# PARAMETRIC OR NONPARAMETRIC: THE FIC APPROACH FOR STATIONARY TIME SERIES

**Gudmund Hermansen, Nils Lid Hjort and Martin Jullum**

**Department of Mathematics, University of Oslo**

Abstract. We seek to narrow the gap between parametric and nonparametric modelling of stationary time series processes. The approach is inspired by recent advances in focused inference and model selection techniques. The paper generalises and extends recent work by developing a new version of the focused information criterion (FIC), directly comparing the performance of parametric time series models with a nonparametric alternative. For a pre-specified focused parameter, for which scrutiny is considered valuable, this is achieved by comparing the mean squared error of the model-based estimators of this quantity. In particular, this yields FIC formulae for covariances or correlations at specified lags, for the probability of reaching a threshold, etc. Suitable weighted average versions, the AFIC, also lead to model selection strategies for finding the best model for the purpose of estimating e.g. a sequence of correlations.

*Key words:* focused inference, model selection, time series modelling, risk estimation

## 1. Introduction and summary

The focused information criterion (FIC) was introduced in Claeskens & Hjort (2003) and is based on estimating and comparing the accuracy of model-based estimators for a chosen focus parameter. This focus, say $\mu$, ought to have a clear statistical interpretation across candidate models. For a given candidate model, $\mu$ is traditionally expressed as a function of this model's parameters. In general, the focus parameter can be any sufficiently smooth and regular function of the underlying model parameters, or more generally its spectral distribution. This includes quantiles, regression coefficients, a specified lagged correlation, but also various types of predictions and data dependent functions, to name some; see Hermansen & Hjort (2015) for a more complete list and discussion of valid focus parameters for time series models.

Suppose there are candidate models $M_1, \ldots, M_k$, leading to focus parameter estimates $\widehat{\mu}_1, \ldots, \widehat{\mu}_k$, respectively. The underlying idea leading to the FIC is to estimate the mean squared error (mse) of $\widehat{\mu}_j$ for each candidate model and then select the model that achieves the smallest value. The mse in question is

$$\text{mse}_j = \text{E}\,(\widehat{\mu}_j - \mu_{\text{true}})^2 = \text{bias}(\widehat{\mu}_j)^2 + \text{Var}\,\widehat{\mu}_j,$$

*Date*: December 2015.

comprising the variance and the squared bias in relation to the true parameter value $\mu_{\rm true}$. Thus the FIC consists of finding ways of assessing, approximating and then estimating the $\text{mse}_j$ for each candidate model. The winning model is the one with smallest $\widehat{\text{mse}}_j$. How this may be done depends on both the candidate models and the focus parameter, as well as on other characteristics of the underlying situation. The FIC apparatus hence leads to different types of formulae in different setups; see Claeskens & Hjort (2008, Ch. 5 & 6) for a fuller discussion and illustrations of such criteria for selection among parametric models.

Most FIC constructions have been derived by relying on a suitably defined local misspecification framework, see again Claeskens & Hjort (2008, Ch. 5 & 6). In such a framework the true model is assumed to gradually shrink with the sample size, starting from the biggest 'wide' model and hitting the simplest 'narrow' model in the limit. In addition, and all candidate models need to lie between these two model extremes. In the various data settings, such frameworks typically result in squared biases and variances of the same asymptotic order, motivating certain approximation formulae for the $\widehat{\text{mse}}_j$ in question. In Hermansen & Hjort (2015) such a framework is used to derive FIC machinery for choosing between parametric time series models within broad classes of time series models. See Section 7.5 for some further remarks.

The aim of the present paper is to derive FIC machinery which will justify comparison and selection among both parametric and nonparametric candidate models. The derivation will be somewhat different from that of Claeskens & Hjort (2003) and Hermansen & Hjort (2015) in that we do not rely on a certain local misspecification framework. We rather take a more direct approach following reasoning similar to the development of Jullum & Hjort (2015), where focused inference and model selection among parametric and nonparametric models are developed for independent observations. By including a nonparametric candidate among the parametric models, we will in particular be able to detect whether our parametric models are off-target. This FIC construction, with a nonparametric alternative, therefore has a built-in insurance mechanism against poorly specified parametric candidates. When one or more parametric models are adequate, such are selected as they typically have lower variance.

Though our methods will be extended to more general setups later, we start our developments with the class of zero-mean stationary Gaussian time series processes. Let $\{Y_t\}$ be such a process. Then the dependency structure, which in such cases determines the entire model, is completely specified by the corresponding covariance function $C(k) = \text{cov}(Y_t, Y_{t+k})$, defined for all lags $k = 0, 1, 2, \ldots$. Here we will, for mathematical convenience, work with the frequency representation, where the covariance function $C(k)$

can be represented by a unique spectral distribution $G$ such that

$$C(k) = \int_{-\pi}^{\pi} e^{ik\omega}\,\mathrm{d}G(\omega) = 2\int_0^{\pi} \cos(k\omega)g(\omega)\,\mathrm{d}\omega, \tag{1.1}$$

provided the corresponding spectral distribution $G$ has a continuous and symmetric density $g$. See among others Brillinger (1975), Priestley (1981) or Dzhaparidze (1986) for a general introduction to time series modelling in the frequency domain. When necessary, we will write $C_g$ to indicate that this is the covariance indexed by the spectral density $g$. Note also that we can obtain the spectral density as the Fourier transform of the covariance function.

The types of parametric models we will consider are typically the classical autoregressive (AR), moving average (MA) and the mixture (ARMA), all of which have clear and well defined corresponding spectral densities; see e.g. Brockwell & Davis (1991) for an introduction to time series modelling with such models. Note that the theory developed here is general, and that there is nothing other than convenience that restricts us to these particular classes of parametric models. For an observed series $y_1, \ldots, y_n$, the raw periodogram

$$I_n(\omega) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} y_t \exp(i\omega t)\right|^2, \quad \text{for } -\pi \le \omega < \pi, \tag{1.2}$$

will be our favourite nonparametric model for the underlying spectral density. The main reason for not considering variations of smoothed or tapered periodogram estimators is that we are interested in focus parameters that involves functions of the integrated spectrum, which essentially is a type of smoothing, rendering the pre-smoothing of the raw periodogram less critical and often unnecessary.

We will start out considering a class of focus functions of the type

$$\mu(G; h_0) = \int_{-\pi}^{\pi} h_0(\omega)\,\mathrm{d}G(\omega), \tag{1.3}$$

where $h_0$ is a piecewise continuous and bounded function on $[-\pi, \pi]$, with potentially a finite number of jump discontinuities. This class includes e.g. the covariance function, which is easily seen from (1.1) above, and allows studying specific parts of the spectral density by using indicator functions; see also Gray (2006) for further illustrations involving quantities of type (1.3).

Finding the best model to estimate the integrated spectrum (or total power/energy) over a specific region, may be an interesting and important applications in several areas of research; like pharmacology, astronomy, oceanography and in the interpretation of seismic data. The reason is that in all of these situations the observed time series is converted into the associated spectra, where the processed spectral density and especially the energy over certain regions of frequencies, have clear interpretations. For example, in

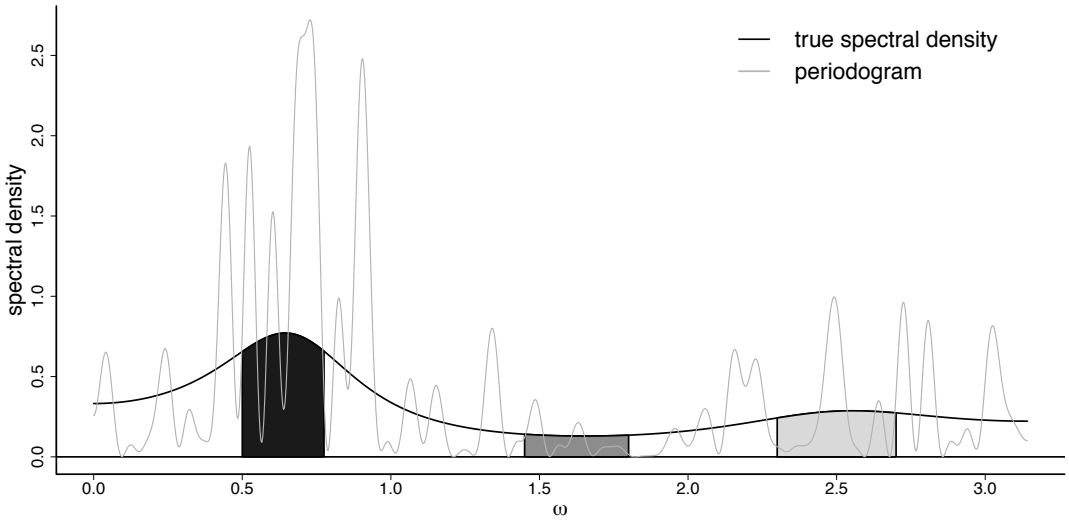


FIGURE 1.1. The true spectral density and the raw periodogram from a simulated autoregressive time series of order 4, with length $n = 100$ and parameters $\rho = (0.2, 0.2, -0.1, -0.2)$ and $\sigma = 1.30$. The shaded regions corresponds to three different focus parameters, namely, the integrated spectrum (or total energy) over that particular region.

pharmacology the spectrum of EEG/ERP signals may be used to quantify certain brain functions, indicating e.g. the effect of a potential drug. In such applications, the different models may not always have clear interpretations as time series, per se. The FIC is nevertheless able to rank the fitted models in terms of estimated precision of estimates, for the focus parameter in question. This general idea and particular usage of the FIC is illustrated in Figures 1.1 and 1.2 using simulated data from an autoregressive model of order 4, for focus parameters

$$\mu_j = \int_0^\pi I(a_j \le \omega < b_j) g(\omega)\, \mathrm{d}\omega = G(b_j) - G(a_j),$$

for $j = 1, 2$ and 3, for the corresponding intervals $(a_j, b_j) \subset [0, \pi)$; which are marked by the shaded regions in Figure 1.1. The candidate models are the autoregressive models of order 0–4 and a nonparametric alternative based on integrating the raw periodogram (1.2). The AR-model of order 0 corresponds to the independence model. Here, the FIC works well: For each focus parameter it prefers models that all results in estimates that are reasonably close to the true value; which in terms if rmse (and absolute deviation from the truth) is not always the nonparametric or true model of order 4. Moreover, this example also illustrates a second and important concept, namely, that one and the same model is not necessarily best for all focus parameters. Note that the FIC prefers an AR(3), AR(4) and AR(1) for the respective regions 1, 2, 3.

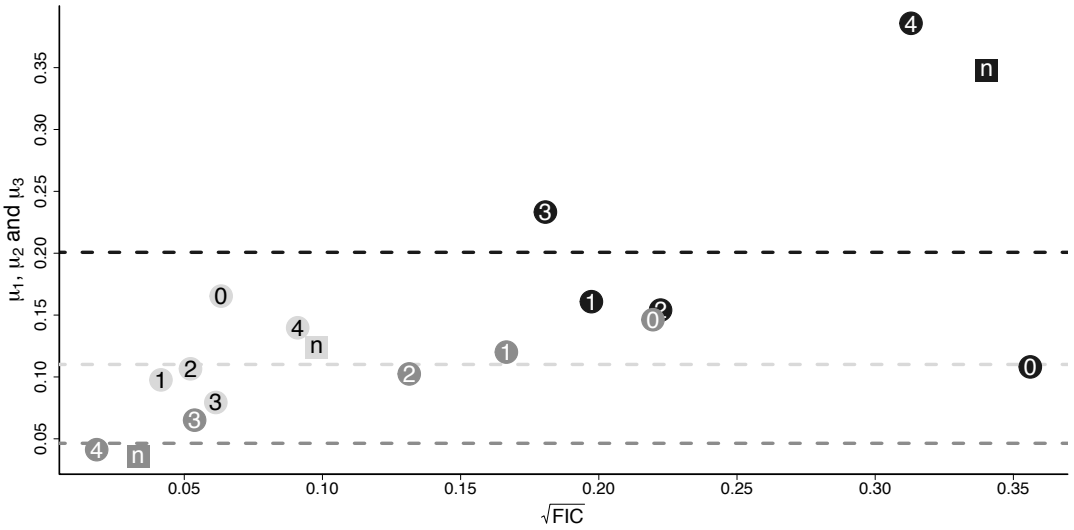


FIGURE 1.2. The horizontal lines indicate the true spectral density over the three shaded regions (of the same colour) shown in Figure 1.1; the three focus parameters $\mu_1, \mu_2$ and $\mu_3$. The corresponding coloured dots show the performance, in terms of the root of the FIC score for the nonparamteric model based on the periodogram (n) and the autoregressive models of order 0–4, where 0 represent the model with independent.

A class of focus parameters wider than that of (1.3) takes focus parameters of the form

$$\begin{aligned}\mu(G;h,H) &= H(\mu(G;h_1),\ldots,\mu(G;h_k)) \\ &= H\Big(\int_{-\pi}^{\pi} h_1(\omega)\,\mathrm{d}G(\omega),\ldots,\int_{-\pi}^{\pi} h_k(\omega)\,\mathrm{d}G(\omega)\Big),\end{aligned} \tag{1.4}$$

for a $k$-dimensional vector function $h(\omega) = (h_1(\omega),\ldots,h_k(\omega))^{\mathrm{t}}$, where each of the $h_j$ is of the above type, and $H(x_1,\ldots,x_k)$ a continuously differentiable function of the $x_j = \mu(G;h_j), j = 1,\ldots,k$. The direct correlations

$$\mathrm{corr}(Y_t,Y_{t+k}) = \frac{\mathrm{cov}(Y_t,Y_{t+k})}{\sigma^2} = \frac{C(k)}{C(0)} = \frac{\int_0^\pi \cos(k\omega)\,\mathrm{d}G(\omega)}{\int_0^\pi \mathrm{d}G(\omega)},$$

for example, are of type (1.4). Another class of estimands captured by (1.4) are conditional threshold probabilities, say $P\{Y_{n+1} \geq y \,|\, Y_n = y_n,\ldots,Y_{n-k} = y_{n-k}\}$, as these are functions of the $(k+1)\times(k+1)$ covariance matrix for $(Y_{n-k},\ldots,Y_n,Y_{n+1})$. Later results will allow us to reach FIC formulae for this more general class.

In Section 2 we provide a brief overview of some standard results needed to obtain good estimates for various mean squared error quantities. Among other aspects we need properties of maximum likelihood- or Whittle approximated estimators outside the model, and some large-sample results regarding the periodogram. Then in Section 3 we motivate and develop such mean squared error estimators, leading to FIC formulae. In Section 4 we

show that under certain conditions, a detrended time series may be handled by our FIC scheme as if it was the original time series. In Section 5 we extend the FIC methodology by deriving an average weighted focused information criterion which aims at selecting the best model for estimating a full set of focus parameters, possibly weighted to reflect their relative importance for the analysis. In Section 6 we discuss certain theoretical behavioural aspects of the derived FIC scheme, and present the results from a simulation study. Some concluding remarks, some of which pointing to future work, are finally provided in Section 7.

## 2. Estimation and approximations

We start out investigating the behaviour of the two most common parametric estimation procedures, those based on the maximum likelihood method and the associated Whittle approximation to the log-likelihood. We also give some basics for nonparametric modelling.

2.1. **Maximum likelihood estimation outside the model.** Let $\underline{y}_n = (y_1, \ldots, y_n)^{\mathrm{t}}$ be a collection of $n$ realisations from a zero mean stationary Gaussian time series process with spectral distribution function $G$ and corresponding spectral density $g$. Furthermore, let the spectral distribution function $F_\theta$ and its corresponding spectral density $f_\theta = f(\cdot;\theta)$ index an arbitrary parametric candidate model, where $\theta$ belongs to some parameter space $\Theta$ of dimension say $p$. The corresponding full log-likelihood is

$$\ell_n(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_n(f_\theta)| - \frac{1}{2}\underline{y}_n^{\mathrm{t}}\Sigma_n(f_\theta)^{-1}\underline{y}_n, \tag{2.1}$$

where $\Sigma_n(f_\theta)$ is the covariance matrix with elements

$$C_{f_\theta}(|s-t|) = 2\int_0^\pi \cos(\omega|s-t|)f_\theta(\omega)\,\mathrm{d}\omega$$

for $s,t = 1,\ldots,n$. Since the class of parametric candidate models is not assumed to necessarily include the true $g$, the maximum likelihood estimator does not converge to a 'true' parameter value. Instead it converges to the so-called least false parameter value, i.e. $\widetilde{\theta}_n = \mathrm{argmax}_\theta\{\ell_n(\theta)\} \to_p \mathrm{argmin}_\theta\{d(g,f_\theta)\} = \theta_0$, where

$$\begin{aligned} d(g,f_\theta) &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\Big\{\frac{g(\omega)}{f_\theta(\omega)} - 1 - \log\frac{g(\omega)}{f_\theta(\omega)}\Big\}\,\mathrm{d}\omega \\ &= -\frac{1}{4\pi}\int_{-\pi}^{\pi}\{\log g(\omega)+1\}\,\mathrm{d}\omega - R(G,\theta), \end{aligned} \tag{2.2}$$

and where

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

may be referred to as the model specific part, see e.g. Dahlhaus & Wefelmeyer (1996) for details. Furthermore, it can be shown that

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d J_0^{-1}U \sim \mathrm{N}_p(0, J_0^{-1}K_0J_0^{-1}), \quad \text{where } U \sim \mathrm{N}_p(0, K_0), \tag{2.3}$$

with $J_0$ and $K_0$ defined by

$$\begin{aligned} J_0 &= J(g, f_{\theta_0}) \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\Big[\nabla\Psi_{\theta_0}(\omega)\nabla\Psi_{\theta_0}(\omega)^{\mathrm{t}}g(\omega) + \nabla^2\Psi_{\theta_0}(\omega)\{f_{\theta_0}(\omega) - g(\omega)\}\Big]\frac{1}{f_{\theta_0}(\omega)}\,\mathrm{d}\omega \end{aligned}$$

and

$$K_0 = K(g, f_{\theta_0}) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\nabla\Psi_{\theta_0}(\omega)\nabla\Psi_{\theta_0}(\omega)^{\mathrm{t}}\Big\{\frac{g(\omega)}{f_{\theta_0}(\omega)}\Big\}^2\,\mathrm{d}\omega,$$

where $\Psi_\theta(\omega) = \log f_\theta(\omega)$. and $\nabla\Psi_\theta(\omega)$ and $\nabla^2\Psi_\theta(\omega)$ are respectively the vector of partial derivatives and matrix of second order partial derivatives with respect to $\theta$, see Dahlhaus & Wefelmeyer (1996, Theorem 3.3). Note that $J_0 = K_0$ under model conditions.

2.2. **The Whittle approximation.** The Whittle pseudo-log-likelihood is an approximation to the full Gaussian log-likelihood $\ell_n$ of (2.1). It was originally suggested by P. Whittle in the 1950s (cf. Whittle (1953)), and is defined as

$$\widehat{\ell}_n(\theta) = -\tfrac{1}{2}n\Big[\log(2\pi) + \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\{2\pi f_\theta(\omega)\}\,\mathrm{d}\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{I_n(\omega)}{f_\theta(\omega)}\,\mathrm{d}\omega\Big], \tag{2.4}$$

where $I_n(\omega) = (2\pi n)^{-1}|\sum_{t\le n} y_t\exp(i\omega t)|^2$ is the periodogram. This approximation is close to the full Gaussian log-likelihood in the sense that $\ell_n(\theta) = \widehat{\ell}_n(\theta) + O_p(1)$ uniformly in $f$, see Coursol & Dacunha-Castelle (1982). More important here, however, is that (2.4) motivates an alternative estimation procedure, namely the Whittle estimator $\widehat{\theta}_n = \mathrm{argmax}_\theta\{\widehat{\ell}_n(f_\theta)\}$. This estimator is easier to work with in practice (both analytically and numerically) and shares several properties with the maximum likelihood estimator. In particular $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ achieves the same limit distribution as in (2.3), with the same least false parameter value $\theta_0$ as defined in relation to (2.2); see Dahlhaus & Wefelmeyer (1996) for details. This means that in a large-sample perspective, the maximum likelihood estimator and the simpler Whittle estimator are equally efficient and essentially interchangeable.

2.3. **Nonparametric modelling.** As mentioned in the introduction, we shall use the periodogram in (1.2) for nonparametric modelling. Under appropriate short memory conditions, it follows from Brillinger (1975, Theorem 5.5.2) that $\mathrm{E}\{I_n(\omega)\} = g(\omega) + O(n^{-1})$, i.e. that the periodogram is asymptotically unbiased as an estimator of the spectral density. We shall thus use

$$\widehat{G}_n(\omega) = \int_{-\pi}^{\omega} I_n(u)\,\mathrm{d}u, \tag{2.5}$$

as a canonical estimator for the spectral distribution $G$; for which

$$\sqrt{n}(\widehat{G}_n(\omega) - G(\omega)) \to_d \mathrm{N}\bigg(0, 4\pi \int_{-\pi}^{\omega} g(u)^2 \,\mathrm{d}u\bigg),$$

see e.g. Taniguchi (1980).

## 3. Parametric versus nonparametric

We shall now obtain large-sample approximations for the focus parameter estimators. These shall then be used to construct approximate mse formulae for each model's estimator of the focus parameter. When estimated these mses then give the FIC formulae.

3.1. **How to compare parametric and nonparametric models?** In completely general terms, let $\mu(G)$ be a focus function, i.e. a functional mapping of the spectral distribution $G$ to a scalar value. This may be estimated parametrically by estimators of the form $\widehat{\mu}_{\rm pm} = \mu(F_{\widehat{\theta}_n})$, or nonparametrically by $\widehat{\mu}_{\rm np} = \mu(\widehat{G}_n)$. Other estimators of $\theta$ and $G$ may also be used, however. Typically, the collection of parametric candidate models does not include the true $G$. The question is then which model should we use – parametric or nonparametric – for estimating the focus parameter.

Assume for the nonparametric and each of the parametric candidate models that

$$\sqrt{n}(\widehat{\mu}_{\rm np} - \mu_{\rm true}) \to_d \mathrm{N}(0, v_{\rm np}) \quad \text{and} \quad \sqrt{n}(\widehat{\mu}_{\rm pm} - \mu_0) \to_d \mathrm{N}(0, v_{\rm pm}),$$

where $\mu_{\rm true} = \mu(G)$ is the true value of the focus parameter and $\mu_0 = \mu(F_{\theta_0})$ is the focus function evaluated under the least false parametric model $F_{\theta_0}$ as discussed in relation to (2.2). Then, without going into details, the large-sample results above motivate the following first-order approximations for the mse of the estimated focus parameters:

$$\mathrm{mse}_{\rm np} = 0^2 + v_{\rm np}/n = v_{\rm np}/n \quad \text{and} \quad \mathrm{mse}_{\rm pm} = b^2 + v_{\rm pm}/n, \tag{3.1}$$

where $b = \mu_0 - \mu_{\rm true}$. The remainder of this section will be used to motivate and obtain good estimators for the mean squared errors in (3.1) with the class of focus paramters of the form $\mu(G; h_0)$ defined in (1.3), and the more general $\mu(G; h, H)$ in (1.4).

3.2. **Deriving unbiased risk estimates.** In the derivation below, the parametric candidates $F_\theta$ will be fitted using the Whittle estimator $\widehat{\theta}_n$ as defined in (2.4), while we will use the canonical periodogoram based estimator in (2.5) for nonparametric estimation of the spectral distribution $G$.

Using the Whittle estimator in collaboration with (2.5) results in a convenient simplification of the derivations below; extending the arguments to full ML estimation is relatively straightforward, using techniques in Dahlhaus & Wefelmeyer (1996). This motivates the

following nonparametric and parametric estimators for focus parameters $\mu(G;h_0)$ on the form of (1.3):

$$\widehat{\mu}_{\rm np} = \int_{-\pi}^{\pi} h_0(\omega) I_n(\omega)\,{\rm d}\omega = \frac{1}{n}\underline{y}_n^{\rm t}\Sigma_n(h_0)\underline{y}_n \quad \text{and} \quad \widehat{\mu}_{\rm pm} = \int_{-\pi}^{\pi} h_0(\omega) f_{\widehat{\theta}_n}(\omega)\,{\rm d}\omega,$$

where $\Sigma_n(h_0)$ is a $n\times n$-dimensional symmetric Toeplitz matrix, having elements of the general form

$$\sigma_{n,s,t}(h_0) = \int_{-\pi}^{\pi} \cos(\omega|s-t|)h_0(\omega)\,{\rm d}\omega.$$

for $s,t=1,\ldots,n$. The following proposition establishes the joint limit distribution for the estimators above (suitably normalised), which in turn will be used to obtain good approximations for their respective mean squared errors.

**Proposition 1.** *Let $y_1,\ldots,y_n$ be realisations from a stationary Gaussian time series model with spectral density $g$ assumed to be uniformly bounded away from both zero and infinity. Suppose $|h_0|$ is bounded in $\omega$, that $f_\theta$ is two times differentiable with respect to $\theta$, and that $f_\theta$ and these derivatives, $\nabla f_\theta$ and $\nabla^2 f_\theta$, are continuous and uniformly bounded in both $\omega$ and $\theta$ in a neighbourhood of the least false parameter value $\theta_0$ as defined in (2.2) above. Then*

$$\begin{pmatrix} \sqrt{n}(\widehat{\mu}_{\rm np}-\mu_{\rm true}) \\ \sqrt{n}(\widehat{\mu}_{\rm pm}-\mu_0)\end{pmatrix} \to_d \begin{pmatrix} X_0 \\ c_0^{\rm t}J(g,f_{\theta_0})^{-1}U\end{pmatrix} \sim {\rm N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} v_{\rm np} & v_c \\ v_c & v_{\rm pm}\end{pmatrix}\right), \tag{3.2}$$

*where*

$$v_{\rm np} = 4\pi\int_{-\pi}^{\pi}\{h_0(\omega)g(\omega)\}^2\,{\rm d}\omega \quad \text{and} \quad v_{\rm pm} = c_0^{\rm t}J(g,f_{\theta_0})^{-1}K(g,f_{\theta_0})J(g,f_{\theta_0})^{-1}c_0,$$

*with $J$ and $K$ as defined below (2.3), and $v_c = c_0^{\rm t}J(g,f_{\theta_0})^{-1}d_0$, where the $c_0$ is the partial derivative of $\mu(F_{\theta_0};h)$ with respect to $\theta$, i.e. $c_0 = \nabla\mu(F_{\theta_0};h) = \int_{-\pi}^{\pi} h_0(\omega)\nabla f_{\theta_0}(\omega)\,{\rm d}\omega$ and*

$$d_0 = {\rm cov}(X,U) = \int_{-\pi}^{\pi}\frac{\nabla f_{\theta_0}(\omega)h_0(\omega)g(\omega)^2}{f_{\theta_0}(\omega)^2}\,{\rm d}\omega.$$

*Proof.* It follows from the results in (Dzhaparidze, 1986, Ch. 2) that $\widehat{\theta}_n-\theta_0 = J(g,f_{\theta_0})^{-1}U_n + o_p(1/\sqrt{n})$, where $U_n = \nabla\widehat{\ell}_n(f_{\theta_0})$ and

$$U_n = -\tfrac{1}{2}\{{\rm Tr}(\Sigma_n(\nabla\Psi_{\theta_0})) - \underline{y}_n^{\rm t}\Sigma_n(\nabla\Psi_{\theta_0}/f_{\theta_0})\underline{y}_n\},$$

where $\Psi_{\theta_0} = \log f_{\theta_0}$ and $\nabla\Psi_{\theta_0}$ is the vector of its partial derivatives. As a consequence, a Taylor expansion motivated by the standard delta method gives $\widehat{\mu}_{\rm pm}-\mu_0 = c_0^{\rm t}J(g,f_{\theta_0})^{-1}U_n + o_p(1/\sqrt{n})$. Since $\sqrt{n}U_n \to_d U$ by the assumptions of the proposition (Dzhaparidze, 1986), the parametric part of the result holds. In addition

$$X_n = (\widehat{\mu}_{\rm np}-\mu_{\rm true}) = \frac{1}{n}\underline{y}_n^{\rm t}\Sigma_n(h_0)\underline{y}_n - \mu_{\rm true},$$

which can be shown, by a modified version of the argument leading to the limit distribution of $U_n$, to have the property that $\sqrt{n}X_n \to_d X_0 \sim \mathrm{N}(0, v_{\mathrm{np}})$. This proves the nonparametric part of the result. We finally need to show that these convergence results hold jointly. Since the two drivers in the derivation of the limit distribution, $\underline{y}_n^{\mathrm{t}}\Sigma_n(h_0)\underline{y}_n/n$ and $U_n$, are quadratic forms, the joint limit distribution is readily obtainable by a Cramér–Wold type of argument. To see how, let $a$ be a vector in $\mathbb{R}^2$ to be used in the Cramér–Wold argument, and define

$$\Lambda_n = a_1\sqrt{n}X_n + a_2\sqrt{n}U_n = \frac{1}{\sqrt{n}}\underline{y}_n^{\mathrm{t}}\Sigma_n(a_1h_0 + a_2\nabla\Psi_{\theta_0}/f_{\theta_0})\underline{y}_n + \gamma_n$$

where $\gamma_n = \sqrt{n}\{a_1\mu_{\mathrm{true}} - a_2\mathrm{Tr}(\Sigma_n(\nabla\Psi_{\theta_0}))/2\}$. The $\gamma_n$ cancels out the mean, here, such that $\Lambda_n$ has mean zero. This is once again just a quadratic form, hence, $\Lambda_n$ is normal under the assumptions of the proposition; see Dzhaparidze (1986) or Hermansen & Hjort (2014b) for derivations of a similar type. The proof is completed by observing that by Dahlhaus & Wefelmeyer (1996, Lemma A.5), the covariances take the relevant form

$$\mathrm{cov}(X_n, U_n) = \frac{2}{n}\mathrm{Tr}\{\Sigma_n(h_0)\Sigma_n(g)\Sigma_n(\nabla\Psi_\theta/f_\theta)\Sigma_n(g)\} \to \int_{-\pi}^{\pi}\frac{\nabla f_{\theta_0}(\omega)h_0(\omega)g(\omega)^2}{f_{\theta_0}(\omega)^2}\,\mathrm{d}\omega.$$

□

We next extend the above proposition to the more general class of We next extend the above proposition to the more general class of focus parameters $\mu(G; h, H)$ in (1.4), being a continuously differentiable function of a finite number of the $\mu(G; h_0)$ functions. The nonparametric and parametric estimators for this class take the form

$$\widehat{\mu}_{\mathrm{np}} = H\big(n^{-1}\underline{y}_n^{\mathrm{t}}\Sigma_n(h_1)\underline{y}_n, \ldots, n^{-1}\underline{y}_n^{\mathrm{t}}\Sigma_n(h_k)\underline{y}_n\big)$$

and

$$\widehat{\mu}_{\mathrm{pm}} = H\Big(\int_{-\pi}^{\pi}h_1(\omega)f(\omega;\widehat{\theta}_n)\,\mathrm{d}\omega, \ldots, \int_{-\pi}^{\pi}h_k(\omega)f(\omega;\widehat{\theta}_n)\,\mathrm{d}\omega\Big).$$

**Proposition 2.** *Under the conditions of Proposition 1 the focus parameters $\mu(G; h, H)$ in (1.4), with estimators and estimands as above, fulfils*

$$\begin{pmatrix}\sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu_{\mathrm{true}})\\ \sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_0)\end{pmatrix} \to_d \begin{pmatrix}\nabla H_{\mathrm{np}}X\\ \nabla H_{\mathrm{pm}}c^{\mathrm{t}}J(g, f_{\theta_0})^{-1}U\end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}v_{\mathrm{np}} & v_c\\ v_c & v_{\mathrm{pm}}\end{pmatrix}\right), \quad (3.3)$$

*where*

$$v_{\mathrm{np}} = \nabla H_{\mathrm{np}}\{4\pi\int_{-\pi}^{\pi}\{h(\omega)g(\omega)\}^2\,\mathrm{d}\omega\}\nabla H_{\mathrm{np}}^{\mathrm{t}} \quad \text{and}$$
$$v_{\mathrm{pm}} = \nabla H_{\mathrm{pm}}c^{\mathrm{t}}J(g, f_{\theta_0})^{-1}K(g, f_{\theta_0})J(g, f_{\theta_0})^{-1}c\nabla H_{\mathrm{pm}}^{\mathrm{t}},$$

*and $v_c = \nabla H_{\mathrm{pm}}c^{\mathrm{t}}J(g, f_{\theta_0})^{-1}d\,\nabla H_{\mathrm{np}}^{\mathrm{t}}$, where $\nabla H_{\mathrm{np}}$ and $\nabla H_{\mathrm{pm}}$ are the gradients of $H$ evaluated at respectively $(\mu(G; h_1), \ldots, \mu(G; h_k))$ and $(\mu(F_{\theta_0}; h_1), \ldots, \mu(F_{\theta_0}; h_k))$, $c$ is the*

*$k \times p$-dimensional matrix with rows given by $\nabla\mu(F_{\theta_0}; h_j), j = 1, \ldots, k$ and*

$$d = \operatorname{cov}(X, U) = \int_{-\pi}^{\pi} \frac{\nabla f_{\theta_0}(\omega) h(\omega) g(\omega)^2}{f_{\theta_0}(\omega)^2} \, \mathrm{d}\omega.$$

*Proof.* By Propostion 1, we see that (3.2) holds for each $\mu(G; h_j)$. Let now $X_{n,j} = \frac{1}{n}\underline{y}_n^{\mathrm{t}} \Sigma_n(h_j) \underline{y}_n - \mu_{\mathrm{true}}$ for $j = 1, \ldots, k$. By extending the Cramér–Wold argument in Propostion 1 to all of $X_{n,1}, \ldots, X_{n,k}, U_n$, we see that there is joint convergence for all these. The standard (multivariate) delta method then completes the proof. □

**Remark 1.** *From the underlying structure of the proof of Propositions 1 and 2, and the arguments (of e.g. Dahlhaus & Wefelmeyer (1996) or Dzhaparidze (1986)) used to show that the Whittle estimator has the same large-sample properties as the maximum likelihood estimator, it is clear that the conclusions of the two propositions stays true if we replace Whittle with full maximum likelihood estimation.*

The nonparametric estimator is by construction unbiased in the limit; an estimate for the risk is therefore easily obtained from the variance formula above. For the parametric candidate, we need in addition an unbiased estimate for the squared bias. Following Jullum & Hjort (2015) we start with $\widehat{b} = \widehat{\mu}_{\mathrm{pm}} - \widehat{\mu}_{\mathrm{np}}$ as an initial estimate for $b = \mu_0 - \mu_{\mathrm{true}}$. Since it follows from (3.2) that $\sqrt{n}(\widehat{b} - b) \to_d c^t J^{-1} U - X \sim \mathrm{N}(0, \kappa)$, where $\kappa = v_{\mathrm{pm}} + v_{\mathrm{np}} - 2v_c$, we have $\mathrm{E}\,\widehat{b}^2 \approx b^2 + \kappa/n + o(1/n)$. This leads to mse estimators of the form

$$\begin{aligned} \mathrm{FIC}_{\mathrm{np}} &= \widehat{\mathrm{mse}}_{\mathrm{np}} = \widehat{v}_{\mathrm{np}}/n, \\ \mathrm{FIC}_{\mathrm{pm}} &= \widehat{\mathrm{mse}}_{\mathrm{pm}} = \widehat{\mathrm{bsq}} + \widehat{v}_{\mathrm{pm}}/n = \max(0, \widehat{b}^2 - \widehat{\kappa}/n) + \widehat{v}_{\mathrm{pm}}/n. \end{aligned} \tag{3.4}$$

For the most general focus parameter formulation in (1.4), the variance and covariance estimators take the form

$$\begin{aligned} \widehat{v}_{\mathrm{np}} &= \nabla \widehat{H}_{\mathrm{np}} \{2\pi \int_{-\pi}^{\pi} h(\omega)^2 I_n(\omega)^2 \, \mathrm{d}\omega\} \nabla \widehat{H}_{\mathrm{np}}^{\mathrm{t}}, \text{ and} \\ \widehat{v}_{\mathrm{pm}} &= \nabla \widehat{H}_{\mathrm{pm}} \widehat{c}^{\,\mathrm{t}} J(I_n, f_{\widehat{\theta}_n})^{-1} K(I_n/\sqrt{2}, f_{\widehat{\theta}_n}) J(I_n, f_{\widehat{\theta}_n})^{-1} \widehat{c}\, \nabla \widehat{H}_{\mathrm{pm}}, \end{aligned}$$

where $\widehat{c} = (\nabla\mu(F_{\widehat{\theta}_n}; h_k), \ldots, \nabla\mu(F_{\widehat{\theta}_n}; h_k))^{\mathrm{t}}$, $\nabla \widehat{H}_{\mathrm{np}}$ and $\nabla \widehat{H}_{\mathrm{pm}}$ are the gradients of $H$ evaluated at respectively $(\mu(\widehat{G}_n; h_1), \ldots, \mu(\widehat{G}_n; h_k))$ and $(\mu(F_{\widehat{\theta}_n}; h_1), \ldots, \mu(F_{\widehat{\theta}_n}; h_k))$, and $J$ and $K$ are as defined in relation to (2.3) – using $I_n(w)^2/2$ as the canonical nonparametric unbiased estimator for $g(w)^2$. These are all consistent according to Taniguchi (1980); Deo & Chen (2000).

With FIC scores as above, representing clear-cut estimates of the risk of the nonparametric and parametric models' estimators of $\mu$, our model selection strategy turns out as follows: Compute the FIC score for each candidate model, rank them accordingly, and select the model and estimator associated with the smallest FIC score. The same $\mathrm{FIC}_{\mathrm{pm}}$

formula (with different estimates and quantities) is used for all of the possibly $m$ different parametric candidate models for simultaneous selection among the $m+1$ models. This is perfectly fine as the $\mathrm{FIC}_{\mathrm{pm}}$ formula does not depend on the other parametric models.

Although we have concentrated on focus functions $\mu(G;h)$ and $\mu(G;h,H)$ given by (1.3-1.4), our focused model selection strategy applies also to more general focus parameters, as long as joint limit distributions like (3.2) and (3.3) may be proven. In completely general terms, our results may be generalised to focus parameters of the form $\mu = T(G)$ for well-behaved functionals $T$ mapping the spectral distribution $G$ to a scalar value. The type of smoothness required for $T$ is in fact that the functional is so-called Hadamard differentiable at $G$ and $F_{\theta_0}$, see e.g. van der Vaart (2000, Theorem 20.8) for further details. This allows us, for instance, to handle focus parameters involving quantiles of the spectral distribution $G$. It is also possible to extend the arguments to other parametric estimation procedures, especially if they are derived as minimisers of the empirical analogue of $\mathrm{argmin}\{R(G,\theta)\}$ for $R$ the model specific part of possibly different divergence measure than in (2.2), see Dahlhaus & Wefelmeyer (1996) and Taniguchi (1980) for alternatives.

## 4. Models with trends

So far we have only considered stationary time series with mean zero. In real applications, this is often an unrealistic assumption to make. Even if the series is stationary, the underlying mean is rarely exactly zero; the common solution in such cases is to detrend the series. In time series modelling, detrending usually refers to the act of removing an estimated or deterministic trend from the observed series before the main analysis. This may be a complex function of time and covariates including seasonal effects, or be as simple as subtracting the arithmetic mean. A common approach is to work with the detrended series, which we will denote by $\widehat{y}_t$, and then analyse this series using models for stationary time series, without factoring in the extra estimation uncertainty involved in the detrending. This is often unproblematic, but even the innocent action of subtracting the mean may have unforeseen consequences (typically for the so-called second order properties). Hermansen & Hjort (2014b) shows that such a simple operation alter the underlying motivation and interpretation of the AIC for stationary Gaussian time series. Thus, special care is required for such an operation.

Suppose the observed series is generated by the model

$$Y_t = m(x_t,\beta) + \varepsilon_t, \tag{4.1}$$

where the $x_t$ are $p$-dimensional covariates, the $m$ is of known parametric structure, and $\{\varepsilon_t\}$ is a zero mean stationary Gaussian time series process with spectral distribution function $G$ and corresponding density $g$. Assume further that we are able to estimate $\beta$ by a suitable $\widehat{\beta}_n$ with reasonable precision. The question is then whether the results of

Section 3 are still valid also with detrended data, such that we may still use the same FIC formulae.

**Proposition 3.** *Suppose the spectral densities $g$ and $f_\theta$ and function $h$ satisfy the conditions of Proposition 1, and that the assumed trend $m$ and corresponding estimator $\widehat{\beta}$ for the unknown $\beta$ are such that $\sqrt{n}(\widehat{\beta}_n - \beta) = O_p(1)$. Assume further that in a neighbourhood of $\beta$ we have*

$$m(x, \widehat{\beta}_n) = m(x, \beta) + \nabla m(x, \beta)^{\mathrm{t}}(\widehat{\beta}_n - \beta) + r_n(x),$$

*with $\max_i |r_n(x_i)| = o_p(1/\sqrt{n})$ and $|\nabla m(x, \beta)|$ bounded in $x$. Then the conclusions of Proposition 1 are still true if we replace $y_t$ with the detrended $\widehat{y}_t = y_t - m(x_t, \widehat{\beta}_n)$.*

*Proof.* We will show that the result follows as a corollary from certain general results regarding limit behaviour of quadratic forms from Hermansen & Hjort (2014a, Section 3).

The argument is structured similarly to that of Proposition 1 and is built around a Cramér–Wold type of argument. Observe that if we replace $y_t$ with the detrended $\widehat{y}_t = y_t - m(x_t, \widehat{\beta}_n)$, we now have $\widehat{X}_n = (\underline{\widehat{y}}_n^{\mathrm{t}} \Sigma_n(h_0) \underline{\widehat{y}}_n - \mu_{\mathrm{true}})$ and similarly

$$\widehat{U}_n = -\tfrac{1}{2}\{\mathrm{Tr}(\Sigma_n(\nabla \Psi_{\theta_0})) - \underline{\widehat{y}}_n^{\mathrm{t}} \Sigma_n(\nabla \Psi_{\theta_0}/f_{\theta_0}) \underline{\widehat{y}}_n\},$$

where $\underline{\widehat{y}}_n = (\widehat{y}_1, \ldots, \widehat{y}_n)^{\mathrm{t}}$. Again, for any $a = (a_1, a_2)$ in $\mathbb{R}^2$, we now have

$$\widehat{\Lambda}_n = a_1\sqrt{n}\widehat{X}_n + a_2\sqrt{n}\widehat{U}_n = \underline{\widehat{y}}_n^{\mathrm{t}} \Sigma_n(a_1 h_0 + a_2 \nabla \Psi_{\theta_0}/f_{\theta_0}) \underline{\widehat{y}}_n/\sqrt{n} + \gamma_n,$$

with $\gamma_n$ as in the proof of Proposition 1. Then, according to Proposition 3.1 of Hermansen & Hjort (2014a),

$$\widehat{\Lambda}_n - \Lambda_n = o_p(n^{-1/2})$$

where $\Lambda_n = \underline{\varepsilon}_n^{\mathrm{t}} \Sigma_n(a_1 h_0 + a_2 \nabla \Psi_{\theta_0}/f_{\theta_0}) \underline{\varepsilon}_n/\sqrt{n} + \gamma_n$, where $\underline{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{t}}$ has elements corresponding to (4.1). Since the limit behaviour of $\Lambda_n$ is what defines the limit distribution in Proposition 1, the argument is essentially complete. □

The above proposition may also be extended to the focus parameter in (1.4), as handled in Proposition 2. Traditionally, the least squares estimator has been the canonical method for estimating $\beta$ in models of the form of (4.1). As an illustration, consider the linear regression model with dependent errors where $Y_t = x_t^{\mathrm{t}}\beta + \varepsilon_t$, for $p$-dimensional covariates $x_t$, and where $\{\varepsilon_t\}$ is a zero mean stationary Gaussian time series process with spectral density $g$. On matrix form this yields $\underline{y}_n = X\beta + \underline{\varepsilon}_n$, where $X$ is the related $n \times p$-dimensional design matrix. The ordinary least squares estimate for $\beta$ is then given by $\widehat{\beta}_n = (X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}\underline{y}_n$. Then, in order for $\widehat{\beta}_n$ to satisfy the conditions of Proposition 3, it is sufficient that $n\mathrm{Var}(\widehat{\beta}_n) = n(X^{\mathrm{t}}X)^{-1}X^{\mathrm{t}}\Sigma_n(g)X(X^{\mathrm{t}}X)^{-1} = o(1)$, which is clearly satisfied if $X^{\mathrm{t}}X/n \to_p Q_1$ and $X^{\mathrm{t}}\Sigma(g)X/n \to_p Q_2$, as $n$ approaches infinity, where $Q_1$ and $Q_2$ are both finite positive definite matrices. These are the standard assumptions

needed to ensure consistency of both standard and generalised least squares for models with correlated errors.

## 5. Average focused information criterion

We have so far concentrated on inference for a single focus parameter $\mu$. A natural generalisation of this is to consider several focus parameters joinly, say correlations of orders 1 to 5. The FIC machinery can easily be lifted to such a situation, involving a weighted average of FIC scores, the AFIC, with weights reflecting importance dictated by the statistician.

Suppose in general terms that estimands $\mu(u)$ are under consideration, for $u$ in some index set. For each of these we have the nonparametric $\widehat{\mu}_{\rm np}(u)$ and one or more parametric estimators $\widehat{\mu}_{\rm pm}(u)$. These typically have versions of Propositions 1 or 2, leading as per (3.1) to

$$\mathrm{mse}_{\rm np}(u) = 0^2 + v_{\rm np}(u) \quad \text{and} \quad \mathrm{mse}_{\rm pm}(u) = b(u)^2 + v_{\rm pm}(u),$$

with $b(u) = \mu_0(u) - \mu_{\rm true}(u)$. These mean squared errors can then be combined, via some suitable cumulative weight function $W(u)$, to

$$\mathrm{risk}_{\rm np} = \int v_{\rm np}(u)\,\mathrm{d}W(u) \quad \text{and} \quad \mathrm{risk}_{\rm pm} = \int \{b(u)^2 + v_{\rm pm}(u)\}\,\mathrm{d}W(u)$$

Here $\mathrm{d}W(\cdot)$ is meant to reflect the relative importance of the different $\mu(u)$, and should stem from the statistician's judgement and the actual context. Based on the data we may now form the following natural estimates of these risk quantities:

$$\begin{aligned} \mathrm{AFIC}_{\rm np} &= \int \widehat{v}_{\rm pm}(u)\,\mathrm{d}W(u), \\ \mathrm{AFIC}_{\rm pm} &= \int \big[\max\{\widehat{b}(u)^2 - \widehat{\kappa}(u)/n\} + \widehat{v}_{\rm pm}(u)\big]\,\mathrm{d}W(u). \end{aligned} \tag{5.1}$$

This operation also needs the covariances $v_c(u)$, as $\widehat{\kappa}(u)$ is to be constructed as the natural estimator of $\kappa(u) = v_{\rm pm}(u) + v_{\rm pm}(u) - 2v_c(u)$.

The AFIC scheme (5.1) can be used in a variety of circumstances. A typical application may involve assessing models for estimating a threshold probability $P\{Y_{n+1} \geq a\}$ over a set of many $a$, again with a weight function $w(a)$ indicating relative importance. Another attractive application is for the task of estimating correlations $\mathrm{corr}(h)$ for lags $h = 1, 2, 3, \ldots$, perhaps with a decreasing $w(h)$. The AFIC method may similarly be applied for comparing the popular autorcorrelation function, such as `acf` in the statistical software package `R` (R Core Team, 2015), with potentially more accurate parametric alternatives.

## 6. Performance

In the present section we will discuss some behavioural aspects of the derived FIC methodology. First we present some theoretical consequences of using our new FIC construction for model selection. Then we discuss some issues related to the more practical performance of this criterion, and illustrate some of these in a simulation study. The goal is not to conduct a broad simulation based investigation, but rather show the potential of having a criterion for selecting among parametric models and a nonparametric alternative in a simple proof of concept type of illustration.

6.1. **FIC under model conditions.** Although we have been working outside specific parametric model conditions when deriving the FIC (and AFIC) above, it is natural to ask how the criteria selects when a parametric model is indeed correct. Consider however first the case where a specific parametric candidate model is incorrect and have bias $b \neq 0$. From the structure of the FIC formulae in (3.4) and the consistency of the involved variance and covariance estimators, we see that $\text{FIC}_{\text{np}} = o_p(1)$, while $\text{FIC}_{\text{pm}} = O_p(1) + o_p(1) = O_p(1)$. I.e. the squared bias term dominates completely, and the probability that the FIC will select this particular parametric model will tend to 0 as $n \to \infty$. If all the parametric candidate models are biased in this sense, then the FIC will eventually prefer the nonparametric model when the sample size increases.

Going more into detail, it is seen from the FIC formulae in (3.4) that the FIC prefers a specific parametric model over the nonparametric whenever

$$\max(\widehat{b}^2 - \widehat{\kappa}/n, 0) + n^{-1}\widehat{v}_{\text{pm}} \leq n^{-1}\widehat{v}_{\text{np}}.$$

Whenever $\widehat{v}_{\text{np}} \geq \widehat{v}_{\text{pm}}$, this is seen to be equivalent to

$$Z_n \leq 2,$$

where $Z_n = (n\widehat{b}^2)/(\widehat{v}_{\text{np}} - \widehat{v}_c)$.

It turns out that under model conditions, we have $v_c = v_{\text{pm}}$. This is rather straightforward to see by investigating the forms of $v_c$ and $v_{\text{pm}}$ involved in Proposition 2, in addition to the forms of $K_0$ and $J_0$. Inserting $g = f_{\theta_0}$ in these formulae reveals that $K_0 = J_0$, $\nabla H_{\text{np}} = \nabla H_{\text{pm}}$ and $c = d$ and thereby $v_c = v_{\text{pm}}$. Now, due to the consistency, we have $\widehat{v}_{\text{np}} - \widehat{v}_c \to_p v_{\text{np}} - v_{\text{pm}}$. Further, the limit distribution result of $\sqrt{n}(\widehat{b} - b)$ given above (3.4) ensures that $Z_n \to_d \chi^2_1$, with $\chi^2_1$ a chi-squared distributed variable with one degree of freedom. That is, the limiting probability that the parametric model will be selected over the nonparametric when it is indeed true is $P\{Z_n \leq 2\} \to P\{\chi^2_1 \leq 2\} \approx 0.843$. Thus, if one of the parametric candidate models is correct, and the others have biases $b \neq 0$, then, for sufficiently large samples, the first parametric model and estimator will be selected

with a probability tending to 84.3%, while the nonparametric will be selected in the other 15.7% proportion.

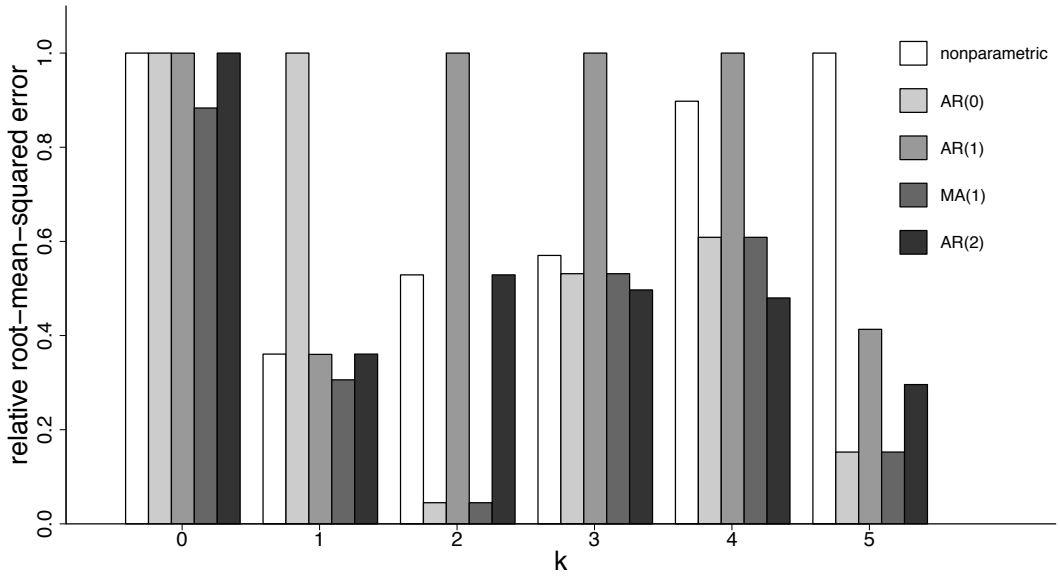


FIGURE 6.1. Relative root-mse for each candidate model fitted to the six focus parameters $\mu_k = C(k)$, for $k = 0, \ldots, 5$. The root-mse is computed based on 5000 simulated AR(2) series of length $n = 100$, with $\sigma = 1.0$ and $\rho = (0.7, -0.6)$, For ease of comparison we have scaled the root-mse to the unit interval.

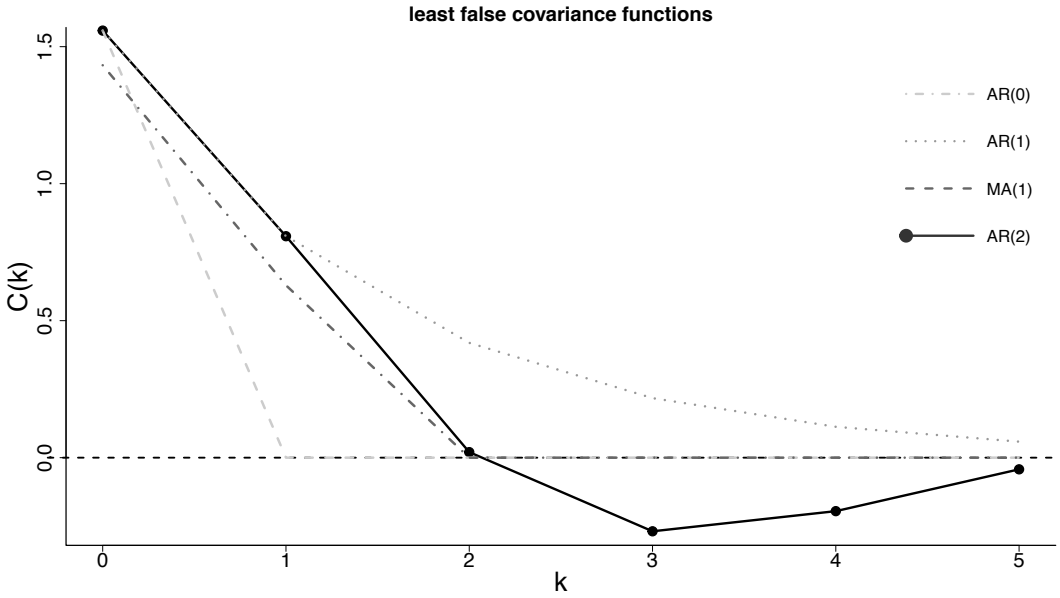


FIGURE 6.2. The five least false covariance functions under the assumption that the true model is an autoregressive model specified by the parameters $\sigma = 1.0$ and $\rho = (0.7, -0.6)$.

6.2. **FIC in practice.** Figure 6.1 shows the relative root-mse for estimating the focus parameter

$$\mu_k = \mu(G; h_k) = \int_{-\pi}^{\pi} \cos(\omega k) g(\omega)\, \mathrm{d}\omega = C_g(k), \qquad \text{for } k = 1, \ldots, 5, \tag{6.1}$$

based on the following five candidates models: the independence model (autoregressive of order zero); the autoregressive of orders one and two; the moving average of order one; and finally the nonparametric one, where nothing more is assumed than saying that the series is stationary with a finite variance. The true model is an autoregressive model specified by the parameters $\rho = (0.7, -0.6)$ and $\sigma = 1.0$. This means that all but two, the autoregressive model of order two and the nonparametric model, are misspecified. The corresponding least false covariance estimates are plotted in Figure 6.2. In the simulation study, we have used $B = 5000$ repetitions of length $n = 100$ to compute the actual relative root-mse values for each candidate. Note that since we have included the true model among our candidates, nonparametric estimation is never the optimal choice; it is however often close and it is the second best choice for lags 1 and 3. For lags 2 and 5, where the true values are close to zero, the simpler models, like AR(0) and MA(1), are highly successful, achieving reasonably low bias and also low variance.

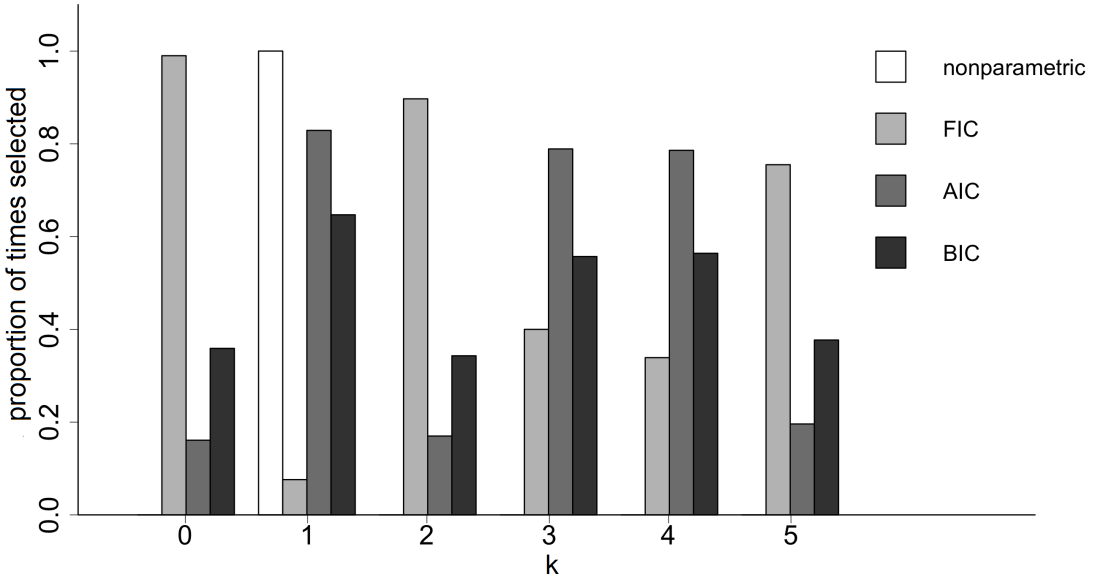


FIGURE 6.3. The proportion for which the different criteria selects the model with the theoretical lowest root-mean-squared error. The model-selectors are always nonparametric, FIC, AIC and BIC. The results are based on 5000 simulated series.

In Figure 6.3 and 6.4 we further investigate the performance of the FIC. Here, we compare our FIC machinery with three other model selection strategies, (i) to always use

the nonparametric model, (ii) select the best parametric model according to the AIC and (iii) the parametric model selected by the BIC. Note that the AIC and BIC tools do not work for the nonparametric model, since there is no likelihood function. In Figure 6.3 we have counted how many times each criterion selects the model that obtains the smallest root-mse value, for each focus parameter $\mu_k$ as defined in (6.1). Figure 6.4 contains the corresponding attained root-mse values. Note that for lag 1 the theoretical root-mse for the autoregressive models are, for all practical purposes, equal to that obtained by the nonparametric model. In all other cases, the nonparametric model has a root-mse larger than the optimal model.

In this illustration, the FIC behaves more or less as intended, by selecting (on average) the models that produces the smallest risk. The amount of evidence is by no means conclusive, but it indicates that the FIC machinery has a real potential.

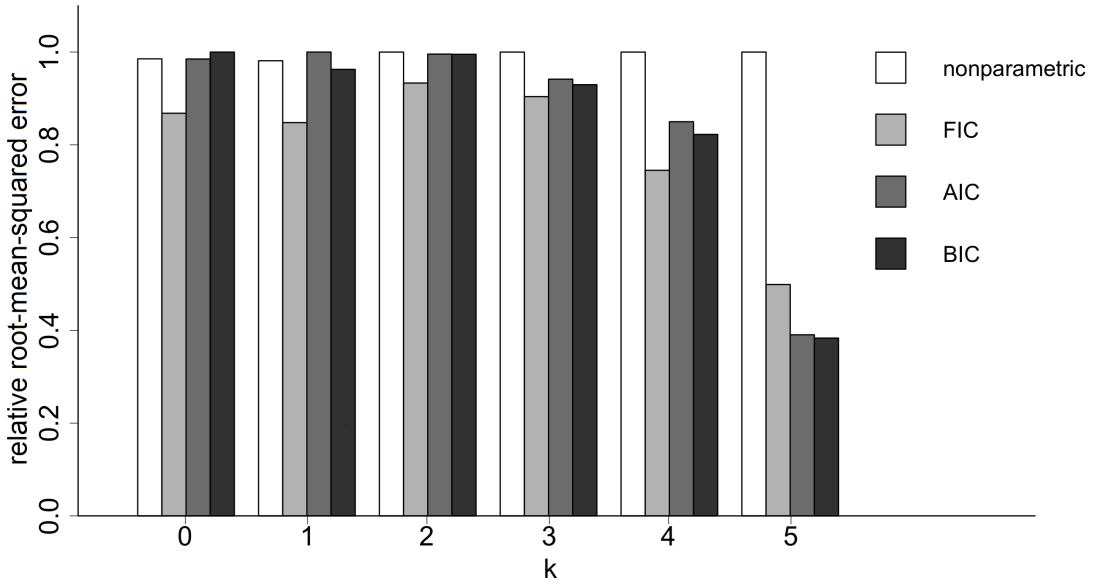


FIGURE 6.4. The relative root-mean-squared (computed in the same simulations) for the models selected by FIC, AIC and BIC, and by always using the nonparametric model.

## 7. Concluding remarks

Here we offer a list of conclucing comments, some pointing to further relevant research.

7.1. **Model averaging.** The FIC scores may also be used to combine the most promising estimators into a model averaged estimator, say $\widehat{\mu}^* = \sum_j c(M_j)\widehat{\mu}_j$, with $c(M_j)$ given higher values for models $M_j$ with higher FIC scores; as discussed in Hjort & Claeskens (2003).

7.2. **The conditional FIC.** For time series processes, several interesting and important focus parameters are naturally related to predictions, are sample size dependent or otherwise formulated conditional on past observations. The classical example is $k$-step ahead predictions. A class of such estimands could take the form

$$\mu(\alpha, \gamma, y_1, \ldots, y_m) = P\{Y_{n+1} > \alpha \text{ and } Y_{n+2} > \gamma \,|\, y_1, \ldots, y_m\}$$

for a suitable choice of $\alpha$. The dependency on previous data requires a new and extended modelling framework, which in Hermansen & Hjort (2015, Sections 5 & 6) led to generalisations and also motivated a conditional focused information criterion (cFIC). In completing the FIC-framework for selecting among parametric and nonparametric time series models, such considerations should also be taken into account.

7.3. **Linear time series processes.** Building on Walker (1964); Hannan (1973); Brillinger (1975), the main results of Section 3 can be extended to more general types of time series processes, like the generalised linear processes (cf. Priestley (1981)); also without the assumption of Gaussian innovation terms.

7.4. **Trends and covariates.** In the presented work, our focus was on the dependency structure only. However, the methods and results of our paper may be generalised to select simultaneously among models with different trends and dependency structures, like $Y_t = m(x_t, \beta) + \varepsilon_t$, with $\varepsilon_t$ a stationary Gaussian time process. These issues, leading to a larger repertoire of FIC formulae, will be returned to in later work. Since it is generally hard to estimate both the trend and dependency structure using a full nonparametric framework, the two main challenges is to extend the existing work to handle the case with various parametric candidates for the trend $m(x_t, \beta)$ and both parametric models and a nonparametric candidate for the dependency, i.e. the spectral distribution (since we are working under the Gaussian assumption). Alternatively, we may assume that the $\varepsilon_t$ belongs to an appropriate width family of parametric stationary time series processes, such as the autoregressive AR, the moving average MA or the mixture ARMA (cf. Brockwell & Davis (1991)) and instead compare a nonparametric method for estimating the trend part of the model, perhaps extending this to functions of the type $m(t, x_i, \beta)$, against a class of parametric alternatives.

7.5. **The local large-sample framework.** As mentioned in the introduction, Hermansen & Hjort (2015) derives FIC for selecting among parametric time series models using a local asymptotics framework. The parametric candidate models then have spectral densities belonging to a parametric family $f(\cdot; \theta, \gamma)$, with a $p$-dimensional protected $\theta$ and a $q$-dimensional open $\gamma$. This constitutes a set of $2^q$ potential parametric candidate models. The full (or wide) model is represented by the spectral density $f(\cdot; \theta, \gamma)$. At the other end of the spectrum, the narrow model corresponds to fixating $\gamma = \gamma_0$, a known

value, with the resulting $f(\cdot;\theta) = f(\cdot;\theta,\gamma_0)$. The local misspecification framework then assumes that the true spectral density takes the form $f(\cdot;\theta_0,\gamma_0+\delta/\sqrt{n})$, for some unknown $q$-dimensional $\delta$ describing the distance to the wide model. This framework causes variances and squared biases to become of the same order of magnitude $O(1/n)$. Those lead to approximation formulae for the mean squared error and FIC formulae for nested parametric models, which are different from those obtained in this paper.

The introduction of the 'asymptotically correct' nonparametric model of the present paper allowed us to derive FIC formulae even when sidestepping the above local misspecification assumption. An alternative approach is to retain the local asymptotics framework and work with spectral densities of the type $f_r(\omega) = f_{\theta_0}(\omega) + r(\omega)/\sqrt{n}$, where $f_{\theta_0}$ is a standard type of parametric model. Such structures have already been worked with in Dzhaparidze (1986), making the extension potentially less cumbersome. This will not be dealt with here, however.

## References


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

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

Claeskens, G. & Hjort, N. L. (2003). The focused information criterion [with discussion and a rejoinder]. *Journal of the American Statistical Association* **98**, 900–916.

Claeskens, G. & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: 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. & Wefelmeyer, W. (1996). Asymptotically optimal estimation in misspecified time series models. *Annals of Statistics* **24**, 952–973.

Deo, R. S. & Chen, W. W. (2000). On the integral of the squared periodogram. *Stochastic processes and their applications* **85**, 159–176.

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

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

Hannan, E. J. (1973). The asymptotic theory of linear time-series models. *Journal of Applied Probability* **10**, 130–145.

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

Hermansen, G. & Hjort, N. L. (2014b). 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.

Hermansen, G. & Hjort, N. L. (2015). Focused information criteria for time series. *Submitted for publication* .

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

Jullum, M. & Hjort, N. L. (2015). Parametric or nonparametric: The FIC approach. *Submitted for publication* .

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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. (2000). *Asymptotic Statistics*. Cambridge University Press.

Walker, A. M. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series. *Journal of the Australian Mathematical Society* **4**, 363–384.

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