from data. To see how, we introduce the periodogram

$$I_n(\omega) = \frac{1}{2\pi n}\left|\sum_{t\leq n} y_t \exp(-i\omega t)\right|^2,$$

which is a common nonparametric estimator for the spectral density. This now leads up to a canonical estimator for R by

$$\widetilde{R}_n(\theta) = -\frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log f_\theta(\omega) + \frac{I_n(\omega)}{f_\theta(\omega)}\right)\mathrm{d}\omega, \tag{3.6}$$

since $|\widetilde{R}_n(\theta) - R(\theta)| \to_{P_g} 0$, uniformly in θ; see the comments in Section 3.3. A natural starting point is therefore to use $\widetilde{R}_n(\widehat{\theta}_n)$ as an estimator for $\mathrm{E}_g\, R(\widehat{\theta}_n)$.

Furthermore, $\widetilde{R}_n$ is in close relation to the Whittle approximation introduced in Whittle (1953), which is defined by

$$\widetilde{\ell}_n(f_\theta) = -\frac{n}{2}\left\{\log 2\pi + \frac{1}{2\pi}\int_{-\pi}^{\pi}\log 2\pi f_\theta(\omega)\,\mathrm{d}\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_n(\omega)}{f_\theta(\omega)}\,\mathrm{d}\omega\right\}, \tag{3.7}$$

where I_n is the periodogram as defined above; for a general introduction and related large-sample properties, see Coursol & Dacunha-Castelle (1982) and Dzhaparidze (1986). Moreover, since we know from Dahlhaus & Wefelmeyer (1996) that the $\widehat{\theta}_n$ and the analogue Whittle estimator $\widetilde{\theta}_n = \arg\max_\theta \widetilde{\ell}_n(\theta)$ have the same large-sample behaviour, even in a misspecified modelling framework, it suggests that we will not commit a substantial error if we instead try to estimate $\mathrm{E}_g\, R(\widehat{\theta}_n)$ by

$$\widetilde{Q}_n = \widetilde{R}_n(\widetilde{\theta}_n) = \frac{1}{n}\widetilde{\ell}_n(\widetilde{\theta}_n) - \frac{1}{2}\log 2\pi. \tag{3.8}$$

In addition to changing the estimator, we will also view $\widetilde{Q}_n$ as an estimator for $Q = \mathrm{E}_g\, R(\widetilde{\theta}_n)$, which is convenient and will make the mathematics more straightforward and the necessary arguments more elegant. With this last change we have seemingly deviated form the original plan, since we have removed the maximum likelihood estimator – the original motivation and starting point – out of the discussion. By the asymptotic equivalence of the estimators these structural changes will not alter the validity of the original argument or claimed objective, however; see the Appendix in Section 9 for details.

3.2. **Derivation of an unbiased estimator from the Whittle approximation.** As in the simple case with i.i.d. observations, the Whittle based estimator $\widetilde{Q}_n$ in (3.8) has a tendency to overshoot its target $Q = \mathrm{E}_g\, R(\widetilde{\theta}_n)$ and a bias correction is therefore needed. In order to determine by how much, observe that the expected bias is given by $b = \mathrm{E}_g\,[\widetilde{Q}_n - Q] = \mathrm{E}_g\,[\widetilde{R}_n(\widetilde{\theta}_n) - R(\widetilde{\theta}_n)]$ and that we may further write

$$\widetilde{R}_n(\widetilde{\theta}_n) - R(\widetilde{\theta}_n) = (\widetilde{R}_n(\widetilde{\theta}_n) - \widetilde{R}_n(\theta_0) - [R(\widetilde{\theta}_n) - R(\theta_0)]) + \widetilde{R}_n(\theta_0) - R(\theta_0) = (\Delta_n - \delta_n) + \epsilon_n, \tag{3.9}$$

where $\Delta_n = \widetilde{R}_n(\widetilde{\theta}_n) - \widetilde{R}_n(\theta_0)$, $\delta_n = R(\widetilde{\theta}_n) - R(\theta_0)$ and $\epsilon_n = \widetilde{R}_n(\theta_0) - R(\theta_0)$. In order to obtain a proper correction, we will study the large-sample properties of $\Delta_n - \delta_n$ and ϵ_n, where the following two lemmas establishes the necessary results.

Lemma 3.3. *Let Δ_n and δ_n be as defined in (3.9). Then if the spectral densities g and f_θ satisfy the conditions of Assumption 3.1,*

$$\Delta_n - \delta_n = \frac{1}{n}\{W_n + o_p(1)\},$$

where $W_n \to_d W = U^{\mathrm{t}} J(g, f_{\theta_0})^{-1} U$, for $U \sim \mathrm{N}(0, K(g, f_{\theta_0}))$. Here

$$J(g, f_\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\nabla\Psi_\theta(\omega)\nabla\Psi_\theta(\omega)^{\mathrm{t}} g(\omega) + \nabla^2\Psi_\theta(\omega)(f_\theta(\omega) - g(\omega))\right]\frac{1}{f_\theta(\omega)}\,\mathrm{d}\omega \tag{3.10}$$

and

$$K(g, f_\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\nabla\Psi_\theta(\omega)\nabla\Psi_\theta(\omega)^{\mathrm{t}}\left[\frac{g(\omega)}{f_\theta(\omega)}\right]^2\mathrm{d}\omega, \tag{3.11}$$

with $\Psi_\theta(\omega) = \log f_\theta(\omega)$ and $\nabla\Psi_\theta(\omega)$ and $\nabla^2\Psi_\theta(\omega)$ being the vector and matrix of partial derivatives with respect to θ, respectively.

Proof. For any candidate model f_θ define

$$\widetilde{U}_n(\theta) = \frac{1}{\sqrt{n}}\nabla\widetilde{\ell}_n(\theta) = \sqrt{n}\,\nabla\widetilde{R}_n(\theta) \quad \text{and} \quad \widetilde{J}_n(\theta) = -\frac{1}{n}\nabla^2\widetilde{\ell}_n(\theta) = -\nabla^2\widetilde{R}_n(\theta),$$

where $\nabla\,\widetilde{R}_n(\theta)_i = \partial/\partial\theta_i\,\widetilde{R}_n(\theta)$ and $\nabla^2\,\widetilde{R}_n(\theta)_{i,j} = \partial^2/\partial\theta_i\partial\theta_j\,\widetilde{R}_n(\theta)$ for all $i, j = 1, \ldots, p$. Then by a standard two-term Taylor expansion of $\widetilde{R}_n$ around θ_0 it follows that

$$\begin{aligned}\widetilde{R}_n(\widetilde{\theta}_n) - \widetilde{R}_n(\theta_0) &= \widetilde{U}_n(\theta_0)^{\mathrm{t}}(\widetilde{\theta} - \theta_0) - \frac{1}{2}(\widetilde{\theta}_n - \theta_0)^{\mathrm{t}}\widetilde{J}_n(\bar{\theta}_n^{(1)})(\widetilde{\theta}_n - \theta_0)\\ &= \frac{1}{n}V_n^{\mathrm{t}}\widetilde{J}_n(\bar{\theta}_n^{(2)})V_n - \frac{1}{2n}V_n^{\mathrm{t}}\widetilde{J}_n(\bar{\theta}_n^{(1)})V_n,\end{aligned} \tag{3.12}$$

since $0 = \nabla\widetilde{\ell}_n(\widetilde{\theta})/n = \nabla\widetilde{R}_n(\theta_0) + \nabla^2\widetilde{R}_n(\bar{\theta}_n^{(2)})(\widetilde{\theta} - \theta_0)$, where $V_n = \sqrt{n}(\widetilde{\theta}_n - \theta_0)$ and $|\bar{\theta}_n^{(j)} - \theta_0| \le |\widetilde{\theta}_n - \theta_0|$ for $j = 1, 2$. Next, let $J(\theta) = J(g, f_\theta) = -\nabla^2 R(\theta)$, then by a similar Taylor expansion of R we obtain

$$R(\widetilde{\theta}_n) - R(\theta_0) = \nabla R(\theta_0)^{\mathrm{t}}(\widetilde{\theta}_n - \theta_0) - \frac{1}{2}(\widetilde{\theta}_n - \theta_0)^{\mathrm{t}}J(\bar{\theta}_n^{(3)})(\widetilde{\theta}_n - \theta_0) = \frac{1}{2n}V_n^{\mathrm{t}}J(\bar{\theta}_n^{(3)})V_n, \tag{3.13}$$

since $\nabla R(\theta_0) = 0$ from the definition of θ_0 and where $|\bar{\theta}_n^{(3)} - \theta_0| \le |\widetilde{\theta}_n - \theta_0|$. Now, by combining (3.12) and (3.13) we may express the difference as

$$\Delta_n - \delta_n = \widetilde{R}_n(\widetilde{\theta}_n) - \widetilde{R}_n(\theta_0) - [R(\widetilde{\theta}_n) - R(\theta_0)] = \frac{1}{n}V_n^{\mathrm{t}}\widetilde{J}_n(\bar{\theta}_n^{(2)})V_n - \frac{1}{2n}[V_n^{\mathrm{t}}\widetilde{J}_n(\bar{\theta}_n^{(1)})V_n - V_n^{\mathrm{t}}J(\bar{\theta}_n^{(3)})V_n].$$

It follows from Dahlhaus & Wefelmeyer (1996) that $V_n = \sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d J(g, f_{\theta_0})^{-1}U$, where $U \sim N_p(0, K(g, f_{\theta_0}))$. Then by arguments similar to those used in Section 3.3 below, it follows that $\widetilde{J}_n(f_{\bar{\theta}_n}) \to_{P_g} J(f_{\theta_0})$, provided $\bar{\theta}_n \to_{P_g} \theta_0$, and we now have the claimed results since

$$\Delta_n - \delta_n = \frac{1}{n}\{W_n + o_{P_g}(1)\}$$

where $W_n = V_n^{\mathrm{t}}\widetilde{J}_n(\bar{\theta}_n^{(2)})V_n \to_d U^{\mathrm{t}}J(g, f_{\theta_0})^{-1}U = W$. Note also that $\mathrm{E}_g\, W = \mathrm{tr}\{J(g, f_{\theta_0})^{-1}K(g, f_{\theta_0})\}$.

□

Lemma 3.4. *Let the spectral density g be continuous and bounded away from zero and infinity. Then for any positive, bounded and continuous function h that is symmetric on $[-\pi,\pi]$,*

$$\mathrm{E}_g \int_{-\pi}^{\pi} h(\omega) I_n(\omega)\,\mathrm{d}\omega - \int_{-\pi}^{\pi} h(\omega) g(\omega)\,\mathrm{d}\omega = -\frac{1}{n}\left[\frac{1}{8\pi^2}\langle g, h\rangle_{1/2} + o(1)\right],$$

where $\langle f_1, f_2\rangle_{1/2}$ is a type of inner product defined for pairs of all spectral densities f_1, f_2 given by the formula

$$\langle f_1, f_2\rangle_{1/2} = \sum_{j=-\infty}^{\infty} |j|\widehat{f}_{1\,j}\widehat{f}_{2\,j}, \text{ where } \widehat{f}_{i\,j} = \int_{-\pi}^{\pi} \exp\{-ij\omega\} f_i(\omega)\,\mathrm{d}\omega, \text{ for } i = 1, 2. \tag{3.14}$$

Proof. The proof follows more or less directly from the results in Coursol & Dacunha-Castelle (1982). In order to see this, we observe that

$$\begin{aligned}\mathrm{E}_g \int_{-\pi}^{\pi} h(\omega) I_n(\omega)\,\mathrm{d}\omega - \int_{-\pi}^{\pi} h(\omega) g(\omega)\,\mathrm{d}\omega &= \frac{1}{n}\left[\mathrm{E}_g\, y^{\mathrm{t}}\Sigma_n(h) y - \mathrm{tr}\{\Sigma_n(hg)\}\right] \\ &= \frac{1}{n}\left[\mathrm{tr}\{\Sigma_n(h)\Sigma_n(g)\} - \mathrm{tr}\{\Sigma_n(hg)\}\right] = -\frac{1}{n}\left[\frac{1}{8\pi^2}\langle g, h\rangle_{1/2} + o(1)\right],\end{aligned}$$

where the last equality is from Coursol & Dacunha-Castelle (1982, Proposition 2). □

Remark 3.5. In the definition of (3.14) used in Coursol & Dacunha-Castelle (1982) there is a small clerical error, where in the sum defining the inner product they use j instead of $|j|$.

By application of Lemma 3.3

$$\widetilde{R}_n(\widetilde{\theta}_n) - R(\widetilde{\theta}_n) = \Delta_n - \delta_n + \epsilon_n = \frac{1}{n} W_n + o_p(1/n) + \epsilon_n, \tag{3.15}$$

where $W_n \to_d W$ and $\mathrm{E}_g W = \mathrm{tr}\{J(g, f_{\theta_0})^{-1} K(g, f_{\theta_0})\}$, with J and K are as defined in (3.10). This is so far quite similar to the classical setup with i.i.d. observations, however, it follows from Lemma 3.4 that $\mathrm{E}_g\,\epsilon_n = O(1/n)$ for the time series processes and not the desired $o(1/n)$, or $\epsilon_n = o_{P_g}(1/n)$, which would have justified neglecting this term at the claimed level of precision. Fortunately, Lemma 3.4 provides the result needed to obtain an estimator at the right level of precision, since

$$\mathrm{E}_g\,\epsilon_n = \frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{g(\omega)}{f_{\theta_0}(\omega)}\,\mathrm{d}\omega - \mathrm{E}_g\,\frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{I_n(\omega)}{f_{\theta_0}(\omega)}\,\mathrm{d}\omega = \frac{1}{n}\left[\frac{1}{8\pi^2}\langle g, 1/f_{\theta_0}\rangle_{1/2} + o(1)\right], \tag{3.16}$$

where $\langle\cdot,\cdot\rangle_{1/2}$ is the inner product defined in (3.14). Then

$$\mathrm{E}_g\,[\widetilde{Q}_n - Q] \approx \frac{p^* + q^*}{n}, \quad \text{where } q^* = \frac{1}{8\pi^2}\langle g, 1/f_{\theta_0}\rangle_{1/2} \quad \text{and} \quad p^* = \mathrm{tr}\{J(g, f_{\theta_0})^{-1} K(g, f_{\theta_0})\},$$

which means that $\widetilde{Q}_n - (p^* + q^*)/n$ is an approximative asymptotic first order unbiased estimator for the target Q. The reason it is only 'approximately' unbiased, is that the expectation of a $o_{P_g}(1/n)$ term is not necessarily of the desired $o(1/n)$ order. It is common practice to ignore the $o_{P_g}(1/n)$ terms and this is also the best we can do without introducing more advanced tools or additional conditions.

This establishes the necessary connection and justifies the criterion

$$\mathrm{AIC}^*_\infty = 2\{\widetilde{\ell}_n(f_{\widetilde{\theta}_n}) - \widetilde{p}^* - \widetilde{q}^*\}, \tag{3.17}$$

where

$$\widetilde{p}^{\,*} = \mathrm{tr}\{J(I_n, f_{\widetilde{\theta}_n})^{-1} K(I_n/\sqrt{2}, f_{\widetilde{\theta}_n})\} \quad \text{and} \quad \widetilde{q}^{\,*} = \frac{1}{8\pi^2}\langle I_n, 1/f_{\widetilde{\theta}_n}\rangle_{1/2}, \tag{3.18}$$

see also Section 3.3 for comments regarding estimation. This is the model robust version where p^* and q^* are estimated from data, the analogue model robust version of the AIC for i.i.d. observations is commonly known as Takeuchi's Information Criterion (TIC) from Takeuchi (1976).

If a particular candidate model f_θ happens to span the true model, it is easy to verify that $J(g, f_{\theta_0}) = K(g, f_{\theta_0})$, meaning that $p^* = \mathrm{tr}(I_p) = p$. This also implies $q^* = q = (2\sqrt{2}\pi^2)^{-2}\langle f_\theta, 1/f_{\theta_0}\rangle_{1/2}$, which gives the alternative formulation

$$\mathrm{AIC}_\infty(f_\theta) = 2\{\widetilde{\ell}_n(f_{\widetilde{\theta}_n}) - p - \widetilde{q}\,\}, \tag{3.19}$$

where $\widetilde{q} = (2\sqrt{2}\pi)^{-2}\langle f_{\widetilde{\theta}_n}, 1/f_{\widetilde{\theta}_n}\rangle_{1/2}$. Taking a more extreme standpoint, where we assume that all candidate models span, or include, the true model, results in $(2\sqrt{2}\pi)^2 q^*(\theta) = \langle g, 1/f_{\theta_0}\rangle_{1/2} = \langle g, 1/g\rangle_{1/2}$ making it constant across models and motivating $\mathrm{AIC}_\infty(f_\theta) = 2\{\widetilde{\ell}_n(f_{\widetilde{\theta}_n}) - p\} = \widetilde{\mathrm{AIC}}_n(f_\theta)$, i.e. the Whittle approximation to the general AIC formula alluded to in Section 1.

The need for the additional correction term obtained in (3.17) and (3.19) was already observed in Findley (1985) for ARMA models, see also Hurvich & Tsai (1991) for some additional comments. The underlying derivation and justification are quite different, however, and in the end the idea of using the Whittle approximation was rejected as generally unmotivated. In our view, the derivation presented here gives the necessary justification needed to motivate the Whittle based AIC_∞ as a rational and coherent model selection strategy. We also note that similar observations are made in Ioannidis (2011) regarding the bias of the general AIC formula. This is in the framework of autoregressive processes and least squares estimation, and there is essentially no overlap with the work present here.

3.3. **Estimation.** The estimation of the correction terms needed for the model robust AIC in (3.17), depends on the ability to consistently estimate integrals involving the true underlying spectral density g. These are easiest estimated following Taniguchi (1980), where it is shown that for a continuous and symmetric function h on $[-\pi, \pi]$ it follows that

$$\int_{-\pi}^{\pi} h(\omega) I_n(\omega)\,\mathrm{d}\omega \xrightarrow{P_g} \int_{-\pi}^{\pi} h(\omega) g(\omega)\,\mathrm{d}\omega \quad \text{and} \quad \frac{1}{2}\int_{-\pi}^{\pi} h(\omega) I_n(\omega)^2\,\mathrm{d}\omega \xrightarrow{P_g} \int_{-\pi}^{\pi} h(\omega) g(\omega)^2\,\mathrm{d}\omega$$

provided the following short memory assumption $\sum_t |t||C_g(t)| < \infty$ is satisfied. As an independent remark, we note that this short memory condition will follow if the spectral density g is continuous and bounded below infinity, see Carslaw (1921, p. 249).

The results are easily strengthened to include functions $h_{\widehat{\gamma}_n}$ that depend on a sequence of estimators $\widehat{\gamma}_n$, provided the sequence converges. This follows since the integrals involving I_n are bounded in P_g-probability, meaning that it is sufficient to observe that

$$\left|\int_{-\pi}^{\pi} h_{\widehat{\gamma}_n}(\omega) I_n(\omega)\,\mathrm{d}\omega - \int_{-\pi}^{\pi} h_{\gamma_0}(\omega) I_n(\omega)\,\mathrm{d}\omega\right| \le \sup_{0\le\omega\le\pi} |h_{\widehat{\gamma}_n}(\omega) - h_{\gamma_0}(\omega)| \left|\int_{-\pi}^{\pi} I_n(\omega)\,\mathrm{d}\omega\right|, \tag{3.20}$$

which will become small provided h_γ is continuous in a neighbourhood of γ_0 and $\widehat{\gamma}_n$ converges in true P_g-probability to γ_0. Also, according to Taniguchi (1979, p. 580) and Taniguchi (1980, p. 74) no other nonparametric estimators can beat the periodogram in terms of variance.

Next, in order to calculate $\widetilde{q}$ we have to estimate the inner products defined in (3.14). To simplify, we observe that for symmetric functions

$$\langle f_1, f_2 \rangle_{1/2} = \sum_{j=-\infty}^{\infty} |j| \widehat{f}_{1\,j} \widehat{f}_{2\,j}, = 2 \sum_{h=1}^{\infty} |h| C_{f_1}(h) C_{f_2}(h),$$

which further implies that

$$\langle f, I_n \rangle_{1/2} = 2 \sum_{h=1}^{n-1} |h| C_f(h) \widehat{C}(h),$$

where $\widehat{C}(\cdot)$ is the classical nonparametric estimate for the covariance function; see e.g. Brillinger (1975). In addition, if f has bounded derivatives up to order k and continuous derivatives up to a order of $k-1$, it follows under additional mild regularity conditions that the j-th Fourier coefficients $\widehat{f}_j$ of the function f will be smaller in absolute value than c/j^{k+1}, where c is a constant independent of j, see Carslaw (1921, Chapter VIII). This means that the value of the inner product are in most cases more or less completely determined by the first few terms in the sum, which provides an strategy for making consistent estimates. In addition,

$$\langle f_1, f_2 \rangle_{1/2} \leq 2 \left| \sum_{h=1}^{\infty} |h| C_{f_1}(h) C_{f_2}(h) \right| \leq 2 \left| \sum_{h=1}^{\infty} |h| C_{f_1}(h)^2 \right|^{1/2} \left| \sum_{h=1}^{\infty} |h| C_{f_2}(h)^2 \right|^{1/2} < \infty,$$

provided the required short memory assumption is satisfied.

This is an important observation, since it shows that the correction term q^* can become large, but will never explode and will therefore in most cases be straightforward to estimate; see Grønneberg & Hjort (2014) for a case where a similar problem is present and is seen to require further exploration.

Finally, in order to apply AIC^*_∞ in practice we need to estimate p^* for each candidate model, the canonical estimator is

$$\widetilde{p}^* = \text{tr}\{J(I_n, f_{\widetilde{\theta}_n})^{-1} K(I_n/\sqrt{2}, f_{\widetilde{\theta}_n})\},$$

which is consistent by the properties of $I_n(\omega)$ and $I_n(\omega)^2/2$. Unfortunately, there is a tendency to underestimate integrals in small samples with $I_n(\omega)^2/2$ as an estimator for $g(\omega)^2$. This is most likely related to an artifact of fitting flexible models, such as the autoregressive, since the estimated parametric spectral density can become very close to the nonparametric periodogram estimate (if the sample size is low compared to the order of the autoregressive process), which means that $J(I_n, f_{\widetilde{\theta}_n}) \approx 2K(I_n/\sqrt{2}, f_{\widetilde{\theta}_n})$ and $\widetilde{p}^* \approx \text{tr}(I_p)/2 = p/2$, and the result is that complex models will be preferred more often that they should.

This should be taken seriously and a finite-sample bias correction may be obtainable from the work in Brillinger (1975, Ch. 4 & 5); a proper solution will require additional work which do not intend to solve here. We will instead briefly present some alternatives. If the models are nested, we may use

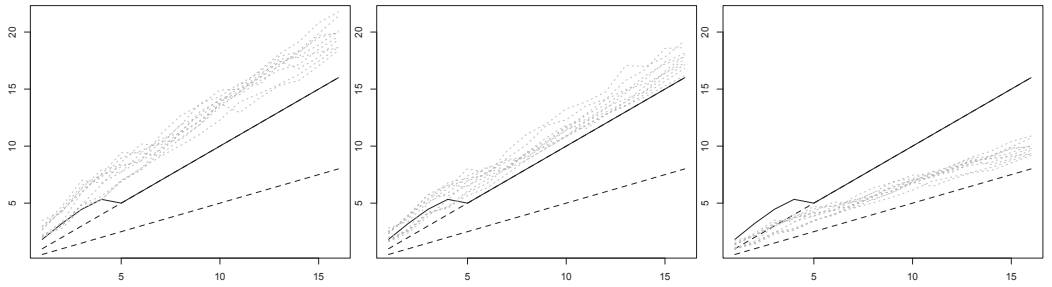

FIGURE 3.1. Estimated p^* in 10 different realizations for size $n = 32$, the candidate models are autoregressive up to an order of 16 and the true model is of order 4 with parameters $(1, 0.4, -0.5, 0.4, -0.5)$. The black solid line indicates the true (average) p^*, note that since the models are nested $p^* = p$ for $p \geq 5$, the two broken lines are p and $p/2$. From left to right we have used $I_n(\omega)^2$, $I_{n,\text{smooth}}(\omega)^2$ and $I_n(\omega)^2/2$ to estimate the unknown $g(\omega)^2$.

the largest model to estimate $g(\omega)$. Moreover, we may take a more semiparametric approach where we combine a nonparametric and parametric estimator with $\widehat{g}(\omega) = I_n(\omega)$ and $\widehat{g}(\omega)^2 = I_n(\omega) f_{\widehat{\theta}_n}(\omega)$. Another idea that seems to work reasonable well in practice, is to use a smooth or tapered periodogram estimator for the true spectral density, i.e. let $\widehat{g}(\omega)^2 = \widehat{f}_{\text{smoothed}}(\omega)^2$; see for example Brillinger (1975) for suggestions.

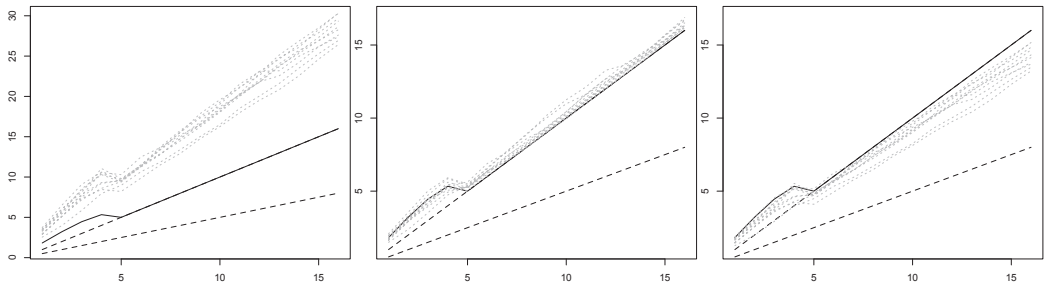

FIGURE 3.2. The same setup as in Figure 3.1, but with $n = 256$.

4. The AIC_n for stationary time series models

The main reason we ended up with the Whittle approximation based AIC_∞ formula (3.17) was the focus on maximum likelihood estimation in a misspecified modelling framework, which in turn resulted in the introduction of the asymptotic Kullback–Leibler divergence (3.3) as a natural measure of discrepancy. In this section we show how to motivate the more standard $\text{AIC}_n = 2(\ell_{n,\max} - p)$ formulation and we will again take inspiration from the simpler case of parametric model fitting for i.i.d. observations.

Let $X_1, \ldots, X_n$ be i.i.d. random variables from the model with density h° and let again h_θ represent a potential candidate. By following the so-called extended likelihood principle of Akaike (1973) we should, in set of competing models, prefer the one that maximises

$$\mathrm{E}_{h^\circ} \log h_{\widehat{\theta}_n}(X) = \mathrm{E}_{h^\circ} \int h^\circ(x) \log h_{\widehat{\theta}_n}(x)\, \mathrm{d}x, \tag{4.1}$$

where $\widehat{\theta}_n$ is the maximum likelihood estimator. Note that the expectation is with respect to both the new random variable X and the estimator $\widehat{\theta}_n$. In order to implement the extended likelihood principle, the scores defined by (4.1) must be estimated from data. The canonical starting point is to use the maximised log-likelihood function, which results in an unbiased estimator, however, the AIC formula then emerges from the derivations of an appropriate bias correction.

Note that (4.1) is actually the expected model specific part of the attained Kullback–Leibler divergence of Section 2. In the original papers by Akaike, the connection to estimation in a misspecified modelling framework is not made explicitly. The principle is instead commonly interpreted in relation to a type of predictive performance and justified via the connection to Kullback–Leibler divergence, which is further connected to information theory and entropy, see Akaike (1973, 1974) for more details. The reasoning is unfortunately somewhat vague, making the general idea harder to grasp. A more intuitive, but heuristic, account is given in Akaike (1976), where the autoregressive models are also briefly discussed.

4.1. **Derivation of an unbiased estimator from the full log-likelihood.** In order to apply this extended likelihood principle in time series and for dependent observations, we need to interpret (4.1) in a multivariate framework, where (4.1) now becomes

$$\mathrm{E}_g\, \ell_n(f_{\widehat{\theta}_n}) = -\tfrac{1}{2}(n \log(2\pi) + \log|\Sigma_n(f_\theta)| + \mathrm{tr}\{\Sigma_n(g)\Sigma_n(f_\theta)^{-1}\}), \tag{4.2}$$

where $\Sigma_n(f_\theta)_{i,j} = C_{f_\theta}(|i-j|)$ and $\widehat{\theta}_n = \arg\max_\theta \ell_n(f_\theta)$. It is easily seen that the model that maximises $\mathrm{E}_g\, \ell_n(f_{\widehat{\theta}_n})$ is the same that maximises $Q_n = \mathrm{E}_g\, R_n(\widehat{\theta}_n)$, where $nR_n(f_\theta) = -\frac{1}{2}(\log|\Sigma_n(f_\theta)| + \mathrm{tr}\{\Sigma_n(g)\Sigma_n(f_\theta)^{-1}\})$ is the model specific part of

$$d_n(g, f_\theta) = \mathrm{KL}_n(g, f_\theta) = \frac{1}{n}\mathrm{E}_g\,[\ell_n(g) - \ell_n(f_\theta)] = \frac{1}{n}\mathrm{E}_g\, \ell_n(g) - \frac{1}{2}\log 2\pi - R_n(f_\theta). \tag{4.3}$$

This means that applying the extended likelihood principle is equivalent to finding the model that minimises the expected attained Kullback–Leibler divergence $\mathrm{E}_g\, \mathrm{KL}_n(g, f_{\widehat{\theta}_n})$, where the expectation is with respect to the maximum likelihood estimator $\widehat{\theta}_n$.

Remark 4.1. From nearness of the Whittle approximation to the full Gaussian log-likelihood, it follows that both $d_n(g, f_\theta)$ and $R_n(f_\theta)$ converge uniformly in θ to the asymptotic Kullback–Leibler divergence d and its model specific part R of Section 3.2, see e.g. Dahlhaus & Wefelmeyer (1996).

A canonical starting point for estimating Q_n is given by the scaled log-likelihood function

$$\widehat{Q}_n = \widehat{R}_n(\widehat{\theta}_n) = \frac{1}{n}\ell_n(f_{\widehat{\theta}_n}) - \frac{1}{2}\log 2\pi, \tag{4.4}$$

since by Lemma A.5 in Dahlhaus (1996) it follows that $|\widehat{R}_n(\theta) - R_n(\theta)| \to_{P_g} 0$ uniformly in θ, which means that we expect $\widehat{Q}_n = \widehat{R}_n(\widehat{\theta}_n) \approx Q_n$. This simple likelihood based estimator turns out to be a little too naive and has a tendency to overshoot its target Q_n. To determine by how much and to give a proper bias correction, we will use a standard two-term Taylor expansion to derive a asymptotic first order unbiased estimator. To simplify this argument we note that $\mathrm{E}_g\,[\widehat{Q}_n - Q_n] = \mathrm{E}_g\,[\widehat{R}_n(\widehat{\theta}_n) - R_n(\theta_n)]$ and that

$$\widehat{R}_n(\widehat{\theta}_n) - R_n(\widehat{\theta}_n) = (\Delta_n - \delta_n) + \epsilon_n, \tag{4.5}$$

where $\Delta_n = \widehat{R}_n(\widehat{\theta}_n) - \widehat{R}_n(\theta_0)$, $\delta_n = R_n(\widehat{\theta}_n) - R_n(\theta_0)$ and $\epsilon_n = \widehat{R}_n(\theta_0) - R_n(\theta_0)$, where it is easily seen from the above that $\mathrm{E}_g\,\epsilon_n = 0$.

Lemma 4.2. *Let Δ_n and δ_n be as defined in (4.5), then if the spectral densities g and f_θ satisfy the conditions of Assumption 3.1*

$$\Delta_n - \delta_n = \frac{1}{n}\{W_n + o_{P_g}(1)\},$$

where $W_n \to_d W = U^{\mathrm{t}} J(g, f_{\theta_0})^{-1} U$, for $U \sim \mathrm{N}(0, K(g, f_{\theta_0}))$, with J and K are as defined in (3.10).

Proof. For any candidate model f_θ let

$$\widehat{U}_n(\theta) = \sqrt{n}\nabla\widehat{R}_n(\theta) = \frac{1}{\sqrt{n}}\nabla\ell_n(f_\theta) \quad \text{and} \quad \widehat{J}_n(\theta) = -\nabla^2\widehat{R}_n(\theta) = -\frac{1}{n}\nabla^2\ell_n(f_\theta),$$

where $\nabla\,\widehat{R}_n(\theta)_i = \partial/\partial\theta_i\,\widehat{R}_n(\theta)$ and $\nabla^2\,\widehat{R}_n(\theta)_{i,j} = \partial^2/\partial\theta_i\partial\theta_j\,\widehat{R}_n(\theta)$ for all $i, j = 1, \ldots, p$. Then by a two-term Taylor expansion (used twice) we have that

$$\begin{aligned}
\Delta_n = \widehat{R}_n(\widehat{\theta}_n) - \widehat{R}_n(\theta_0) &= (\widehat{\theta}_n - \theta_0)^{\mathrm{t}}\nabla\widehat{R}_n(\theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}}[\nabla^2\widehat{R}_n(\bar{\theta}_n^{(1)})](\widehat{\theta}_n - \theta_0) \\
&= \frac{1}{n}V_n^{\mathrm{t}}U_n(\theta_0) - \frac{1}{2n}V_n^{\mathrm{t}}\widehat{J}_n(\bar{\theta}_n^{(1)})V_n \\
&= \frac{1}{n}V_n^{\mathrm{t}}\widehat{J}_n(\bar{\theta}_n^{(2)})V_n - \frac{1}{2n}V_n^{\mathrm{t}}\widehat{J}_n(\bar{\theta}_n^{(1)})V_n,
\end{aligned}$$

where $V_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)$ and $|\bar{\theta}_n^{(i)} - \theta_0| \le |\widehat{\theta}_n - \theta_0|$ for $i = 1, 2$.

Define $U_n(\theta) = \nabla R_n(\theta)/\sqrt{n}$ and $J_n(\theta) = -\nabla^2 R_n(\theta)$, then by a second two-term Taylor expansion, we obtain

$$\delta_n = R_n(\widehat{\theta}_n) - R_n(\theta_0) = \frac{1}{\sqrt{n}}V_n^{\mathrm{t}}U_n(\theta_0) - \frac{1}{2n}V_n^{\mathrm{t}}J_n(\bar{\theta}_n^{(3)})V_n, \tag{4.6}$$

where $|\bar{\theta}_n^{(3)} - \theta_0| \le |\widehat{\theta}_n - \theta_0|$. Summarising the above, it follows that

$$\Delta_n - \delta_n = \frac{1}{n}V_n^{\mathrm{t}}\widehat{J}_n(\bar{\theta}_n^{(2)})V_n - \frac{1}{\sqrt{n}}V_n^{\mathrm{t}}U_n(\theta_0) - \frac{1}{2n}[V_n^{\mathrm{t}}\widehat{J}_n(\bar{\theta}_n^{(1)})V_n + V_n^{\mathrm{t}}J_n(\bar{\theta}_n^{(3)})V_n]. \tag{4.7}$$

Since θ_0 is not the minimiser of R_n we have that $U_n(\theta_0) \neq 0$. Fortunately, it follows from Lemma A.6 in Dahlhaus & Wefelmeyer (1996) that $U_n(\theta_0) = o_{P_g}(n^{-1/2})$, which is exactly the order needed to neglect (at the claimed level of precision) the second term in (4.7).

Furthermore, it follows from Theorem 3.3 in Dahlhaus & Wefelmeyer (1996) that $V_n \to_d V = J(g, f_{\theta_0})^{-1}U$, where $U \sim N(0, K(g, f_{\theta_0}))$, and that $\widehat{J}_n(\bar{\theta}_n^{(i)}) = -\nabla^2 \widehat{R}_n(\bar{\theta}_n^{(i)}) \to_{P_g} J(g, f_{\theta_0})$, for $i = 1, 2, 3$, where $J(g, f_{\theta_0})$ and $K(g, f_{\theta_0})$ are as defined in (3.10). This provides the tools needed to establish the claimed result

$$\Delta_n - \delta_n = \frac{1}{n}\{W_n + o_{P_g}(1)\}, \tag{4.8}$$

where $W_n = V_n^{\mathrm{t}} \widehat{J}_n(\bar{\theta}_n^{(3)}) V_n \to_d U^{\mathrm{t}} J(g, f_{\theta_0}) U = W$ and $\mathrm{E}_g\, W = \mathrm{tr}\{J(g, f_{\theta_0})^{-1} K(g, f_{\theta_0})\}$. □

Remark 4.3. There is a small error in the second part of the Lemma A.6 in Dahlhaus & Wefelmeyer (1996), the big O notation used needs to be a small o for the Lemma to work as they intended.

By application of Lemma 4.2 it now follows that

$$\mathrm{E}_g\,[\widehat{Q}_n - Q_n] = \mathrm{E}_g[\Delta_n - \delta_n + \epsilon_n] \approx \frac{1}{n}\mathrm{E}_g W = \frac{p^*}{n}, \quad \text{where } p^* = \mathrm{tr}\{J(g, f_{\theta_0})^{-1} K(g, f_{\theta_0})\}. \tag{4.9}$$

which furthermore establishes $\widehat{Q}_n - p^*/n$ as an approximative asymptotic first order unbiased estimator for Q. This provides the rational motivation needed to establish

$$\mathrm{AIC}_n^* = 2(\ell_n(\widehat{\theta}_n) - \widehat{p}^*), \quad \text{where } \widehat{p}^* = \mathrm{tr}\{J(I_n, f_{\widehat{\theta}_n})^{-1} K(I_n/\sqrt{2}, f_{\widehat{\theta}_n})\}, \tag{4.10}$$

see Section 3.3 for discussion regarding estimation. Note that if the model is correctly specified, it follows that $p^* = p$ and the above formula simplifies to $\mathrm{AIC}_n = 2(\ell_n(\widehat{\theta}_n) - p)$.

Remark 4.4. It is possible to construct consistent estimates for J and K that are based more directly on the full Gaussian log-likelihood. In order to see this, let $[\nabla \Sigma_n(f_\theta)^{-1}]_i$, for $i = 1, \ldots, p$, be the i-th component of the derivative of the inverse covariance matrix, then elements of $K_n(\theta) = n\,\mathrm{Var}_{P_g} \nabla R_n(\theta)$ is then given by

$$K_n(g, f_\theta)_{i,j} = \frac{2}{n}\,\mathrm{tr}\{[\nabla \Sigma_n(f_\theta)^{-1}]_i \Sigma_n(g) [\nabla \Sigma_n(f_\theta)^{-1}]_j \Sigma_n(g)\}, \quad \text{for } i, j = 1, \ldots, p,$$

and by Lemma A.5 in Dahlhaus (1996) and Taniguchi (1980) it is easily seen that $K_n(I_n/\sqrt{2}, f_{\widehat{\theta}_n}) \to_{P_g} K(g, f_{\theta_0})$, see also the discussion in Section 3.3. Note that a similar argument can be used to show that an estimate for $J(g, f_{\theta_0})$ can be obtained from the second derivative of the log-likelihood function evaluated in the maximum likelihood estimator.

4.2. **Motivating AIC_n from maximum likelihood estimation.** The justification for using the AIC_n as a model selection strategy depends largely on how rational the Kullback–Leibler divergence is as a measure of discrepancy. In order to establish a coherent argument, the Kullback–Leibler divergence is, as already commented on, often justified by connecting it to entropy and information theory with references to Shannon (1948) and Kullback (1959), however, we believe the most direct and elegant motivation follows from the more direct connection to maximum likelihood estimation in a misspecified modelling framework.

In Section 3, the maximum likelihood estimator $\widehat{\theta}_n$ was shown to converge to the least false parameter value $\theta_0 = \arg\min_\theta d(g, f_\theta)$, i.e. the minimiser of the asymptotic Kullback–Leibler divergence (3.3). Note that we can by similar arguments claim that $\widehat{\theta}_n$ aims at $\theta_{0,n} = \arg\min_\theta d_n(g, f_\theta)$, in the sense that $\|\widehat{\theta}_n - \theta_{0,n}\| \to_{P_g} 0$, which means that the maximum likelihood estimator is instead close to the minimiser of the Kullback–Leibler divergence (4.3). Moreover, under the additional assumption that the sequence $\{\theta_{0,n}\}_{n\geq 1}$ exists, a corresponding modified version of Lemma 4.2 can be shown to hold, which in turn motivates the criterion AIC_n^* in (4.10). As a final remark, the limit distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_{0,n})$ is the same as $\sqrt{n}(\widehat{\theta}_n - \theta_0)$, which can be seen to follow from

$$\sqrt{n}\nabla\ell_n(f_{\theta_{0,n}}) = \sqrt{n}\nabla\ell_n(f_{\theta_0}) + \sqrt{n}(\nabla\ell_n(f_{\theta_{0,n}}) - \nabla\ell_n(f_{\theta_0})),$$

since $\theta_{0,n} \to \theta_0$ implies that $\sqrt{n}(\nabla\ell_n(f_{\theta_{0,n}}) - \nabla\ell_n(f_{\theta_0})) \to_{P_g} 0$ by the results of Lemma A.5 and A.6 in Dahlhaus & Wefelmeyer (1996).

5. Two additional variations of the AIC for the time series processes

In the discussions above we have presented what we see as the most natural strategies for deriving AIC-like criteria for the stationary time series processes. In this section we will investigate two additional cases. Here we will use the maximised log-likelihood $\ell_{n,\max}$ as an estimator for the expected attained asymptotic Kullback–Leibler divergence (3.3). We will also see what happens if the Whittle log-likelihood is used as an approximation for the expected model specific part of (4.3). This results in additional AIC-like criteria with even more correction terms, which emphasises the idea that some strategies may be seen as more natural than others.

Consider the case where the scaled log-likelihood is used to estimate the expected model specific part of the asymptotic Kullback–Leibler divergence (3.3), i.e. we wish to use $\widehat{Q}_n = \widehat{R}_n(\widehat{\theta}_n) = \ell_n(f_{\widehat{\theta}_n})/n - (\log 2\pi)/2$ to estimate $Q = \mathrm{E}_g\, R(\widehat{\theta}_n)$, where $R(\theta) = \lim_{n\to\infty} \mathrm{E}_g\, \widetilde{\ell}_n(f_\theta)/n - (\log 2\pi)/2$. This seems reasonable, since the asymptotic Kullback–Leibler divergence is a natural choice of discrepancy from the connection with the least false parameter value, as the target for the maximum likelihood estimator. In addition, there may also be good reasons to prefer the full log-likelihood and maximum likelihood estimation. Not surprisingly, $\widehat{Q}_n$ is also a biased estimator for Q. In order to give a proper correction we have to make some small changes to the arguments used to prove Lemma 3.3 and 4.2.

Let $\widehat{U}_n(\theta_0) = \sqrt{n}\nabla\widehat{R}_n(\theta_0) = \nabla\ell_n(\theta_0)/\sqrt{n}$ and $\widehat{J}_n(\theta_0) = -\nabla^2\widehat{R}_n(\theta_0) = -\nabla^2\ell_n(\theta_0)/n$. Then by arguments similar to those already presented, a two-term Taylor expansion reveals that

$$\widehat{R}_n(\widehat{\theta}_n) - \widehat{R}_n(\theta_0) \doteq \frac{1}{\sqrt{n}}\widehat{U}_n(\theta_0)^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}}\widehat{J}_n(\theta_0)(\widehat{\theta}_n - \theta_0) \quad \text{and}$$

$$R(\widehat{\theta}_n) - R(\theta_0) \doteq U(\theta_0)^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}}J(\theta_0)(\widehat{\theta}_n - \theta_0),$$

where $U(\theta_0) = \nabla R(\theta_0) = 0$ and $J(\theta_0) = -\nabla^2 R(\theta_0)$. The only part that really differs and will therefore need extra care, is the calculation of the expectation $\mathrm{E}_g\, \epsilon_n$, which by Lemma 3.4 is easily seen to satisfy

$$\begin{aligned} n\mathrm{E}_g\, \epsilon_n &= n\mathrm{E}_g\, [\widehat{R}_n(\theta_0) - R(\theta_0)] \\ &= \mathrm{E}_g\, [\ell_n(f_{\theta_0}) - \widetilde{\ell}_n(f_{\theta_0})] + n\mathrm{E}_g\, [\widetilde{R}_n(\theta_0) - R(\theta_0)] \\ &= -\frac{1}{8\pi^2}\left[\frac{1}{2}\langle \log f_{\theta_0}, \log f_{\theta_0}\rangle_{1/2} + \langle \log f_{\theta_0}, g/f_{\theta_0}\rangle_{1/2} + \langle g, 1/f_{\theta_0}\rangle_{1/2}\right] + q^* + o(1) \\ &= -\frac{1}{8\pi^2}\left[\frac{1}{2}\langle \log f_{\theta_0}, \log f_{\theta_0}\rangle_{1/2} + \langle \log f_{\theta_0}, g/f_{\theta_0}\rangle_{1/2}\right] + o(1) \\ &= -r^* + o(1), \end{aligned} \tag{5.1}$$

where the brackets indicate the inner product defined in (3.14) and q^* is as define in (3.17). The result is a new criterion

$$\widehat{\mathrm{AIC}}^*_\infty = 2(\ell_n(\widehat{\theta}_n) - \widehat{p}^* + \widehat{r}^*) \quad \text{and} \quad \widehat{\mathrm{AIC}}_\infty = 2(\ell_n(\widehat{\theta}_n) - p + \widehat{r}\,), \tag{5.2}$$

where p^* is as defined in (3.17) and the values of p^* and r^* are estimated in accordance with the previous discussion.

The second approach is to use the Whittle approximation as it was 'intended', as an approximation for the full Gaussian log-likelihood, where $\widetilde{Q}_n = \widetilde{R}_n(\widetilde{\theta}_n)$, as defined in (3.6), is used to estimate $Q_n = \mathrm{E}_g\, R_n(\widehat{\theta}_n)$. It is essentially easy to motivate this case, since there might be situations where the standard Kullback–Leibler divergence is preferred, but it is impractical, or not even possible, to do the calculations required.

Again, a few changes are needed to the original arguments (the proofs of Lemma 3.3 and 4.2) and if $\widetilde{U}_n(\theta_0) = \sqrt{n}\nabla \widetilde{R}_n(\theta_0) = \nabla \ell_n(\theta_0)/\sqrt{n}$ and $\widetilde{J}_n(\theta_0) = -\nabla^2 \widetilde{R}_n(\theta_0) = -\nabla^2 \ell_n(\theta_0)/n$, a two-term Taylor expansion shows that

$$\begin{aligned} &\widetilde{R}_n(\widetilde{\theta}_n) - \widetilde{R}_n(\theta_0) \doteq \frac{1}{\sqrt{n}}\widetilde{U}_n(\theta_0)^{\mathrm{t}}(\widetilde{\theta}_n - \theta_0) - \frac{1}{2}(\widetilde{\theta}_n - \theta_0)^{\mathrm{t}}\widetilde{J}_n(\theta_0)(\widetilde{\theta}_n - \theta_0) \quad \text{and} \\ &R_n(\widehat{\theta}_n) - R_n(\theta_0) \doteq U_n(\theta_0)^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}} J_n(\theta_0)(\widehat{\theta}_n - \theta_0). \end{aligned}$$

Then, since we already know that $U_n(\theta_0) = o_{P_g}(1/\sqrt{n})$, the only part that needs additional care is the calculation of the expectation

$$\begin{aligned} n\mathrm{E}_g\, \epsilon_n &= n\mathrm{E}_g[\widetilde{Q}_n - \widehat{Q}_n] + n\mathrm{E}_g[\widehat{R}_n(\theta_0) - R_n(\theta_0)] \\ &= \mathrm{E}_g[\widetilde{\ell}_n(f_{\theta_{0\,n}}) - \ell_n(f_{\theta_{0\,n}})] \\ &= \frac{1}{8\pi^2}\left[\frac{1}{2}\langle \log f_{\theta_0}, \log f_{\theta_0}\rangle_{1/2} + \langle g, 1/f_{\theta_0}\rangle_{1/2} + \langle \log f_{\theta_0}, g/f_{\theta_0}\rangle_{1/2} + o(1)\right], \end{aligned}$$

see Lemma 3.4. The new corrected unbiased estimator needs all the previous correction terms and the result is a Whittle approximated AIC^*_n formula given by

$$\widetilde{\mathrm{AIC}}^*_n = 2(\widetilde{\ell}_n(\widetilde{\theta}_n) - \widetilde{p}^* - \widetilde{q}^* - \widetilde{r}^*) \quad \text{and} \quad \widetilde{\mathrm{AIC}}_n = 2(\widetilde{\ell}_n(\widetilde{\theta}_n) - p - \widetilde{q} - \widetilde{r}\,) \tag{5.3}$$

where q^* and r^* are as defined above and estimates are obtained in accordance with the discussion of Section 3.3.

6. Generalising and correcting the FPE and related discrepancy measures

In this section use the methodology developed to explore some additional strategies. In particular, we will discuss a generalisation of the final prediction error (FPE; Akaike (1969, 1970)) introduced in Linhart & Göttingen (1985). This extends the FPE to a broader class of time series models (the criterion was originally developed for autoregressive models). In the derivations they overlooked the bias introduced by the periodogram as an estimator for the unknown spectral density, however. The error is observed by the authors and there is a short comment in Linhart & Zucchini (1986, Section A.2.6), but it is incorrectly attributed to be completely caused by the use of the discrete approximation of the integral. This is a remark of interest, but simple simulation studies indicate that the scaled expected difference is zero; a complete discussion will require additional work.

6.1. Correcting the FPE for a general class of time series models.

In order to easily extend the FPE argument, the goal of minimising the one-step-ahead predictions in a finite sample, is changed to minimising the mean square error of the one-step-ahead prediction in the limit experiment. To make the paper more self-contained, we will provide a brief overview of the main results and show how we can obtain a discrepancy measure from this general idea; see also Linhart & Zucchini (1986, Ch. 12).

To simplify, we reorganise the observed series such that the task is now to predict y_0 given the observations $y_{-n}, \ldots, y_{-1}$, for $n \geq 1$, where $\widehat{y}_0 = \widehat{y}_0(y_{-1}, \ldots, y_{-n})$ is the predictor for the unobserved y_0. Under the conditions of Assumption 3.1, it follows from Azencott & Dacunha-Castelle (1986, Section 4.2) that if the model is correctly specified, then

$$\lim_{n\to\infty} \mathrm{E}_g\, |\widehat{y}_0 - y_0|^2 = 2\pi \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log g(\omega)\, \mathrm{d}\omega\right\} = \sigma_0^2(g). \tag{6.1}$$

The limit is shown by a neat trick using the determinant of the covariance matrix and the first Szegő theorem. If the model is not correctly specified, however, it follows from Grenander & Rosenblatt (1957, Section 8.1) that

$$\lim_{n\to\infty} \mathrm{E}_g |\widehat{y}_0 - y_0|^2 = \frac{\sigma_0^2(f_\theta)}{2\pi}\int_{-\pi}^{\pi} \frac{g(w)}{f_\theta(\omega)}\, \mathrm{d}\omega > 0, \tag{6.2}$$

where $\sigma_0^2(f_\theta)$ is as defined in (6.1). Since

$$\begin{aligned}\frac{\sigma_0^2(f_\theta)}{2\pi}\int_{-\pi}^{\pi} \frac{g(w)}{f_\theta(\omega)}\, \mathrm{d}\omega &= \sigma_0^2(g) \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\frac{f_\theta(\omega)}{g(\omega)}\, \mathrm{d}\omega\right\}\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{g(w)}{f_\theta(\omega)}\, \mathrm{d}\omega \\ &= \sigma_0^2(g) + \sigma_0^2(g)\left[\exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\frac{f_\theta(\omega)}{g(\omega)}\, \mathrm{d}\omega\right\}\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{g(w)}{f_\theta(\omega)}\, \mathrm{d}\omega - 1\right]\end{aligned}$$

the limit (6.2) motivates

$$d_0(g, f_\theta) = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log\frac{f_\theta(\omega)}{g(\omega)}\, \mathrm{d}\omega\right\}\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{g(w)}{f_\theta(\omega)}\, \mathrm{d}\omega - 1 \tag{6.3}$$

as a discrepancy measure. It is fairly easy to show that this is a valid discrepancy, since by Jensen's inequality the first half of (6.3) is greater or equal to one, with equality if and only if f_θ is equal to g almost everywhere, see Grenander & Rosenblatt (1957, p. 261).

The least false parameter value is (with a slight misuse of notation) $\theta_0 = \arg\min_\theta d_0(g, f_\theta) = \arg\min_\theta R_0(\theta)$, where

$$R_0(\theta) = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f_\theta(\omega)\,\mathrm{d}\omega\right\}\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{g(w)}{f_\theta(\omega)}\,\mathrm{d}\omega = \frac{\sigma_0^2(f_\theta)}{2\pi}\int_{-\pi}^{\pi}\frac{g(w)}{f_\theta(\omega)}\,\mathrm{d}\omega \tag{6.4}$$

takes the role of the model specific part of (6.3). Note that, in accordance with the related literature, we will now aim at minimising the model specific part instead of maximising it, but we could just as well have been working with $-R_0$ to retain the more familiar framework.

An estimator for (6.4) is obtained by replacing the unknown true spectral density with a non-parametric estimate. The natural choice is to use the periodogram, which leads to

$$\widetilde{R}_0(\theta) = \frac{\sigma_0^2(f_\theta)}{2\pi}\int_{-\pi}^{\pi}\frac{I_n(w)}{f_\theta(\omega)}\,\mathrm{d}\omega,$$

and also introduces the estimator $\overline{\theta}_n = \arg\min_\theta d_0(I_n, f_\theta) = \arg\min_\theta \widetilde{R}_0(\theta)$ for the unknown θ_0. Following the general theme of the paper, this also give rise to a model selection strategy by preferring the model that, in a set of competing models, attains the smallest value of $\mathrm{E}_g\, R_0(\overline{\theta}_0)$. As usual, the initial starting point $\widetilde{R}_0(\overline{\theta}_n)$ is biased and a model information criterion emerges from the derivation of a suitable first order bias correction.

Corollary 6.1. *Under the conditions of Assumption 3.1 the model robust version of the FPE is given by*

$$\mathrm{FPE}^*(f_\theta) = \frac{\sigma_0^2(f_{\overline{\theta}_n})}{2\pi}[1 + \mathrm{tr}\{J_0(I_n, f_{\overline{\theta}_n})^{-1}K_0(I_n/\sqrt{2}, f_{\overline{\theta}_n})\} + \langle I_n, 1/f_{\overline{\theta}_n}\rangle], \tag{6.5}$$

where $J_0(g, f_\theta) = 2\pi[\nabla^2 R_0(\theta)]/\sigma_0^2(f_\theta)$ *and*

$$K_0(g, f_\theta) = 4\pi\int_{-\pi}^{\pi}\left\{\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\nabla\Psi_\theta(\omega)\,\mathrm{d}\omega\right]\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\nabla\Psi_\theta(\omega)\,\mathrm{d}\omega\right]^{\mathrm{t}} - \nabla\Psi_\theta(\omega)\nabla\Psi_\theta(\omega)^{\mathrm{t}}\right\}\left[\frac{g(\omega)}{f_\theta(\omega)}\right]^2\mathrm{d}\omega.$$

If the model is correctly specified, meaning that f_{θ_0} *spans the true* g*, it follows that* $K_0 = 2\sigma_0^2(f_\theta)J_0$ *and the* FPE^* *formula simplifies accordingly.*

Proof. From Linhart & Göttingen (1985) we know that

$$\sqrt{n}(\overline{\theta}_n - \theta_0) \xrightarrow{d} J_0(g, f_{\theta_0})^{-1}U, \quad \text{where } U \sim \mathrm{N}_p\left(0, \left\{\frac{\sigma_0^2(f_\theta)}{2\pi}\right\}^2 K_0(g, f_{\theta_0})\right)$$

with J_0 and K_0 as defined above. Let $\widetilde{Q}_0 = \widetilde{R}_0(\overline{\theta}_n)$ be an estimate for the unknown $Q_0 = \mathrm{E}_g\, R_0(\overline{\theta}_n)$, then by a two-term Taylor expansion we obtain that

$$\begin{aligned}\widetilde{R}_0(\overline{\theta}_n) - \widetilde{R}_0(\theta_0) &= (\overline{\theta}_n - \theta_0)^{\mathrm{t}} \nabla \widetilde{R}_0(\theta_0) + \frac{1}{2}(\overline{\theta}_n - \theta_0)^{\mathrm{t}} \nabla^2 \widehat{R}_0(\bar{\theta}_n^{(1)})(\overline{\theta}_n - \theta_0) \\ &= \frac{1}{n} V_n^{\mathrm{t}} [\nabla^2 \widetilde{R}_0(\bar{\theta}_n^{(2)})^{-1}] V_n + \frac{1}{2n} V_n^{\mathrm{t}} [\nabla^2 \widetilde{R}_0(\bar{\theta}_n^{(1)})] V_n \\ &= \frac{1}{n} W_n + \frac{1}{2n} V_n^{\mathrm{t}} [\nabla^2 \widetilde{R}_0(\bar{\theta}_n^{(1)})] V_n\end{aligned}$$

since $0 = \nabla \widetilde{R}_0(\overline{\theta}_n) = \nabla \widetilde{R}_0(\theta_0) + (\overline{\theta}_n - \theta_0)^{\mathrm{t}} \nabla^2 \widetilde{R}_0(\bar{\theta}_n^{(2)})$, where $V_n = \sqrt{n}(\overline{\theta}_n - \theta_0)$ and $|\bar{\theta}_n^{(i)} - \theta_0| \leq |\overline{\theta}_n - \theta_0|$ for $i = 1, 2$. In addition, since by definition $\nabla R_0(\theta_0) = 0$ we may also write

$$R(\overline{\theta}_n) - R(\theta_0) = \frac{1}{2n} V_n^{\mathrm{t}} [\nabla^2 R_0(\bar{\theta}_n^{(3)})] V_n,$$

where $|\bar{\theta}_n^{(3)} - \theta_0| \leq |\overline{\theta}_n - \theta_0|$. By combining the above results, we may express the expected bias of $\widetilde{Q}_0$ as

$$\mathrm{E}_g\, \{\widetilde{Q}_0 - Q_0\} = \mathrm{E}_g\, \{\widetilde{R}_0(\overline{\theta}_n) - R_0(\theta_0)\} = \frac{1}{n} \mathrm{E}_g\, \{\epsilon_n - W_n + o_{P_g}(1)\}, \tag{6.6}$$

where $W_n \to_d W = U^{\mathrm{t}} [\nabla^2 R_0(\theta_0)] U$, and $\mathrm{E}_g\, W = \sigma_0^2(f_\theta)\, \mathrm{tr}\{J_0(g, f_{\theta_0})^{-1} K_0(g, f_{\theta_0})\}/2\pi$, see also Linhart & Göttingen (1985). Furthermore, in order to stay true to the claimed level of precision, we have to correct for the bias introduced by the periodogram, i.e. since

$$\mathrm{E}_g\, \epsilon_n = \mathrm{E}_g\, n(\widetilde{R}_0(\theta_0) - R_0(\theta_0)) = \frac{\sigma_0^2(f_\theta)}{2\pi} n \mathrm{E}_g \int_{-\pi}^{\pi} \frac{I_n(\omega) - g(\omega)}{f_\theta(\omega)}\, \mathrm{d}\omega = -\frac{\sigma_0^2(f_\theta)}{16\pi^3} \langle g, 1/f_{\theta_0} \rangle + o(1),$$

see also the derivation of q^* in Section 3.2. Summarising the results we now have that

$$n \mathrm{E}_g\, \{\widetilde{Q}_{1\,n} - Q_1\} \approx -\frac{\sigma_1^2(f_\theta)}{2\pi} \left[\mathrm{tr}\{J_0'(g, \theta_0)^{-1} K_0'(g, \theta_0)\} + \frac{1}{8\pi^2} \langle g, 1/f_{\theta_0} \rangle \right],$$

which is what we intended to show.

Finally, in order to obtain expressions for the two matrixes J_0 and K_0, it is sufficient to work with the first and second derivative of R_0. From

$$\nabla R_0(\theta) = \frac{\sigma_0^2(f_\theta)}{2\pi} \int_{-\pi}^{\pi} \left\{ \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \nabla \Psi_\theta(\omega)\, \mathrm{d}\omega \right] - \nabla \Psi_\theta(\omega) \right\} \frac{g(\omega)}{f_\theta(\omega)}\, \mathrm{d}\omega,$$

we are able to obtain K_0, see Taniguchi (1980) for details. Secondly, $J_0(g, f_\theta) = 2\pi[\nabla^2 R_0(\theta)]/\sigma_0^2(f_\theta)$, where

$$\begin{aligned}\nabla^2 R_0(\theta) = \frac{\sigma_0^2(f_\theta)}{2\pi} \Bigg\{ &\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \nabla \Psi_\theta(\omega)\, \mathrm{d}\omega \right] \int_{-\pi}^{\pi} \left(\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \nabla \Psi_\theta(\omega)\, \mathrm{d}\omega \right] - 2\nabla \Psi_\theta(\omega) \right) \frac{g(\omega)}{f_\theta(\omega)}\, \mathrm{d}\omega \\ &- \int_{-\pi}^{\pi} \left(\left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{g(\omega)}{f_\theta(\omega)}\, \mathrm{d}\omega \right] - \frac{g(\omega)}{f_\theta(\omega)} \right) \nabla^2 \Psi_\theta(\omega)\, d\omega - \int_{-\pi}^{\pi} \nabla \Psi_\theta(\omega) \nabla \Psi_\theta(\omega)^{\mathrm{t}} \frac{g(\omega)}{f_\theta(\omega)}\, \mathrm{d}\omega \Bigg\}.\end{aligned}$$

Note that in order to obtain consistent estimates, we simply replace the unknown true spectral density by its periodogram; see also Taniguchi (1980) for additional comments and discussion. □

6.2. **A general model selection strategy for the frequency domain.** The generalisation of the FPE in (6.5) and the derivation of AIC_∞ in Section 3 can be seen as members of a common family of model information criteria. To see this, let g and f_θ be spectral densities that again represent the true model and a parametric candidate from the set of competing models. Furthermore, suppose

$$D(g, f_\theta) = D(g) + R(g, f_\theta), \quad \text{where } R(g, f_\theta) = R(f_\theta) + \int_{-\pi}^{\pi} h_\theta(\omega) g(\omega)\, \mathrm{d}\omega, \tag{6.7}$$

is a valid discrepancy measure, i.e. $D(g, f_\theta) \geq 0$ with equality if and only if f_θ is equal to g almost everywhere, with $R(g, f_\theta)$ taking on the role as the model specific part and where $D(g)$ does not depend on the candidate model. This class is easily generalised to other types of measures, e.g. in Dahlhaus & Wefelmeyer (1996) the authors suggest working with functions of the form

$$D(g, f_\theta) = \int_{-\pi}^{\pi} T(f_\theta, g(\omega), \omega)\, \mathrm{d}\omega,$$

for a suitable smooth and interesting functional T. We will not go into details, however; see also Taniguchi (1979) for similar ideas.

With a little abuse of notation, assume that there exists a unique minimiser $\theta_0 = \arg\min_\theta D(g, f_\theta)$ and a nonparametric estimator $\widehat{g}_n$ such that $\widehat{\theta}_n = \arg\min_\theta D(\widehat{g}_n, f_\theta)$ is consistent for this least false parameter value θ_0. This motivates an alternative estimation procedure which has the potential to focus the estimation to optimise certain important features, which may be quite different to those obtained by maximum likelihood, e.g. minimising one-step-ahead predictions errors. Moreover, this also introduces a general family of model selection strategies by preferring the model that minimises the expected attained discrepancy $D(g, f_{\widehat{\theta}_n})$. The corresponding criterion is then derived following the familiar recipe of bias correcting the initial estimator $\widehat{D}_n = D(\widehat{g}_n, f_{\widehat{\theta}_n})$ for the unobtainable $\mathrm{E}_g\, D(g, f_{\widehat{\theta}_n})$; see also Linhart & Zucchini (1986) and in Taniguchi & Kakizawa (2000, Ch. 3.2.5) for some similar ideas.

A natural estimator for $Q = \mathrm{E}_g\, D(g, f_{\widehat{\theta}_n})$ is

$$\widehat{Q}_n = R(I_n, f_{\widehat{\theta}_n}) = R(f_{\widehat{\theta}_n}) + \int_{-\pi}^{\pi} h_{\widehat{\theta}_n}(\omega) I_n(\omega)\, \mathrm{d}\omega.$$

Moreover, suppose the expansions

$$\begin{aligned} R(I_n, f_{\widehat{\theta}_n}) - R(I_n, f_{\theta_0}) &\doteq \nabla R(I_n, f_{\theta_0})^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}}[-\nabla^2 R(I_n, f_{\theta_0})](\widehat{\theta}_n - \theta_0) \text{ and} \\ R(g, f_{\widehat{\theta}_n}) - R(g, f_{\theta_0}) &\doteq R(g, f_{\theta_0}) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}}[-\nabla^2 R(g, f_{\theta_0})](\widehat{\theta}_n - \theta_0), \end{aligned}$$

hold, with $\nabla^2 R(I_n, f_{\theta_0}) \to_{P_g} \nabla^2 R(g, f_{\theta_0}) = -J$ and $\nabla R(I_n, f_{\theta_0})^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) \to_d J^{-1}U$, for $U \sim \mathrm{N}_p(0, K)$, for corresponding matrices J and K. Then the arguments of Section 3.3 now justifies

$$\mathrm{E}_g\,[\widehat{Q}_n - Q] \approx \frac{1}{n}\left[\mathrm{tr}(J^{-1}K) - \frac{1}{8\pi^2}\langle g, h_{\theta_0}\rangle\right],$$

which in turn motivates a family of properly corrected model robust 'spectral information criteria' by the formula

$$\mathrm{SIC}^*(f_\theta) = nD(\widehat{\theta}_n, I_n) - \mathrm{tr}(\widehat{J}_n^{-1}\widehat{K}_n) + \frac{1}{8\pi^2}\langle I_n, h_{\widehat{\theta}_n}\rangle, \tag{6.8}$$

where $\widehat{J}_n$ and $\widehat{K}_n$ are consistent estimates for the matrices J and K above.

7. Illustrations

The general AIC formula is rarely used for selecting among time series models. The reason is that the AIC is commonly believed to have a tendency to prefer unnecessarily complex models. The effect is often shown for autoregressive models in small or moderate samples, as discussed in Hurvich & Tsai (1989). This artifact is also observed to become even worse if the models are fitted using conditional maximum likelihood, i.e. a trick used to simplify the likelihood function for the autoregressive models. The idea is to condition on the first $k_{\max}$ observations, where $k_{\max}$ is equal to the model order of the largest autoregressive model, which can be shown to simplify the likelihood and making it computational equivalent to a standard regression model, see McQuarrie & Tsai (1998, Ch. 3) for details.

The preference for large models is illustrated in the simulated example in Figure 7.1 (right), where the AIC scores based on conditional maximum likelihood are computed for autoregressive models of increasing order (dotted line). The true order is low, $k_0 = 3$, but it is easily seen that the higher order models are preferred and according to the AIC the best model is of order $k = 7$.

There are some attempts to correct for this and the best known adjustment is perhaps the corrected AIC formula (AIC_c) suggested in Hurvich & Tsai (1989). The difference between the corrected and uncorrected AIC is shown in Figure 7.1 (left), where the standard AIC (dotted line) is outperformed by AIC_c (dashed line) and AIC_u (solid line), the latter is a further refinement of AIC_c suggested in McQuarrie et al. (1997).

If we instead use standard maximum likelihood estimation and the full Gaussian model to calculate the scores, the performance of the AIC improves considerable, which is clearly illustrated in Figure 7.1 (middle). Note that the scores attained by the models of order $k \in \{3, 5, 8\}$ are almost identical making the evidence for the 'best' model of order $k = 3$ a little less convincing. A further improvement is obtained, however, from implementing the model robust AIC_n^* as shown in Figure 7.1 (middle, dashed black line). The reason is illustrated in Figure 7.1 (right, dashed black line) where the classical correction with p is compared with the model robust estimate $\widehat{p}^*$ from Section 4. Here we will expect (on average) that $\widehat{p}^*$ is close to p for all models with $k \geq 3$, since the models are nested, but for this particular realisations the larger models are given a slightly larger correction. Note the small dip in $\widehat{p}^*$ for the correct model, i.e. where $p = k + 1 = 4$.

The higher order models, estimated by conditional maximum likelihood, are often seen to provide a quite poor fit in comparison with the estimated periodogram, or the raw non-parametric covariance

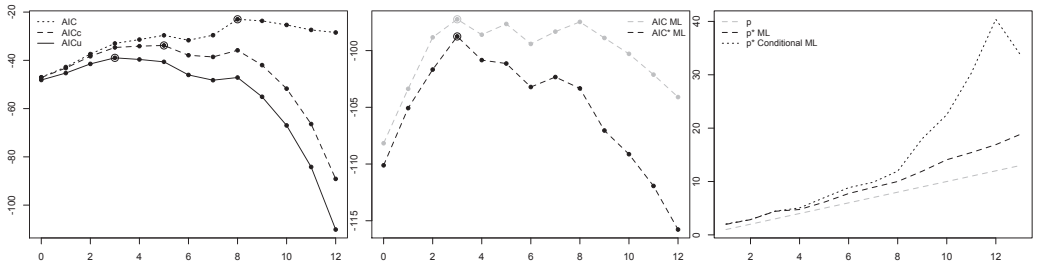

FIGURE 7.1. A simulated sample of size $n = 32$ of an autoregressive model with true order $k_0 = 3$, specified by the parameters $\sigma = 1$ and $\rho = (0.8, -0.5, 0.2)$, see Brockwell & Davis (1991) for conditions.

estimates. This is not always reflected by the AIC scores or the log-likelihood values. A more direct examination of these models would often reject them, however. This is actually the case for the illustration in Figure 7.1 (left), where the lack of proper fit is revealed by the corresponding estimated p^*. The reason is that the estimated J and K become very different and the value of $\widehat{p}^*$ is therefore much larger than p, as seen in Figure 7.1 (right, dotted black line).

Moreover, if the fitted model is close to the true model, we would expect that $\widehat{p}^*$ is close to p. This suggests that $\widehat{p}^*$ may be used as a model selection procedure on its own, where we prefer the model that minimises $|\dim(M) - \widehat{p}^*(M)|$. Finally, we have observed in simulations that if the estimated p^* explodes for some of the fitted models, it is often an indication that something has gone wrong in the estimation procedure, with the fitted model quite different from the non-parametric estimate. The problem often occurs when fitting high-dimensional models in small samples and is also sometimes caused by numerical instability in the optimisation procedure, which provides an additional argument for the model robust construction as a quick model check.

8. Concluding remarks

A. AIC_n *vs* AIC_∞. The two discrepancy measures KL_n vs KL_∞ defined in (1.3) are perhaps best interpreted as measures concerned with different parts of the underlying model. The KL_n measures a type of average performance in a new sample of the same size, while the limiting Kullback–Leibler divergence KL_∞ is concerned with the performance of the entire process. This last observation can be seen from the general structure, since it aims at measuring a type of discrepancy between spectral densities, which specifies the entire model. In this perspective, the two AIC formulations AIC_∞ and AIC_n are based on different underlying discrepancy measures and therefore aims at answering different problems.

B. Smoothed periodograms and tapers. The additional correction q emerged in AIC_∞ as a result of the first order bias introduced with the periodogram as an estimator for the unknown spectral

density. According to Taniguchi (1979, p. 580) and Taniguchi (1980, p. 74) no other nonparametric estimators can beat the periodogram at this in terms of variance. Our problem is with the bias, however, which suggests that smoothed and taper periodogram estimators should also be studied to see how this affect the overall bias of the estimate for $Q = \mathrm{E}_g\, R(\widetilde{\theta}_n)$.

C. Cross-validation. There is a connection between leave-one-out cross-validation, the AIC and the model robust AIC*, see e.g. Claeskens & Hjort (2008, Ch. 2.9). In the simple framework with i.i.d. observations this follows from the attempt to motivate a model selection strategy by preferring the model that maximises

$$\mathrm{xv}_n = \frac{1}{n}\sum_{i\le n} \log f(Y_i, \widehat{\theta}_{(i)}),$$

where $\widehat{\theta}_{(i)}$ is the maximum likelihood estimator based on the reduced dataset omitting observation Y_i. From this alternative large-sample approximation approach, it follows that $\mathrm{xv}_n \doteq \ell_n(\widehat{\theta}_n) - \widehat{p}^*$, where $p^* = \mathrm{tr}(J^{-1}K)$ for the corresponding matrices J and K, and where $\widehat{p}^*$ is obtained using suitable consistent estimates for both J and K. For the time series processes, it is now interesting to see whether such types of cross-validation approaches to model selection relate to either AIC_n or AIC_∞; this will require additional work and we will not go into details, however.

D. Detrending. The series worked with so far are all assumed to be stationary and with zero mean. In real life applications this is usually not the case and a common strategy is to detrend the observed series before analysing the dependency. This complicates the overall AIC argument considerably, as we now intend to show.

To easily see how, we consider the simplest non-trivial case where $y_1, \ldots, y_n$ are observations from the model $Y_t = \beta_0 + \epsilon_t$, where ϵ_t are elements from a stationary Gaussian time series with mean zero. In order to remove the effect of β_0, which is unknown, it is common to mean correct the observed series and work with $\widehat{y}_t = y_t - \bar{y}_n = \epsilon_t - \bar{\epsilon}_n$ as if this is actually stationary with mean zero. This seemingly innocent correction turns out to have a profound effect on the AIC argument.

Let $\widehat{I}_n(\omega) = (2\pi n)^{-1}|\sum_t \widehat{y}_t \exp\{-i\omega t\}|^2$ and I_n be the periodogram based on ϵ_t and observe that

$$\begin{aligned}\int_{-\pi}^{\pi} h(\omega)\widehat{I}_n(\omega)\,\mathrm{d}\omega &= \frac{1}{2\pi n}(\epsilon - \bar{\epsilon}_n)^{\mathrm{t}}\Sigma_n(h)(\epsilon - \bar{\epsilon}_n)\\ &= \int_{-\pi}^{\pi} h(\omega)I_n(\omega)\,\mathrm{d}\omega - \frac{1}{\pi n}\bar{\epsilon}_n(\epsilon - \bar{\epsilon}_n)^{\mathrm{t}}\Sigma_n(h)1 - \frac{1}{2\pi n}\bar{\epsilon}_n^2 1^{\mathrm{t}}\Sigma_n(h)1.\end{aligned}$$

The expectation of the analogue integral in the truly mean zero case is what introduces the correction term q and is a crucial component in the derivation of AIC_∞. To see how the detrended series affect this part of the bias, we observe that from Lemma A.5 in Dahlhaus (1996) it follows (under slightly stronger model conditions) that $1^{\mathrm{t}}\Sigma_n(h)1/(2\pi n) = h(0) + O(n^{-2/3}\log^{2k+2} n)$ provided h is sufficient

smooth. Furthermore,

$$\frac{1}{\pi n}\mathrm{E}_g[\epsilon^{\mathrm{t}}\Sigma_n(h)1\bar{\epsilon}_n] = \frac{1}{\pi n^2}\mathrm{E}_g[\epsilon^{\mathrm{t}}\Sigma_n(h)\mathbf{1}\epsilon^{\mathrm{t}}]$$
$$= \frac{1}{\pi n^2}\operatorname{tr}\{\Sigma_n(h)\mathbf{1}\Sigma_n(g)\} = \frac{1}{\pi n^2}1^{\mathrm{t}}\Sigma_n(h)\Sigma_n(g)1 = \frac{4\pi h(0)g(0) + O(n^{-2/3}\log^{2k+2} n)}{n},$$

which in summary results in

$$\mathrm{E}_g\int_{-\pi}^{\pi} h(\omega)\widehat{I}_n(\omega)\,\mathrm{d}\omega = \mathrm{E}_g\int_{-\pi}^{\pi} h(\omega)I_n(\omega)\,\mathrm{d}\omega + \frac{h(0)}{n}\{C_g(0) + 4\pi g(0)\} + o(1/n).$$

indicating that even a modest mean correction actually introduces additional correction terms and more complexity.

E. Alternative motivation for the Whittle log-likelihood. The connection with the asymptotic Kullback–Leibler divergence (3.3) establishes an independent motivation for the Whittle estimator, since it can be viewed as a direct attempt at estimating the least false parameter value θ_0 from the limit discrepancy measure, i.e. $\widetilde{\theta}_n = \arg\min_\theta d(I_n, f_\theta)$, see Taniguchi (1979) and Dahlhaus & Wefelmeyer (1996) for related ideas and comments.

9. Appendix

The following is an argument for why we are allowed to make the change from full Gaussian log-likelihood and maximum likelihood estimation to the Whittle approximation and related estimates in Section 3. Remember the original setup where we viewed $\widetilde{Q}_n$ as an estimator for $\mathrm{E}_g\, R(\widehat{\theta}_n)$. Then the Taylor expansion in (3.13) results in

$$R(\widehat{\theta}_n) = R(\theta_0) + \nabla R(\theta_0)^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) - \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}}J(\bar{\theta}_n^{(3)})(\widehat{\theta}_n - \theta_0) \doteq R(\theta_0) - \frac{1}{2n}V_n^{\mathrm{t}}J(\theta_0)V_n,$$

which does not alter the result in the final derivation in any way. In addition, if we use $\widetilde{R}_n(\widehat{\theta}_n)$ as an estimator for Q we still have the same version of the AIC formula. This is seen from the Taylor expansion in (3.12), which in the current framework gives

$$\widetilde{R}_n(\widehat{\theta}_n) \doteq \widetilde{R}_n(\theta_0) + \nabla\widetilde{R}_n(\theta_0)^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^{\mathrm{t}}\nabla^2\widetilde{R}_n(\theta_0)(\widehat{\theta}_n - \theta_0).$$

The first and the third term to the right do not cause any real problems to the general derivation. In order to obtain full control on the second term we apply a result form the proof of Theorem 3.3 in Dahlhaus & Wefelmeyer (1996) that (essentially) states that $\sqrt{n}(\nabla\ell_n(\theta_0) - \nabla\widetilde{\ell}_n(\theta_0)) \to_{P_g} 0$. This means that

$$\nabla\widetilde{R}_n(\theta_0)^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) \doteq \nabla\widehat{R}_n(\theta_0)^{\mathrm{t}}(\widehat{\theta}_n - \theta_0) \doteq (\widehat{\theta}_n - \theta_0)^{\mathrm{t}}J(\theta_0)(\widehat{\theta}_n - \theta_0),$$

where $\widehat{R}_n(\theta) = \ell_n(f_\theta)/n + \frac{1}{2}\log 2\pi$ and we get the same limit and expectation since the two estimators converges to the same distribution.

References

Akaeël, J. (1982). Fitting models in time series analysis. *Statistics: A Journal of Theoretical and Applied Statistics* **13**, 121–143.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, vol. 1. Springer.

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

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. *Computational Methods for Modeling of Nonlinear Systems* **126**, 27.

Azencott, R. & Dacunha-Castelle, D. (1986). *Series of Irregular Observations: Forecasting and Model Building*. Springer.

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston.

Brockwell, P. J. & Davis, R. (1991). *Time Series: Theory and Methods*. Springer.

Carslaw, H. S. (1921). *Introduction to the Theory of Fourier's Series & Integrals*. Glasgow University Press.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

Coursol, J. & Dacunha-Castelle, D. (1982). Remarks on the approximation of the likelihood function of a stationary Gaussian process. *Theory of Probability and Its Applications* **27**, 162–167.

Dahlhaus, R. (1996). Maximum likelihood estimation and model selection for locally stationary processes. *Nonparametric Statistics* **6**, 171–191.

Dahlhaus, R. & Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Annals of Statistics* **24**, 952–973.

Dzhaparidze, K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer.

Findley, D. (1985). On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *Journal of Time Series Analysis* **6**, 229–252.

Gray, R. (2006). *Toeplitz and Circulant Matrices: A Review*. Now Publisher.

Grenander, U. & Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series*. Wiley.

Grønneberg, S. & Hjort, N. L. (2014). The copula information criteria. *Scandinavian Journal of Statistics* **41**, 436–459.

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

Hermansen, G. H. & Hjort, N. L. (2014). Focused information criteria for time series. Tech. rep., University of Oslo and Norwegian Computing Centre.

Hurvich, C. & Tsai, C. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499.

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

Ioannidis, E. (2011). Akaike's information criterion correction for the least-squares autoregressive spectral estimator. *Journal of Time Series Analysis* **32**, 618–630.

Kullback, S. (1959). *Information Theory and Statistics*. Courier Dover Publications.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Linhart, H. & Göttingen, P. V. (1985). On a criterion for selection of models for stationary time series. *Metrika* **32**, 181–196.

Linhart, H. & Zucchini, W. (1986). *Model Selection*. Wiley.

McQuarrie, A., Shumway, R. & Tsai, C.-L. (1997). The model selection criterion AICu. *Statistics & Probability Letters* **34**, 285–292.

McQuarrie, A. D. R. & Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific Publishing.

Ogata, Y. (1980). Maximum likelihood estimates of incorrect Markov models for time series and the derivation of AIC. *Journal of Applied Probability* **17**, 59–72.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423.

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

Taniguchi, M. (1979). On estimation of parameters of Gaussian stationary processes. *Journal of Applied Probability* **16**, 575–591.

Taniguchi, M. (1980). On estimation of the integrals of certain functions of spectral density. *Journal of Applied Probability* **17**, 73–80.

Taniguchi, M. & Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer New York.

Whittle, P. (1953). The analysis of multiple stationary time series. *Journal of the Royal Statistical Society. Series B (Methodological)* **15**, 125–139.

Norwegian Computing Centre, University of Oslo, P.O. Box 114 Blindern, NO-0314 Oslo, Norway
E-mail address: gudmund.hermansen@nr.no

Department of Mathematics, University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway
E-mail address: nils@math.uio.no