

FOCUSED INFORMATION CRITERIA FOR TIME SERIES

GUDMUND HORN HERMANSEN^{1,2} AND NILS LID HJORT²

ABSTRACT. We extend the model selection methodology of the focused information criterion (FIC) to large classes of time series models. Compared to other classical information criteria, the FIC allows the precise intention of the analysis to be taken into account when selecting the model. The goal is not to find the one ‘correct’ model which does best in some broad overall sense; the FIC aims instead at finding the model which is best suited for answering focused questions, one focus at a time. For time series data, 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 leads to a proper generalisation of the original criterion and motivates a conditional focused information criterion (cFIC). In addition to the classical stationary Gaussian time series models, we develop a FIC as a model selection strategy for the broader family of locally stationary processes. This general class of models, which in particular includes those built up via smooth trend functions with errors of the stationary time series kind, provides a unified framework for interpretation and focused model selection. Our methodology is illustrated for real data problems.

1. INTRODUCTION AND SUMMARY

There is a long tradition of selecting models for time series data via so-called information criteria. These carry a considerable appeal by typically being simple in structure and use, yielding scores which then can be used to rank candidate models from best to worst in accordance to some predefined measure of discrepancy. There are a number of such criteria, derived and justified from different points of departure. Among the earlier and more popular ones are the final prediction error (FPE; Akaike (1969)), Akaike’s information criterion (AIC; Akaike (1973)) and the Bayesian information criterion (BIC; Schwarz (1978)); see Claeskens & Hjort (2008) for a general overview. For a more time series oriented discussion, see Akaike (1982) or the final chapters of Linhart & Zucchini (1986). Moreover, a comprehensive survey of the autoregressive models, with focus on consistency, i.e. the ability to find the correct order, or true model, can be found in McQuarrie & Tsai (1998).

In the present paper we develop focused information criteria (FIC) for a large class of time series models. The FIC was originally developed in Hjort & Claeskens (2003) and Claeskens & Hjort (2003) for models without dependency, e.g. classical models for i.i.d. observations, different types of regression models and the Cox hazard regression model, see Hjort & Claeskens (2006), Hjort (2008) and Claeskens

Key words and phrases. Autoregressive processes, focused inference, locally stationary processes, model selection, stationary time series, time series with trend,

¹Norwegian Computing Centre.

²Department of Mathematics, University of Oslo.

& Hjort (2008). 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.

The model preferred by the FIC will depend on a particular focus parameter of interest μ , or focus function $\mu(\cdot)$. For stationary time series models the canonical example would be h -step ahead predictions, i.e. the task of finding the best model for predicting future outcomes. This is already well studied and for the autoregressive processes, with focus on one-step ahead predictions, this leads to the FPE criterion of Akaike (1969); see also Linhart & Göttingen (1985) and Hermansen & Hjort (2014b) for a generalisation to more universal and also potentially misspecified time series models. In general, a wide variety of other focused questions emerges naturally for time series processes, for example estimation of conditional probabilities for thresholds being exceeded, determination of confidence bounds, the influence of certain covariates, etc., or more direct features like certain covariance lags, properties of the spectral density for frequencies close to zero, and so on.

In the first half of our paper, Sections 3–7, we focus mainly on standard stationary Gaussian time series. These will be seen to be a special case of the locally stationary models of Dahlhaus (1997) discussed in Section 9. The class of locally stationary processes provides a good framework for interpretation and inference, serving also as a basis for a unified theory. Moreover, it includes the important class of stationary models with trends, studied in Section 8, with

$$Y_t = m(t/n, \beta) + \epsilon_t, \quad \text{for } t = 1, \dots, n, \quad (1.1)$$

where $m(u, \beta)$ is a smooth function of u , defined on the unit interval, and ϵ_t is a stationary time series process with zero mean, defined by a parametric covariance function $C_\theta(h)$, for all lags $h \geq 0$. The machinery we develop will be able to pick the best models of this type, selecting among both trends and covariance functions, for a given focus question.

The potential candidate models are assumed to be nested between some specified ‘narrow’ and ‘wide’ models. The narrow model is typically some standard or baseline model, with the wide model extending the narrow one by including additional parameters. Let S represent an arbitrary submodel, defined by including some and excluding others of these extra parameters, and define $\hat{\mu}_S$ to be the corresponding submodel estimate for μ_{true} , i.e. the focus function evaluated in the true model. The basic strategy, which will lead to our FIC, is to aim for the candidate model S that attains the smallest mean squared error. This is then solved in a large-sample framework, where the FIC formula will emerge as an unbiased, large-sample approximation for the respective mean squared errors. The model with the smallest FIC score is hence the model that obtains the smallest estimated mean squared

error. Note that for such comparisons of submodel estimators to be relevant, the focus μ worked with has to be meaningful across all competing models.

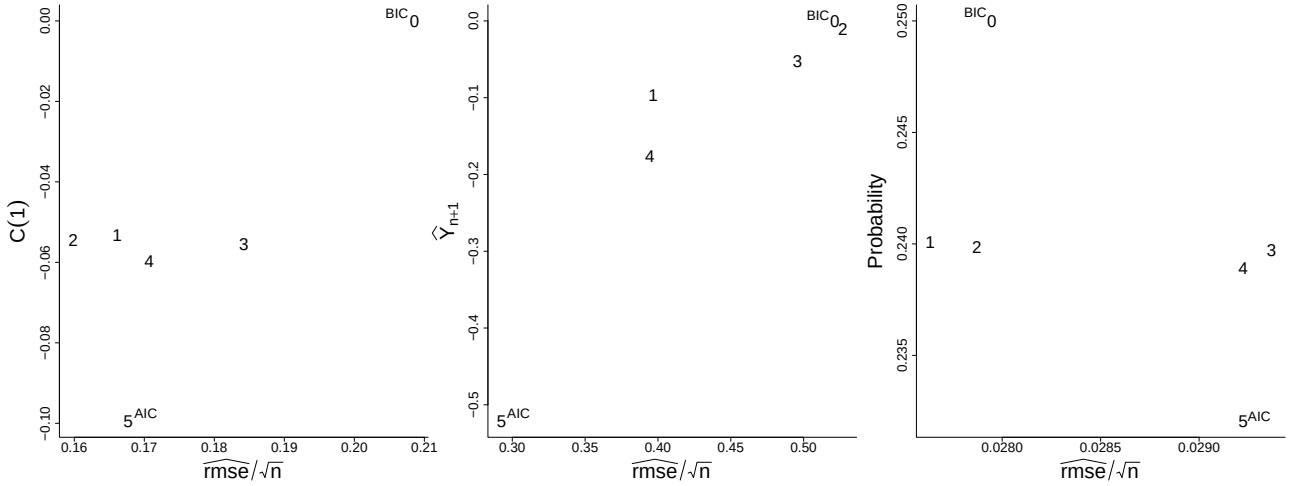


FIGURE 1.1. The different submodel estimators for the three foci of (1.2) (left–right) are plotted against the estimated precision, i.e. $\widehat{\text{mse}}(S) = \text{FIC}(S)$ for each submodel S . The estimates are based on 100 simulated realisations from an autoregressive model defined by the parameters $\sigma = 1$ and $\rho = (-0.04, 0.21, -0.22, -0.10, 0.33)$, see Brockwell & Davis (1991) for technical details regarding the model assumptions. The numbers represent the different model orders, where ‘0’ corresponds to the simple independence model.

A consequence of a focused model selection strategy is that different models can be preferred for different foci. This is in sharp contrast to 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. In general, we anticipate the FIC to prefer simple models when these suffice and only more complex models when such are required by the focused question and the given data (also, more sophisticated models would be selected with more data). This effect is illustrated in Figure 1.1 with a simulated dataset from an autoregressive model of order five. The candidates are the autoregressive models of order zero to five, where order zero corresponds to the simple independence model. Here we consider three different foci:

$$(i) \mu_1(1) = C(1), \quad (ii) \mu_2(Y_{n+1}) = Y_{n+1} \quad \text{and} \quad (iii) \mu_3(0) = \Pr\{Y_{n+1} < 0, Y_{n+2} < 0\}, \quad (1.2)$$

where $C(1) = \text{Cov}(Y_i, Y_{i+1})$ and (ii) represent the one-step ahead prediction. The methodology required for handling these is developed in Section 3–6. Here the AIC, which does not differentiate between the different foci, turns out to prefer the widest model, the autoregressive model of order five, whereas the BIC chooses the smallest model with independence. As seen via the FIC plots, the FIC selects different best models for the three different foci.

Our examples and illustrations often use these autoregressive models, partly for reasons of convenience. The methodology developed and our computer scripts (for the R software) are very general and allow for a variety of time series models, including the so-called moving average (MA) and the mixture with the autoregressive (ARMA) models; see for example Brockwell & Davis (1991) for details on these.

In Claeskens & Hjort (2003) only models for independent realisations were considered. Therefore, in order to develop FIC methods for the time series framework, parts of the original theory must be redeveloped to fully accommodate the time series processes. In addition, several interesting foci are more natural if formulated as conditional on the already observed series, as when predicting annual global mean temperature for the coming years, given the recently observed past. As another illustration, assume we observe a speedskater run an Olympic 10k race, and wish to assess, two laps before the finish, that he will set a new world record; we will return to these issues in Example 8.1 below. Handling such questions is not covered by the existing theory. In a properly extended framework, we derive a generalisation of the FIC which is capable of using such ‘data-dependent’ focus functions and we will also introduce the conditional focused information criterion (cFIC), both of which take the dependency on data in the focused questions properly into account, see Section 6 below.

Although the FIC has received attention and gained decent popularity, it has not really been embraced in the time series communities. To our knowledge there are only a few attempts that use this kind of methodology in time series analysis. In Claeskens et al. (2007) a version of the FIC was derived for autoregressive models, which can also include a polynomial trend component. They considered only types of h -step ahead predictions as their foci, however. A related study is given by Rohan & Ramanathan (2011), where these ideas are extended to the class of ARMA models. In our view, by considering only h -step ahead predictions, these studies do not take full advantage of the underlying potential and flexibility of the FIC idea. The ability to easily change and consider several foci is one of its main strengths, and should be exploited further.

In the present work we extend the original framework developed in Hjort & Claeskens (2003) and Claeskens & Hjort (2003), and derive generalisations that are valid for a large class of time series processes. We start with a motivating illustration in Section 2, and in Section 3 we work with the standard mean-zero stationary Gaussian time series processes. Tools for model averaging and in particular FIC averaging (AFIC) are developed in Section 4. The class of foci is then extended in Sections 5 and 6, where we first introduce sample size dependent functions and then extend these further to so-called data-dependent focus functions. This motivates, in turn, a second generalisation and also a conditional version of the original FIC construction. In Section 7 an alternative derivation is given, which shows that the FIC based on the Whittle approximation to the Gaussian log-likelihood function, cf. Whittle (1953), works just as well as with the full likelihood version. In addition to being

interesting in its own right, this provides an alternative, simpler and numerically more robust method for estimating the FIC scores. Several examples and illustrations are given throughout the paper, most of which are based on real data sets. In all of these the calculations are done using the full Gaussian log-likelihood. In the allude simulation studies, however, both the full log-likelihood and the Whittle approximation are tested, resulting in similar conclusions. In Section 8 we extend the FIC arguments to include stationary models with a smooth trend component. The most general results of our paper are then derived within the framework of locally stationary processes, in Section 9. Finally, some concluding remarks are offered in Section 10, some of them pointing to further research.

2. THE NUMBER OF SKIING DAYS AT BJØRNHOLT

As an introductory illustration, we consider the number of skiing days in a winter season, i.e. the number of days with at least 25 cm snow, at the particular location of Bjørnholt in Oslo's skiing and recreation area Nordmarka. 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. In Figure 2.1 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. The models we employ for this data set are all autoregressive models of order less than or equal to three, with and without a linear trend, giving a total of $2^4 = 16$ candidate models.

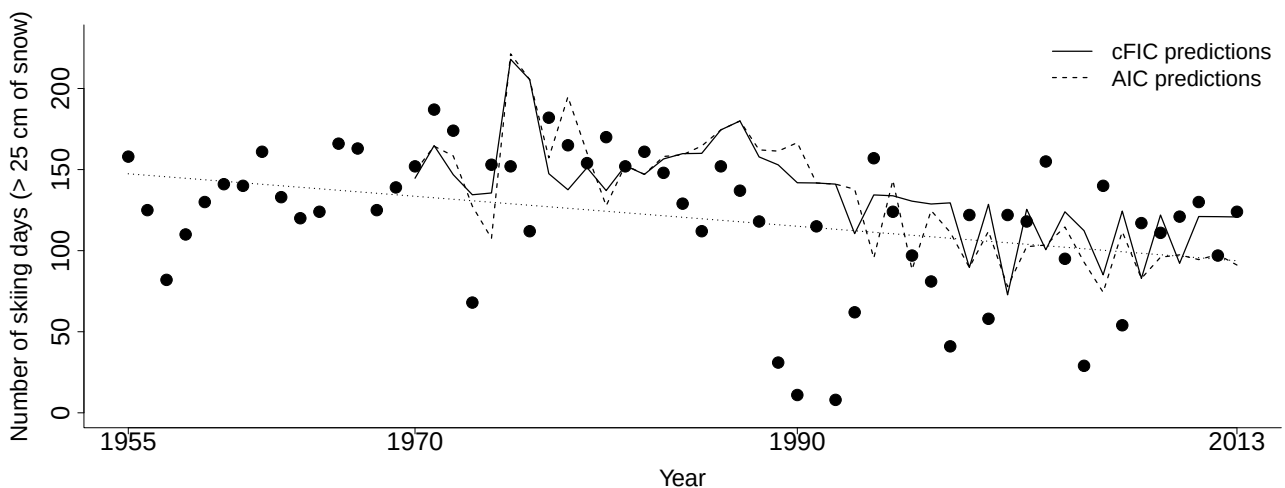


FIGURE 2.1. The number of skiing days for the winter seasons 1954-55 to 2012-13 at Bjørnholt, a location in the countryside just outside Oslo, are plotted with sequential one year a head predictions from the models preferred by the cFIC, introduced in Section 6 below, and AIC. The estimated root-mean-squared errors for the predictions are given by $\widehat{\text{rmse}}_{\text{cFIC}} = 0.845$, $\widehat{\text{rmse}}_{\text{AIC}} = 0.856$, and $\widehat{\text{rmse}}_{\text{BIC}} = 0.840$. As an additional interesting observation we note that the FIC prefers the simplest models; on average the FIC uses models with 3.2 parameters, the AIC likes models with 4 and the BIC 3.5 parameters.

The sequential one year ahead forecasts from the 1970-71 season and onwards, based on the current history, are computed and compared for models chosen for each season by the cFIC (solid line) and the classical AIC (dashed line) in Figure 2.1.

The AIC selects models according to the global performance, which seems to produce forecasts that are quite sensitive to the large scale behaviour of data. This is particularly evident for the last five seasons from 2008-09 to 2013-14. It appears that cFIC is less sensitive to this and select models that rely more on recent behaviour, which on average results in more accurate predictions. The effect is also illustrated in Figure 2.2 (left panel), where the submodel forecasts for the season 2014-15 is plotted against the root-cFIC values; the more to the left in the cFIC plot, the better is the estimate.

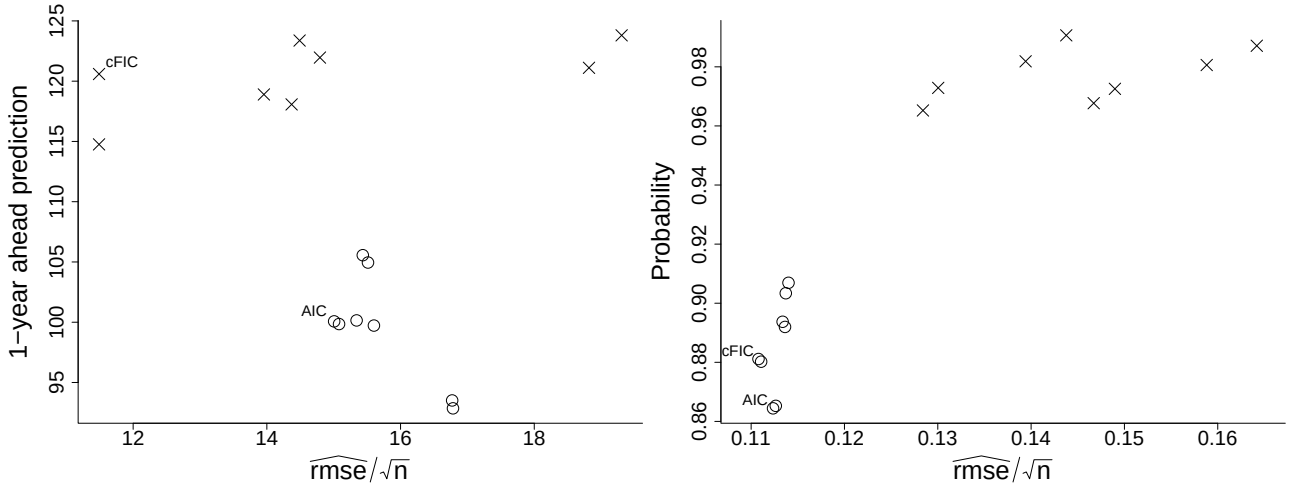


FIGURE 2.2. A continuation of the illustration in Figure 2.1. The figure shows the different submodel estimates plotted against the precision, which is estimated using the cFIC machinery. The foci are the number of skiing days for the winter of 2014-2015 (left panel) and the probability of having at least one season, within the next three years, with a minimum of 100 skiing days (right panel). The models with constant trend are represented with an 'x' and the models with linear trend are marked with an 'o'. The AIC prefers the autoregressive model with a linear trend, the cFIC prefers the independent model with constant trend (left panel) and the independent model with linear trend (right panel).

For simple and local phenomena (e.g. in time), such as one-step ahead predictions, low order models will often produce accurate or reasonable forecasts, but are not necessarily expected to capture the underlying (large scale) dynamics in data. This can be seen in Figure 2.1 and 2.2 (left panel), with the absence of a linearly decreasing trend in several of the best models, however. By preferring simpler models, the cFIC selects in this case models that are on average more suitable for short-term predictions.

The hope is that focusing the model selection procedure will result in models that capture more of the relevant features, with respect to the chosen focus, which may help to centre the attention on what is of most importance and hopefully assist in revealing more of the underlying mechanisms

and dynamics in real world problems. This is indicated in Figure 2.2 (right), where the change from one-step ahead predictions to the more involved problem of estimating the probability of having at least one winter with a minimum of 100 skiing days within the following three seasons, yields the linear model with independent errors as the best model.

3. THE FIC FOR STATIONARY TIME SERIES

In this section we introduce the focused information criterion (FIC) for stationary Gaussian time series processes. The derivation we give will also work as a stepping stone, introducing the main ideas and the basic notation for the following sections, however, we first need to establish some preliminary results and notation.

3.1. Background material and regularity conditions. Let Y_t be a zero mean stationary Gaussian time series process. The dependency structure, which in this case determines the entire model, is completely specified by the covariance function $C(h) = \text{Cov}(Y_{t+h}, Y_t)$, defined for all lags $h = 0, 1, 2, \dots$. The covariance function is nonnegative-definite, which means that for all $n \geq 1$ the following holds: $a^t \Sigma_n a \geq 0$ for all vectors $a \in \mathbb{R}^n$, where Σ_n is the covariance matrix with elements $C(|i - j|)$, for $i, j = 1, \dots, n$. By Wold's theorem, see e.g. Priestley (1981), $C(h)$ is such a covariance function if and only if

$$C(h) = \int_{-\pi}^{\pi} e^{i\omega h} dG(\omega) \text{ for } h = 0, 1, 2, \dots,$$

where the spectral distribution function G needs to be nondecreasing and right-continuous, with $G(-\pi) = 0$ and $G(\pi)$ finite. For real valued time series, the formula above simplifies and since we shall assume that G has a symmetric spectral density f , it follows that $C(h) = 2 \int_0^\pi \cos(\omega h) f(\omega) d\omega$. Note that the spectral density can be obtained as the Fourier transform of the covariance function, i.e.

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} C(h) e^{-i\omega h} = \frac{\sigma^2}{2\pi} + \frac{1}{\pi} \sum_{h=1}^{\infty} C(h) \cos(\omega h), \quad \text{for } 0 \leq \omega \leq \pi, \quad (3.1)$$

see e.g. Brillinger (1975), Priestley (1981) or Dzhaparidze (1986). As an illustration, the spectral density of the autoregressive model of order k is

$$f_{\sigma, \rho_1, \dots, \rho_k}(\omega) = \frac{\sigma^2}{2\pi} \left| \sum_{j=1}^k \rho_j e^{-i\omega j} \right|^{-2} = \frac{\sigma^2}{2\pi} \left[\left(1 - \sum_{j=1}^k \rho_j \cos(\omega j) \right)^2 + \left(\sum_{j=1}^k \rho_j \sin(\omega j) \right)^2 \right]^{-1},$$

see Brockwell & Davis (1991, Theorem 4.4.2). Note that some technical conditions are needed to ensure that the corresponding process is stationary and the reader is referred to Brockwell & Davis (1991, Ch. 3) for a complete account. If $k = 1$ matters simplify and

$$f_{\sigma, \rho}(\omega) = \frac{\sigma^2}{2\pi(1 - 2\rho \cos \omega + \rho^2)}, \quad (3.2)$$

which with positive σ and $|\rho| < 1$ corresponds to the stationary process with covariance function $C(h) = \sigma^2 \rho^h / (1 - \rho^2)$.

In our framework, the candidate models are nested between a narrow or baseline model and a wide model, which extends the narrow one by including additional parameters. The widest model has spectral density of the form $f_{\text{wide}} = f_{\theta, \gamma}$, for (θ, γ) in a natural parameter space. The narrow model corresponds to $\gamma = \gamma_0$, where γ_0 is a known vector determined by context. For standard autoregressive models, γ_0 will typically be a vector of zeros that cancel higher order dependencies from the narrow model. 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 $\{1 \dots, q\}$, ranging from the model, with $S = \emptyset$, to the full wide model where $S = \{1 \dots, q\}$. In practice we include only those submodels we regard as sufficiently interesting or plausible, however.

The observed series $y_1, \dots, y_n$ is assumed to be a realisation from a model with spectral density

$$f_{\text{true}} = f_{\theta_0, \gamma_0 + \delta/\sqrt{n}}, \quad (3.3)$$

where (θ_0, γ_0) is an inner point in a compact parameter space $\Theta \times \Gamma$. The idea behind studying 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 Hjort & Claeskens (2003) and Claeskens & Hjort (2008, Ch. 5 & 6) for a more complete discussion and additional comments.

A broader class of focus functions will be introduced in Sections 5 and 6, but for now assume that the focus $\mu = \mu(\theta, \gamma)$ depends only on the underlying model through the model parameters and that the derivatives with respect to (θ, γ) exist and are continuous in a neighbourhood of (θ_0, γ_0) . The unknown model parameters are estimated by maximum likelihood and the different submodel estimates of the focus parameter are then given by

$$\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0, S^c}), \quad (3.4)$$

where $S \subseteq \{1 \dots, q\}$ and S^c is the complement of S , i.e. the indexes of the parameters we do not estimate for that model. The idea leading to the FIC is the derivation of a large-sample approximation for the mean squared error of $\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}})$ 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.

3.2. The limit experiment and the FIC. Let $y_1, \dots, y_n$ be a realisation from the model defined in (3.3) with log-likelihood function

$$\ell_n(\theta, \gamma) = -\frac{1}{2}[n \log(2\pi) + \log |\Sigma_n(f_{\theta, \gamma})| + \mathbf{y}_n^t \Sigma_n(f_{\theta, \gamma})^{-1} \mathbf{y}_n], \quad (3.5)$$

where $\Sigma_n(f_{\theta, \gamma})$ is the covariance matrix with elements $2 \int_0^\pi \cos(\omega|i-j|) f_{\theta, \gamma}(\omega) d\omega$ for $i, j = 1, \dots, n$, and $\mathbf{y}_n^t = (y_1, \dots, y_n)$. Let $\nabla = \nabla_{\theta, \gamma} = \partial/\partial(\theta, \gamma)$ and define

$$Z_n = \sqrt{n}(U_n, V_n) = \frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0, \gamma_0) \quad (3.6)$$

the vector of scaled score functions evaluated at the null-point (θ_0, γ_0) . Note that the absence of a subscript, as in $\nabla \ell_n(\theta_0, \gamma_0)$, will in general indicate that the operation is with respect to all available model parameters.

The limit distribution of Z_n under the true model (3.3) is a key component in the derivation of the FIC. We shall need some preliminary results. Under the conditions of Assumption 3.1 below it follows from Lemma 11.1 in the Appendix that

$$J_{\text{wide}} = \lim_{n \rightarrow \infty} \text{Var}_{P_0} Z_n = \lim_{n \rightarrow \infty} \frac{1}{2n} \text{tr}\{(\Sigma_0^{-1}[\nabla \Sigma_0])^2\} = \frac{1}{4\pi} \int_{-\pi}^{\pi} \nabla \Psi_0(\omega) \nabla \Psi_0(\omega)^{\dagger} d\omega, \quad (3.7)$$

where $\Sigma_0 = \Sigma_n(f_0)$, $\Psi_0(\omega) = \log f_0(\omega)$, $f_0 = f_{\theta_0, \gamma_0}$ and $P_0 = P_{0,n}$ represents the distribution associated with a Gaussian vector of length n from the model with spectral density f_0 . In general, a zero in the subscript indicates evaluation at (θ_0, γ_0) . Finally, we define

$$J_{\text{wide}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}, \text{ with inverse } J_{\text{wide}}^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}, \quad (3.8)$$

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

Assumption 3.1. In a neighbourhood of (θ_0, γ_0) , the spectral density $f_{\theta, \gamma}$ is continuous and bounded away from zero and infinity and also two times differentiable in (θ, γ) , with derivatives that are differentiable in ω with uniformly continuous derivatives.

The following proposition is a key component in the derivation of the FIC, and establishes the limit distribution of the normalised score functions under the true model (3.3). Since the stationary models of this section can be seen as a special case of the processes discussed in Section 9, the result below can be seen to be a special case of Theorem 9.3 and is therefore given without a separate proof. This is also the reason the conditions in Assumption 3.1 are somewhat stronger than what is commonly seen in classical time series literature, see Remark 7.2 for an alternative argument under weaker and more standard conditions.

Proposition 3.2. *Let $y_1, \dots, y_n$ be realisations from the model (3.3) and let $P_{\delta} = P_{\delta, n}$ be the associated probability measure. Then if $f_{\theta, \gamma}$ satisfies the conditions of Assumption 3.1*

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

with J_{wide} as defined in (3.7). Also, $\widehat{J}_n(\bar{\theta}_n, \bar{\gamma}_n) = -\nabla^2 \ell_n(\bar{\theta}_n, \bar{\gamma}_n)/n \rightarrow_{P_{\delta}} J_{\text{wide}}$, provided $(\bar{\theta}_n, \bar{\gamma}_n) \rightarrow_{P_{\delta}} (\theta_0, \gamma_0)$.

With the above results in place we are almost ready to derive the limiting mean squared error for the normalised submodels estimators from (3.4). A bit more notation is needed first. Define $Q = J^{11} = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$ and for each submodel $S \subset \{1, \dots, q\}$ define the projection matrix π_S

that maps the vector $v = (v_1, \dots, v_q)^t$ to the subvector $\pi_S v = v_S$ with components v_j for $j \in S$. Let $Q_S = J^{11,S} = (\pi_S Q^{-1} \pi_S^t)^{-1}$, $G_S = \pi_S^t Q_S \pi_S Q^{-1}$ and finally

$$\nu = J_{10} J_{00}^{-1} \nabla_{\theta} \mu(\theta_0, \gamma_0) - \nabla_{\gamma} \mu(\theta_0, \gamma_0) \quad \text{and} \quad \tau_0^2 = \nabla_{\theta} \mu(\theta_0, \gamma_0)^t J_{00}^{-1} \nabla_{\theta} \mu(\theta_0, \gamma_0). \quad (3.9)$$

By the arguments of Section 3.2 in Hjort & Claeskens (2003), it now follows from Proposition 3.2 above that $D_n = \sqrt{n}(\hat{\gamma}_n - \gamma_0) \rightarrow_d D \sim N_q(\delta, Q)$. Moreover, by standard application of the delta method

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{d} \Lambda_S = \Lambda_0 + \nu^t(\delta - G_S D), \quad (3.10)$$

where $D \sim N_q(\delta, Q)$ and $\Lambda_0 \sim N(0, \tau_0^2)$ are independent random variables, see Lemma 3.3 in Hjort & Claeskens (2003). The mean squared error of (3.10) is then

$$\text{mse}(S, \delta) = \tau_0^2 + \nu^t G_S Q G_S^t \nu + \nu^t (I_q - G_S) \delta \delta^t (I_q - G_S)^t \nu. \quad (3.11)$$

In the limit experiment, the quantities appearing in (3.11) are known, apart from δ , for which the statistical information is $D \sim N_q(\delta, Q)$, with known variance matrix Q . Since DD^t has mean $\delta \delta^t + Q$, the unbiased estimator of $\delta \delta^t$ is $DD^t - Q$. Thus an unbiased risk estimate here is

$$r(S) = \tau_0^2 + \nu^t G_S Q G_S^t \nu + \nu^t (I_q - G_S) (DD^t - Q) (I_q - G_S)^t \nu.$$

Estimating the required quantities from the data at hand we arrive at

$$\text{FIC}(S) = \hat{\tau}_0^2 + \hat{\nu}^t G_S \hat{Q} G_S^t \hat{\nu} + \hat{\nu}^t (I_q - \hat{G}_S) (D_n D_n^t - \hat{Q}) (I_q - \hat{G}_S)^t \hat{\nu}, \quad (3.12)$$

seen as an estimate of the limiting risk $\text{mse}(S, \delta)$ of (3.11). We then select the model with smallest $\text{FIC}(S)$. Note that the first term is common to all models and may hence be omitted, as far as the model comparison and selection is concerned, but it is useful to keep the information intact and to compute and display estimates of risk, as opposed to a function of the risk. See also Claeskens & Hjort (2008, Ch. 5 & 6) for a more detailed discussion.

The FIC of (3.12) is well-motivated and works well in practice. A useful alternative emerges by truncating the implied estimate of squared bias to zero, in cases where that term is negative. This modified FIC is hence

$$\text{FIC}^*(S) = \hat{\tau}_0^2 + \hat{\nu}^t G_S \hat{Q} G_S^t \hat{\nu} + \max\{h_n(S), 0\}, \quad (3.13)$$

where $h_n(S) = \hat{\nu}^t (I_q - \hat{G}_S) (D_n D_n^t - \hat{Q}) (I_q - \hat{G}_S)^t \hat{\nu}$. Our FIC plots in Figure 1.1 and elsewhere have $\text{FIC}^*(S)^{1/2}/\sqrt{n}$ on the x-axis and estimates $\hat{\mu}_S$ on the y-axis, with root-FIC scores to be interpreted as estimates of root mean squared error (rmse).

3.3. Estimation and detrending. To compare the performance of the submodel estimators using the FIC machinery of (3.12), or the bias corrected (3.13), we need consistent estimates for τ_0, ν, G_S and Q_S , which follows via consistent estimates of J_{wide} and $\nabla\mu$. Since the components of $\nabla\mu$ are assumed to be continuous in a neighbourhood of (θ_0, γ_0) , the unknown $\nabla\mu_0$ is estimated using the standard plug-in principle with a P_δ -consistent estimator for (θ_0, γ_0) . This is also the case for J_{wide} , since under the conditions of Assumption 3.1 it follows by Lemma 11.1 in the Appendix that

$$\hat{J}_{\text{wide}} = J_{\text{wide}}(\bar{\theta}_n, \bar{\gamma}_n) \xrightarrow{P_\delta} J_{\text{wide}}(\theta_0, \gamma_0) = J_{\text{wide}},$$

provided $(\bar{\theta}_n, \bar{\gamma}_n)$ is a P_δ -consistent. With the maximum likelihood estimator this can be accomplished at the usual root- n rate. Alternatively, J_{wide} can be estimated from the second derivative of the log-likelihood function, which from the results of Proposition 3.2 is easily seen to result in a consistent estimate. Note that by the local large-sample framework, estimates obtained from both wide and narrow model can be used to estimate J_{wide} and $\nabla\mu_0$, however, to be model robust we usually prefer estimates under the wide model.

Let ϵ_t be a stationary time series process and suppose we observe $y_1, \dots, y_n$ from the model $Y_t = m(x_t, \beta_0) + \epsilon_t$, where $x_t \in \mathbb{R}^d$, is a sequence of covariates and $m(x_t, \beta)$, for $\beta \in \mathbb{R}^p$, is a function of known parametric form. Let $\hat{y}_t = y_t - m(x_t, \hat{\beta}_n)$, where $\hat{\beta}_n$ is a consistent estimates for β_0 . Then under additional mild regularity assumptions, the detrended series $\hat{y}_t$ will behave as the underlying stationary process ϵ_t with respect to the results of Section 3.2, see Hermansen & Hjort (2014a). This shows that the detrended series, in the FIC machinery, possesses a type of oracle property, meaning that we can remove the estimated trend (as if this was the actual true trend function) and do not risk breaking the conclusion of the final analysis. Note that this is not true in general, for example, mean correcting a stationary Gaussian time series process will change the structure and interpretation of the standard AIC formula, see Hermansen & Hjort (2014b).

4. POST-SELECTION ESTIMATORS, MODEL AVERAGING AND FIC AVERAGING

For a focus parameter μ , consider a model averaging estimator of the general form

$$\hat{\mu} = \sum_S c(S | D_n) \hat{\mu}_S,$$

where the $c(S | d)$ functions sum to one over all candidate models S . This class in particular encompasses post-selection estimators, where the $c(S | D_n)$ is 1 for the winning model and 0 for the others, as long as the model selection scheme being used can be expressed in terms of D_n , as is the case for our FIC. With arguments essentially similar to those used in Section 7.3 of Claeskens & Hjort (2008), it holds that

$$\sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \rightarrow_d \Lambda_0 + \nu^t \{\delta - \hat{\delta}(D)\}, \quad (4.1)$$

where $\widehat{\delta}(D) = \sum_S c(S|D)G_S D$. The underlying reasons for this general result to hold are that (3.10) holds jointly, for all submodels S , and also jointly with $D_n \rightarrow_d D$.

The limit distributions (4.1) are nonlinear mixtures of normals, and are often highly non-normal, e.g. featuring skewnesses and multimodality. We note that result (4.1) is needed when it comes to actual performance analysis for the implied post-selection estimators arrived at via our FIC methods, or other model selectors. Though this feature is a vital component of model selection too, we concentrate in this article on motivating and building the model selectors themselves. This also includes the smoothed FIC case, where $c(S|D_n)$ is proportional to $\exp\{-\lambda \text{FIC}(S)\}$ and λ is a tuning parameter.

The FIC machinery we have developed is perfectly suited to situations where the analyst focuses on one aspect at the time, so to speak. In other situations it is more practical and suitable to consider several foci jointly, say an above-threshold type probability of the type $\mu(a) = \Pr\{Y_{n+1} \geq a, Y_{n+2} \geq a | \text{data}\}$, but where a can be one of several values. This calls for an extension of the FIC to a weighted or averaged FIC, which we term the AFIC. It requires a list of focus parameters, say $\mu(u)$, suitably indexed by some $u \in \mathcal{U}$, and also a weight function $w(u)$, indicating the relative importance of doing well when estimating $\mu(u)$. The focus parameters may form a continuum, but for convenience of presentation we assume that there is a finite set only.

First note that result (3.10) works for each $\mu(u)$, when comparing the different estimators $\widehat{\mu}_S(u)$, so that

$$\sqrt{n}\{\widehat{\mu}_S(u) - \mu_S(u)\} \rightarrow_d \Lambda_0(u) + \nu(u)^\top(\delta - G_S D).$$

Also, with loss $L_n(S) = n \sum_u \{\widehat{\mu}_S(u) - \mu(u)\}^2 w(u)$, we have under mild conditions that the corresponding risk converges,

$$r_n(S) = \mathbb{E} L_n(S) \rightarrow r(S) = \mathbb{E} \sum \{\Lambda_0(u) + \nu(u)^\top(\delta - G_S D)\}^2 w(u).$$

Using earlier results, this may be expressed as

$$r(S) = \sum \tau_0(u)^2 w(u) + \sum \nu(u)^\top G_S Q G_S^\top \nu(u) w(u) + \sum \nu(u)^\top (I_q - G_S) \delta \delta^\top (I_q - G_S)^\top \nu(u) w(u).$$

The first term does not depend on S , leaving us with

$$\text{tr}(G_S Q G_S^\top A) + \text{tr}\{(I_q - G_S) \delta \delta^\top (I_q - G_S)^\top A\},$$

featuring the matrix $A = \sum \nu(u) \nu(u)^\top w(u)$. Estimating these quantities, as in Section 3.3, this leads to

$$\text{AFIC}(S) = \text{tr}(\widehat{G}_S \widehat{Q} \widehat{G}_S^\top \widehat{A}) + \max[0, \text{tr}\{(I_q - \widehat{G}_S)(D_n D_n^\top - \widehat{Q})(I_q - \widehat{G}_S)^\top \widehat{A}\}]. \quad (4.2)$$

As an illustration of the AFIC concept, suppose we have a list of time series model candidates. Let μ_j be the covariance $C(j)$, along with user defined importance weights $w(j)$, typically decreasing since covariances for smaller j are more important than covariances for bigger j . The AFIC apparatus above then makes it possible to select the best of the candidate models, with this covariance perspective

in mind. Similarly an AFIC algorithm can easily be constructed to select the best model from the perspective of having precise estimates for $\mu_j = f(\pi j/j_{\max})$, with a suitable $j_{\max}$.

Illustration 4.1. An related idea is to use the FIC methodology above to construct a nonparametric estimate for the underlying covariance function. Building this from a sequence of estimated covariance lags, e.g. $\hat{\mu}_j = \hat{C}(j)$ computed for $j \leq j_{\max}$ by the model that wins each FIC competition, will not work in general, since can not guarantee that the set of estimates will become a valid covariance function, i.e. produces positive-semidefinite covariance matrixes. By applying Wold's theorem we are able to work around this problem. Essentially, this result tells us that any right-continuous, non-decreasing and bounded spectral distribution function G will provide us with a valid covariance function, see Section 3.1. Therefore, if we change the above perspective slightly and focus on estimating the spectral density for a set of frequencies, we may apply a similar strategy as above with $\mu_j = f(\pi j/j_{\max})$. The collection of estimates $\hat{f}_1, \dots, \hat{f}_{j_{\max}}$ with linear interpolation, can then be used as a nonparametric estimate for underlying spectral density, which in turn, via application of (3.1), will result in valid nonparametric estimate for the covariance function, see Figure 4.1 for a illustration.

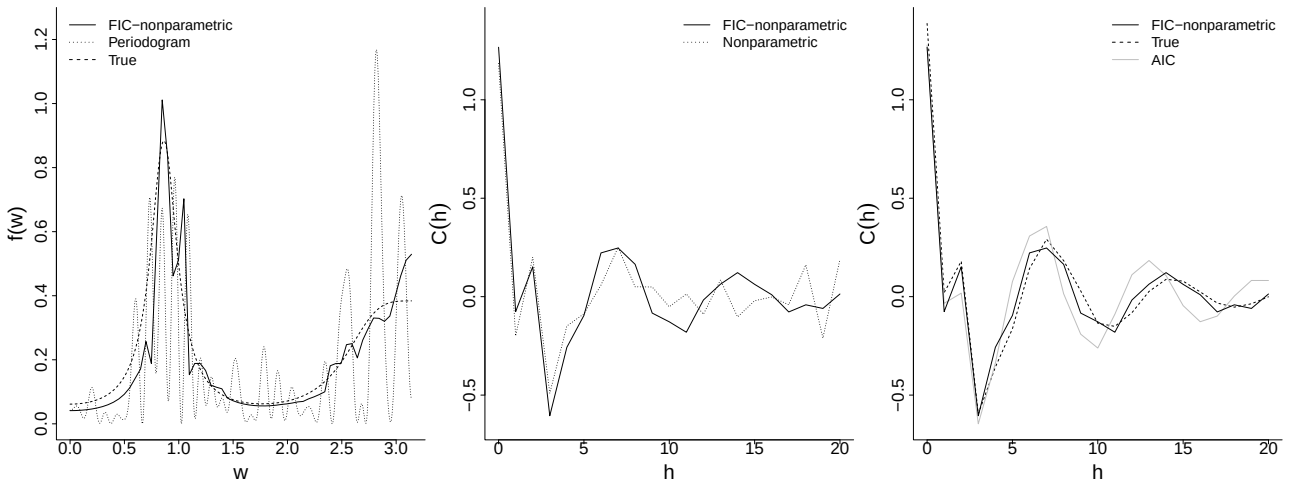


FIGURE 4.1. The estimation is based on a sample size 64 from an autoregressive model of order four with parameters $\rho = (-0.06, 0.17, -0.41, -0.30)$ and $\sigma = 1$. The set of candidate models are all submodels of the true model, i.e. of order less or equal to four. The nonparametric estimate, a mixture of autoregressive models, is shown with periodogram (a classical nonparametric estimate, see Section 7) and the true spectral density (left panel).

5. SAMPLE SIZE DEPENDENT FOCUS FUNCTIONS

The foci discussed in Sections 3 and 4 were all assumed to be functions depending on the underlying model only through model parameters. In the time series framework this is seen to be too restrictive, since it excludes several common and interesting focused questions. In this section we consider a simple extension which opens the possibility for using foci which in addition to the model parameters also

depends on the particular sample size. With the parametric spectral density f_θ , a simple prototype example is $\mu_n = (1/n) \log |\Sigma_n(f_\theta)|$, the scaled log-determinant of the covariance matrix, which can be seen as a general measure for the underlying spread. For the autoregressive model of order one, there is an simple explicit formula and it can easily be shown that $\mu_n = (1/n) \log |\Sigma_n(f_{\sigma,\rho})| = \log \sigma^2 - (\log(1 - \rho^2))/n$.

Since this larger class of foci evolve with sample size n , the large-sample arguments used earlier are not as straightforward anymore. There is however a simple repair operation for this, and the following corollary establishes the required conditions.

Corollary 5.1. *Let $y_1, \dots, y_n$ be realisations from the model in (3.3) and suppose the focus function $\mu_n = \mu(\theta, \gamma, n)$, in addition to the model parameters, depends on the particular sample size n . Then for μ_n to be a ‘valid’ focus function, i.e. that the limit distribution (3.10) follows from the large-sample properties of the score functions (e.g. Proposition 3.2), it is sufficient that μ_n has existing first order derivatives $\nabla \mu_n$, for all $n \geq 1$, which converge uniformly to a limit $\nabla \mu_0$ in a neighbourhood of the point (θ_0, γ_0) .*

Proof. The result follows by extending Lemma 3.3 in Hjort & Claeskens (2003), where the conditions stated are easily seen to be sufficient to ensure the correct limit results. $\square$

Remark 5.2. The result of Corollary 5.1 is more general than stated here and can be generalised to other models under standard regularity conditions, especially, it is easily seen to be true for the models of Sections 8 and 9.

The stronger conditions needed are canonical and easy to motivate, but may be hard to verify in practice, which is somewhat unfortunate, since this removes some of the elegance and applicability of the general FIC construction. In our running example, it is straightforward to show that the conditions of Corollary 5.1 are satisfied, since $\nabla \mu_n = \text{tr}\{\Sigma_n(f_\theta)^{-1} \nabla \Sigma_n(f_\theta)\}/n$ and

$$\nabla \mu_n \rightarrow \nabla \mu_0 = \frac{1}{\pi} \int_0^\pi \frac{\nabla f_\theta(\omega)}{f_\theta(\omega)} d\omega,$$

by standard limit results on Toeplitz matrices, see e.g. Dzhaparidze (1986) or Lemma 11.1 in the Appendix below, which also ensures uniform convergence.

Note that the limit point $\nabla \mu_0$ of Corollary 5.1 does not have to be the derivative of some focus function μ_0 that corresponds to a proper statistical question in the limit experiment, it only has to exist, for the delta method to work. Therefore, if μ_n has limit μ_0 , where μ_0 is independent of sample size and otherwise valid, i.e. the derivatives of μ_0 , with respect to (θ, γ) , exist and are continuous in a neighbourhood of (θ_0, γ_0) , it motivates the idea of using μ_0 as an approximation of μ_n . This removes the dependency on data, meaning that we are back to the framework of Section 3. The idea works

well with the running log-determinant example, since

$$\mu_n \rightarrow \mu_0 = \frac{1}{\pi} \int_0^\pi \log\{2\pi f_\theta(\omega)\} d\omega$$

by Lemma 11.1 in the Appendix (e.g.). Moreover, it is easily shown that for the autoregressive models of order one

$$\mu_0 = \frac{1}{\pi} \int_0^\pi \{2 \log \sigma - \log(1 - 2\rho \cos(\omega) + \rho^2)\} d\omega = \log \sigma^2,$$

which fits well with the exact formula stated above. To our understanding, this is true for autoregressive models in general, meaning that $\mu_0 = \log \sigma^2$, independent of the model order.

Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of the covariance matrix $\Sigma_n(f_\theta)$. Then a variety of foci of this form are motivated by the convergence

$$\mu_n = \frac{1}{n} \sum_{i=1}^n H(\lambda_i) \rightarrow \mu_0 = \frac{1}{2\pi} \int_0^{2\pi} H(f_\theta(\omega)) d\omega,$$

for functions H that are continuous on the range of f_θ , see Gray (2006) for additional regularity conditions, applications and details. As a final illustration, this also introduces the idea of using $\mu_0 = \min_\omega f_\theta(\omega)$ and $\mu_0 = \max_\omega f_\theta(\omega)$ as focus parameters, with the interpretation as upper and lower bounds for the eigenvalues.

6. DATA-DEPENDENT FOCUS FUNCTIONS AND THE CONDITIONAL FIC

Let $y_1, \dots, y_n$ be realisations from the stationary models of Section 3 and consider the general type of ‘data-dependent’ focus $\mu(\text{data}) = \mathbb{E}\{g(Y_{n+1}, \dots, Y_{n+k}) \mid \text{data}\}$, for a suitable function g and a finite number k . A simple prototype example would be $g(Y_{n+1}, Y_{n+2}, Y_{n+3}) = I(Y_{n+1} > a, Y_{n+2} > a, Y_{n+3} > a)$ for some threshold a . Note that these depend on both the model parameters and the observed data. The aim of the present section is to extend the previous FIC machinery to cover such focus functions too. This will be seen to involve a nontrivial extension of the FIC criteria worked with earlier in the literature, since these previous versions essentially all involve independent data. Moreover, it also motivates a conditional version of the focused information criterion (cFIC; Section 6.2). Though the following development can be similarly worked with also in the more general framework of Sections 8 and 9, we focus attention here on the stationary models of Section 3.

In the lemma below we show that working with the data-dependent foci $\mu(\text{data})$ of the above form, is related to working with the prediction version of the problem, that of predicting $g(Y_{n+1}, \dots, Y_{n+k})$.

Lemma 6.1. *Let $y_1, \dots, y_n$ be realisations from the model of (3.3). Then the submodel S , with submodel predictor $\hat{g}_S$ for $g(Y_{n+1}, \dots, Y_{n+k})$, which minimises the mean squared error of $g(Y_{n+1}, \dots, Y_{n+k})$, is the same model, with submodel estimator $\hat{\mu}_S$, which minimises the mean squared error of $\mu_{\text{true}}(\text{data}) = \mathbb{E}\{g(Y_{n+1}, \dots, Y_{n+k}) \mid \text{data}\}$.*

Proof. The result follows essentially from the observation that

$$\begin{aligned} \mathbb{E} \{ \widehat{g}_S - g(Y_{n+1}, \dots, Y_{n+k}) \}^2 &= \mathbb{E} \{ \widehat{g}_S - \mu_{\text{true}}(\text{data}) \}^2 \\ &\quad - 2\mathbb{E} (\widehat{g}_S - \mu_{\text{true}}(\text{data})) (g(Y_{n+1}, \dots, Y_{n+k}) - \mu_{\text{true}}(\text{data})) \\ &\quad + \mathbb{E} \{ g(Y_{n+1}, \dots, Y_{n+k}) - \mu_{\text{true}}(\text{data}) \}^2, \end{aligned}$$

where the rightmost term does not depend on the particular estimator and the cross-product is seen to be zero, via iterated expectations. $\square$

Remark 6.2. The conclusion of Lemma 6.1 is more general than stated above and is easily generalised to other models under standard regularity conditions. In addition, we note that an analogous result still holds if the underlying perspective is changed from mean squared error to conditional mean squared error, which will be important for the discussion in Section 6.2.

Since the submodel estimators $\widehat{\mu}_S$ for the data-dependent focus functions $\mu(\text{data})$ will depend on the observed series through the estimated parameters, $(\widehat{\theta}_S, \widehat{g}_S)$, and the function itself, the large-sample arguments (3.10)–(3.12) will not work as smoothly. Moreover, it is at the outset not clear how to make sense of the large-sample approximations with these data-dependent functions. Take the h -step ahead predictions as an example, in a way it is a little unclear how to interpret $\mu(\text{data}) = \mathbb{E} \{ Y_{n+h} \mid y_1, \dots, y_n \}$ in the limit experiment, since in the current framework it seems like Y_{n+h} should eventually be observed and its value known, or the focus will constantly change its meaning, making it a somewhat poorly defined question.

Before we clear up these ambiguities we consider a motivating illustration. Let $y_1, \dots, y_n$ be realisations from the model (3.3), where the widest model is an autoregressive model of finite order m . Suppose furthermore that our interest is in estimating the probability that two consecutive observations will be below a given threshold a . This focus can be worked with from different perspectives. For example, we may consider $\mu = \Pr\{Y_{n+1} < a, Y_{n+2} < a\}$. Since future outcomes depend on the present state, however, it will often be more appropriate to work with

$$\mu(\text{data}) = \Pr\{Y_{n+1} < a, Y_{n+2} < a \mid \text{data}\} = \Pr\{Y_{n+1} < a, Y_{n+2} < a \mid Y_{n-m}, \dots, Y_n\},$$

where the second equality follows from the Markov property of finite-order autoregressive models. With this data-dependent focus, we can now search for the submodel S that minimises either

$$(i) \quad \mathbb{E} n(\widehat{\mu}_S - \mu_{\text{true}}(\text{data}))^2 \quad \text{or} \quad (ii) \quad \mathbb{E} \{ n(\widehat{\mu}_S - \mu_{\text{true}}(\text{data}))^2 \mid Y_{n-m} = y_{n-m}, \dots, Y_n = y_n \}. \quad (6.1)$$

Note that (i) is the mean of (ii).

To be specific, consider the task of estimating the probability that the final two laps in a 10k speed skating race are both below 30.2 seconds. In this case, we might be interested in a particular race, the one we are watching right now, which fits well with the conditional mean squared error

view (ii) above. Alternatively, we might be more interested in the general dynamics, the underlying features that influence the final two lap times, which is more in the spirit of the unconditional mean squared error (i) of (6.1), see also Example 8.1 below.

Motivated from the scenarios sketched out in (6.1), we first extend the FIC to handle the data-dependent foci and then develop a conditional focused information criterion (cFIC), derived from the conditional mean squared error approach to model assessment.

6.1. Introducing the FIC with data-dependent focus functions. To simplify the discussion we reindex the observed series and let the set of realisations be given by $y_{-n}, \dots, y_{-1}$, where our interest is the behaviour of ‘future’ time points $t \geq 0$. The idea is to construct a large-sample framework where the data-dependent focus function can be thought of as fixed, with respect to the sample size, such that the limit arguments, analogue to those of (3.10)–(3.12), have time to kick in by imagining the series extending to the infinite past, without interfering with the chosen foci.

Define the recent history $\mathcal{H}_m = \{y_{-m}, \dots, y_{-1}\}$ to be the ‘last’ m observations and assume that there exists a finite number m such that

$$\mu(\theta, \gamma, \text{data}) = \mu(\theta, \gamma, \mathcal{H}_m), \quad (6.2)$$

where the derivatives of $\mu(\theta, \gamma, \mathcal{H}_m)$, with respect to the model parameters (θ, γ) , are continuous in a neighbourhood of (θ_0, γ_0) with probability one. The size of m may depend on the problem at hand, as with the autoregressive models, where m is equal to the order of the largest model. The construction in (6.2) will in many cases be a reasonable approximation, or alternatively, there might be good reasons to restricting the history to the last m observations.

The aim is to find the submodel S that minimises

$$\mathbb{E} n(\hat{\mu}_S - \mu_{\text{true}}(\mathcal{H}_m))^2 = \mathbb{E} \{ \mathbb{E} \{ n(\hat{\mu}_S - \mu_{\text{true}}(\mathcal{H}_m))^2 \mid \mathcal{H}_m \} \}, \quad (6.3)$$

where $\mu_{\text{true}}(\mathcal{H}_m) = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n}, \mathcal{H}_m)$ and $\hat{\mu}_S = \mu_n(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c}, \mathcal{H}_m)$. If the estimated parameters are based the reduced data set, $y_{-n}, \dots, y_{-(m+1)}$, it follows by application of the delta method that

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \mid \mathcal{H}_m \xrightarrow{d} \Lambda_S(\mathcal{H}_m) = \Lambda_0(\mathcal{H}_m) + \nu(\mathcal{H}_m)^\top (\delta - G_S D), \quad (6.4)$$

where $\nu(\mathcal{H}_m) = J_{10} J_{00}^{-1} \nabla_\theta \mu_0(\mathcal{H}_m) - \nabla_\gamma \mu_0(\mathcal{H}_m)$ and $\mu_0(\mathcal{H}_m) = \mu(\theta_0, \gamma_0, \mathcal{H}_m)$, and where $D \sim N_q(\delta, Q)$ and $\Lambda_0(\mathcal{H}_m) \mid \mathcal{H}_m \sim N(0, \tau_0(\mathcal{H}_m)^2)$, with $\tau_0(\mathcal{H}_m)^2 = \nabla_\theta \mu_0(\mathcal{H}_m)^\top J_{00}^{-1} \nabla_\theta \mu_0(\mathcal{H}_m)$, are independent random variables; see also the arguments in Section 3.2 leading up to the limit distribution in (3.10). Note that we continue to use $\sqrt{n}$, although $\sqrt{n-m}$ is actually more appropriate in the current framework, however, since m is finite the change is irrelevant with respect to the large sample arguments. The mean squared error of (6.4), condition on $\mathcal{H}_m$, is therefore given by

$$\text{mse}(S, \delta, \mathcal{H}_m) = \tau_0(\mathcal{H}_m)^2 + \nu(\mathcal{H}_m)^\top G_S Q G_S^\top \nu(\mathcal{H}_m) + \nu(\mathcal{H}_m)^\top (I_q - G_S) \delta \delta^\top (I_q - G_S)^\top \nu(\mathcal{H}_m) \quad (6.5)$$

and continuing the derivation from (6.3), it now follows that

$$\begin{aligned} \mathbb{E} \{n(\hat{\mu}_S - \mu_{\text{true}}(\mathcal{H}_m))^2\} &\approx \mathbb{E} \{\text{mse}(S, \delta, \mathcal{H}_m) \mid \mathcal{H}_m\} \\ &= \mathbb{E} \{\tau_0(\mathcal{H}_m)^2 + \nu(\mathcal{H}_m)^t G_S Q G_S^t \nu(\mathcal{H}_m) \\ &\quad + \nu(\mathcal{H}_m)^t (I_q - G_S) \delta \delta^t (I_q - G_S)^t \nu(\mathcal{H}_m)\}. \end{aligned}$$

for large n . This entails that

$$\begin{aligned} \mathbb{E} \{n(\hat{\mu}_S - \mu_{\text{true}})^2\} &\approx \tau_0^{*2} + \nu^{*t} \{G_S Q G_S^t + (I_q - G_S) \delta \delta^t (I_q - G_S)^t\} \nu^* \\ &\quad + \text{tr}(N \{G_S Q G_S^t + (I_q - G_S) \delta \delta^t (I_q - G_S)^t\}), \end{aligned} \quad (6.6)$$

with $\tau_0^{*2} = \mathbb{E} \tau_0(\mathcal{H}_m)^2 = \nabla \mu^{*t} J_{00}^{-1} \nabla \mu^* + \text{tr}\{J_{00}^{-1} M_{00}\}$, $\nu^* = \mathbb{E} \nu(\mathcal{H}_m) = J_{10} J_{00}^{-1} \nabla_{\theta} \mu^* - \nabla_{\gamma} \mu^*$ and

$$N = \text{Var} \nu(\mathcal{H}_m) = J_{10} J_{00}^{-1} M_{00} J_{00}^{-1} J_{01} + M_{11} - J_{10} J_{00}^{-1} M_{01} - M_{01} J_{00}^{-1} J_{10},$$

where

$$\nabla \mu^* = \mathbb{E} \nabla \mu(\theta_0, \gamma_0, \mathcal{H}_m) \text{ and } M = \begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix} = \text{Var} \nabla \mu_n(\theta_0, \gamma_0, \mathcal{H}_m).$$

Then, by arguments similar to those presented in previous sections, the results above leads to a new FIC formula for the data-dependent focus functions

$$\text{FIC}(S, \mathcal{H}_m) = \hat{\tau}_0^{*2} + \text{tr}\{(\hat{\nu}^* \hat{\nu}^{*t} + N)(I_q - \hat{G}_S)(D_n D_n^t - \hat{Q})(I_q - \hat{G}_S)^t\} + \text{tr}\{(\hat{\nu}^* \hat{\nu}^{*t} + N)G_S Q G_S^t\}. \quad (6.7)$$

The estimates of G_S , D_n and Q are obtained in accordance with the discussion of Section 3.3 and the mean and variance of $\nabla \mu(\theta_0, \gamma_0, \mathcal{H}_m)$ are easiest estimated using standard parametric or nonparametric bootstrapping techniques. In simulated examples, both parametric and nonparametric bootstrap provide accurate results, even in moderate samples, indicating good convergence rates.

6.2. The conditional focused information criterion (cFIC). In several real applications, the main interest regarding the data-dependent foci, is with the performance of the submodel estimators given a particular set of observations, those we have actually observed. This encourages a slightly alternative approach, namely to find the submodel S that minimises

$$\mathbb{E} \{n(\hat{\mu}_S - \mu_{\text{true}}(\mathcal{H}_m))^2 \mid \mathcal{H}_m\},$$

with μ_{true} , μ_n , $\hat{\mu}_S$ and m as discussed in relation to (6.2). It is this change, from mean squared error to conditional mean squared error, that leads to the conditional focused information criterion (cFIC). From the results of Section 6.1 above, this criterion is now easily motivated as an estimate for the conditional mean squared error $\text{mse}(S, \delta, \mathcal{H}_m) \mid \mathcal{H}_m$ of (6.4), i.e. $\text{cFIC}(S, \mathcal{H}_m) = \widehat{\text{mse}}(S, \delta, \mathcal{H}_m) \mid \mathcal{H}_m$, where

$$\text{cFIC}(S, \mathcal{H}_m) = \hat{\tau}_{0n}^2 + \hat{\nu}_n^t G_S \hat{Q} G_S^t \hat{\nu}_n + \hat{\nu}_n^t (I_q - G_S)(D_n D_n^t - \hat{Q})(I_q - G_S)^t \hat{\nu}_n, \quad (6.8)$$

for $\hat{\tau}_{0n} = \hat{\tau}_0(\mathcal{H}_m)$ and $\hat{\nu}_n = \hat{\nu}(\mathcal{H}_m)$. The above formula is essentially the same as the one derived in (3.12), the difference lies in the interpretation, modulo the change to a (potentially) reduced dataset for

the parameter estimation; see Remark 6.3 below. Also, the conditional mean squared error framework is how we should interpret the FIC-scores if we neglect the fact that the foci now depend on the observed series and use the results and formulae of Section 3.2 without any changes; see also the comment in Remark 6.2. Moreover, the arguments presented can also be used to motivate a bias corrected cFIC* and the model averaging and AFIC of Section 4.

Remark 6.3. In the derivations above, we assumed that a reduced sample was used for estimating the unknown parameters. The main reason for this is that it removes the dependency between the focus function and the estimated model parameters, which in turn simplifies the large-sample arguments. In practice, however, the question is whether to follow this or use all observations in the estimation. In simple simulation experiments, with small samples sizes, we observe that by using all observations we gain some precision, with respect to the actual mean squared error, but the calculated FIC values have larger variance. To reach general conclusions and recommendations additional efforts are needed, however.

Illustration 6.4. As an illustration of both the conditional and unconditional mean squared error perspective, we compare the independent and the autoregressive model of order one at making one-step ahead predictions. Let $y_{-n}, \dots, y_{-1}$ be realisations from the wide model, which is parametrised by $(\theta_0, \gamma_0 + \delta/\sqrt{n}) = (\sigma_0, \delta/\sqrt{n})$, since $\gamma_0 = 0$. The focus is to predict the outcome of y_0 given the past observations, which from Lemma 6.1 above is equivalent to comparing estimators for the conditional expectation $E\{y_0 \mid \text{data}\}$. Since

$$Y_0 \mid y_{-1}, \dots, y_{-n} \sim N(\delta y_{-1}/\sqrt{n}, \sigma_0^2),$$

we have two alternative submodel estimates

$$\hat{\mu}_{\text{narr}} = \hat{\mu}_\emptyset = 0 \quad \text{or} \quad \hat{\mu}_{\text{wide}} = \hat{\mu}_1 = \hat{\rho}_{n-1} y_{-1},$$

where $\hat{\rho}_{n-1}$ is the maximum likelihood estimator based on the reduced sample. To compare the performance of these submodel estimators, we have to choose between the mean squared error of

$$\sqrt{n}(\hat{\mu}_S - y_0) \mid y_{-1} \quad \text{or} \quad \sqrt{n}(\hat{\mu}_S - y_0),$$

for $S \in \{\emptyset, 1\}$. We shall now consider these two. To make the derivations more clear, we will use the exact expressions for conditional and unconditional mean squared error instead of the large-sample approximations above. The optimal model, in view of the unconditional mean squared error, is the one that has the smallest value of

$$E[(\hat{\mu}_1 - y_0)^2] = E[y_{-1}^2] E[(\hat{\rho}_{n-1} - \delta/\sqrt{n-1})^2 \mid y_{-1}] = \sigma_0^2 E[(\hat{\rho}_{n-1} - \delta/\sqrt{n-1})^2]$$

and

$$E[(\hat{\mu}_\emptyset - y_0)^2] = \sigma_0^2 \delta,$$

meaning that we prefer the wide model if $E[n(\hat{\rho}_{n-1} - \delta/\sqrt{n-1})^2] < \delta$. Moreover, for the conditional mean squared error, the comparison is between

$$E[(\hat{\mu}_1 - y_0)^2 | y_{-1}] = y_{-1}^2 E[(\hat{\rho}_{n-1} - \delta/\sqrt{n-1})^2]$$

and

$$E[(\hat{\mu}_0 - y_0)^2 | y_{-1}] = y_{-1}^2 \delta,$$

which turns out to give exactly the same conclusion, i.e. we prefer the largest model in both cases if $E[n(\hat{\rho}_{n-1} - \delta/\sqrt{n-1})^2] < \delta$. The equality of the two mean squared error perspectives follows as a consequence of the simple structure of the models and the particular focus and we do not expect it to be true in general.

7. THE FIC FROM THE WHITTLE APPROXIMATION

We will now give an alternative derivation of the FIC machinery motivated from the Whittle approximation for the Gaussian log-likelihood of stationary processes, introduced by P. Whittle in a series of works (cf. Whittle (1953)) from the 1950s. This results in an alternative, numerically simpler and more stable version of the focused information criterion, say FIC_w , which from a large-sample perspective is precisely as well-motivated as the version employing the exact log-likelihood. We note that this is at the outset not to be taken for granted. In Hermansen & Hjort (2014b) it is shown that working towards a model-robust AIC type formula, starting from respectively the full log-likelihood and the Whittle approximation, leads to two slightly different schemes, regarding structure and interpretation.

The Whittle approximation for the log-likelihood in (3.5) is

$$\tilde{\ell}_n(\theta, \gamma) = -\frac{n}{2} \left\{ \log 2\pi + \frac{1}{2\pi} \int_{-\pi}^{\pi} \log(2\pi f_{\theta, \gamma}(\omega)) d\omega + \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{I_n(\omega)}{f_{\theta, \gamma}(\omega)} d\omega \right\}, \quad (7.1)$$

where $I_n(\omega) = |\sum_{t \leq n} y_t \exp(-i\omega t)|^2 / (2\pi n)$ is the periodogram. The approximation also motivates an alternative estimation strategy, analogue to the maximum likelihood principle, by $(\tilde{\theta}_n, \tilde{\gamma}_n) = \arg \max_{\theta, \gamma} \{\tilde{\ell}_n(\theta, \gamma)\}$, often referred to as the Whittle estimator. Furthermore, the Whittle estimators just defined have the same large-sample properties as the corresponding maximum likelihood estimators, in terms of limit distributions and efficiency, see e.g. Dzhaparidze (1986).

The derivation of a FIC from (7.1) follows along similar arguments as presented in Section 3. To avoid redeveloping previous results, we will only provide the main result needed, i.e. the analogue of Proposition 3.2, which is summarised in the theorem below. This result also motivates an alternative proof for Proposition 3.2, which will be shown to be true under weaker conditions than those given in Assumption 3.1, see Remark 7.2 below.

Theorem 7.1. *Let $y_1, \dots, y_n$ be realisations from the model (3.3) and define $\tilde{Z}_n = \sqrt{n}(\tilde{U}_n, \tilde{V}_n) = \nabla \tilde{\ell}_n(\theta_0, \gamma_0) / \sqrt{n}$, i.e. the Whittle version of the scaled score functions. If the spectral density $f_{\theta, \gamma}$ is*

continuous and bounded away from both zero and infinity, moreover, if $f_{\theta,\gamma}$ is two times differentiable in (θ, γ) , with derivatives that are continuous and uniformly bounded in both ω and (θ, γ) in a neighbourhood of (θ_0, γ_0) . Then

$$\tilde{Z}_n = \sqrt{n} \begin{pmatrix} \tilde{U}_n \\ \tilde{V}_n \end{pmatrix} \rightarrow_d \begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix} + Z, \quad \text{where } Z \sim N_{p+q}(0, J_{\text{wide}}),$$

with J_{wide} as defined in (3.7). Also, $\tilde{J}_n(\bar{\theta}_n, \bar{\gamma}_n) = -\nabla^2 \tilde{\ell}_n(\bar{\theta}_n, \bar{\gamma}_n)/n \rightarrow_{P_\delta} J_{\text{wide}}$, provided $(\bar{\theta}_n, \bar{\gamma}_n) \rightarrow_{P_\delta} (\theta_0, \gamma_0)$.

Proof. First of all, observe that

$$\begin{aligned} \tilde{Z}_n = \sqrt{n} \begin{pmatrix} \tilde{U}_n \\ \tilde{V}_n \end{pmatrix} &= -\frac{\sqrt{n}}{4\pi} \int_{-\pi}^{\pi} \frac{\nabla f_0(\omega)[f_0(\omega) - I_n(\omega)]}{f_0(\omega)^2} d\omega \\ &= -\frac{\sqrt{n}}{4\pi} \int_{-\pi}^{\pi} \nabla \Psi_0(\omega) d\omega + \frac{1}{4\pi\sqrt{n}} \mathbf{y}_n^t \Sigma_n (\nabla \Psi_0/[2\pi f_0]) \mathbf{y}_n, \end{aligned} \quad (7.2)$$

where Σ_n is the covariance matrix as defined in (3.5), $\Psi_0 = \log f_0$ and $f_0 = f_{\theta_0, \gamma_0}$. Then, the expectation of $\tilde{Z}_n$ in the limit experiment is

$$\begin{aligned} E_{P_\delta} \tilde{Z}_n &= -\frac{\sqrt{n}}{4\pi} \left[\int_{-\pi}^{\pi} \nabla \Psi_0(\omega) d\omega - \frac{1}{n} \text{tr}\{\Sigma_\delta \Sigma_n (\nabla \Psi_0/[2\pi f_0])\} \right] \\ &= -\frac{\sqrt{n}}{4\pi} \left[\int_{-\pi}^{\pi} \nabla \Psi_0(\omega) d\omega - \int_{-\pi}^{\pi} \frac{f_\delta(\omega) \nabla f_0(\omega)}{f_0(\omega)^2} d\omega + O(n^{-2/3} \log^4 n) \right] \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\nabla f_0(\omega) \delta^t [\nabla_\gamma f_{\theta_0, \bar{\gamma}_n}(\omega)]}{f_0(\omega)^2} d\omega + O(n^{-1/6} \log^4 n) \rightarrow \begin{pmatrix} J_{01}\delta \\ J_{11}\delta \end{pmatrix}, \end{aligned}$$

by application of Lemma 11.1 in the Appendix, where $\|\bar{\gamma}_n - \gamma_0\| < \|\delta\|/\sqrt{n}$, $\Sigma_\delta = \Sigma_n(f_\delta)$ and $f_\delta = f_{\theta_0, \gamma_0 + \delta/\sqrt{n}}$. Moreover, by similar arguments it is seen that

$$\begin{aligned} \text{Var}_{P_\delta} \tilde{Z}_n &= \frac{2}{(4\pi)^2 n} \text{tr}\{[\Sigma_\delta \Sigma_n (\nabla \Psi_0/[2\pi f_0])]^2\} \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{f_\delta(\omega)^2 \nabla \Psi_0(\omega) \nabla \Psi_0(\omega)^t}{f_0(\omega)^2} d\omega + O(n^{-2/3} \log^6 n) \rightarrow J_{\text{wide}}, \end{aligned}$$

see again Lemma 11.1 for details. In order to show that the normalised $\tilde{Z}_n$ has a Gaussian limit distribution, we apply a Cramér–Wold type of argument. For any $\mathbf{a} \in \mathbb{R}^{p+q}$ let $\tilde{A}_n = \mathbf{a}^t \tilde{Z}_n$ and observe that

$$\tilde{A}_n = -\frac{\sqrt{n}}{4\pi} \int_{-\pi}^{\pi} g(\omega) d\omega + \frac{1}{4\pi\sqrt{n}} \mathbf{y}_n^t \Sigma_n (g/[2\pi f_0]) \mathbf{y}_n, \quad (7.3)$$

where $g(\cdot) = a_1 \nabla_1 \Psi_0(\cdot) + \dots + a_{p+q} \nabla_{p+q} \Psi_0(\cdot)$. From results in Hermansen & Hjort (2014a) it now follows that $(A_n - E_{P_\delta} A_n)/\text{Var}_{P_\delta}(A_n)^{1/2}$ is asymptotically normal if and only if

$$\max_{i \leq n} \{\lambda_i\} / \left(\sum_{i=1}^n \lambda_i^2 \right)^{1/2} \rightarrow 0 \quad (7.4)$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $\Sigma_\delta \Sigma_n(g/[2\pi f_0])$, which is easiest shown by establishing bounds on the set of eigenvalues.

The largest eigenvalue, and therefore also the nominator of (7.4), is bounded by the model conditions, since it follows from the properties of this particular type of Toeplitz matrices that $\max_{i \leq n} \{\lambda_i\} \leq \sup_\omega f_\delta(\omega) \sup_\omega g(\omega)$, see e.g. Gray (2006).

If the covariance functions defining $\Sigma_n(f_\delta)$ and $\Sigma_n(g)$ are absolutely summable, i.e. if $\sum_j |C_{f_\delta}(j)| < \infty$ and $\sum_j |C_g(j)| < \infty$, in a neighbourhood of (θ_0, γ_0) , then f_δ and g are in the Wiener class and by Theorem 5.3 in Gray (2006) the matrix product $\Sigma_n(f_\delta)\Sigma_n(g)$ is asymptotically equivalent to the circulant matrix $C(f_\delta g)$, defined by the eigenvalues $f_\delta(\omega_j)g(\omega_j)$, for $\omega_j = 2\pi j/n$ and $j = 0, \dots, n-1$. Then, by the properties of asymptotically equivalent matrices, it follows from the assumed model conditions that $\lim_{n \rightarrow \infty} n^{-1} \sum_i \lambda_i^2 = \lim_{n \rightarrow \infty} n^{-1} \sum_i [f_\delta(\omega_i)g(\omega_i)]^2 = c$, for a finite constant c , which is enough to establish (7.4) and we have the claimed limit distribution.

To show that $\Sigma_n(f_\delta)$ and $\Sigma_n(g)$ has absolutely summable elements. Let h be an arbitrary spectral density with derivative h' (with respect to ω) that is bounded away from infinity and can be divide into a finite number of open subintervals $0 = w_0 < \dots < w_k = \pi$, such that h' is monotonic on each section (w_{j-1}, w_j) and $j \leq k$. Then integration by parts leads to

$$C_h(j) = 2 \int_0^\pi \cos(\omega j) h(\omega) d\omega = -\frac{2}{j} \int_0^\pi \sin(\omega j) h'(\omega) d\omega,$$

for any integer $j > 0$. Moreover, by the second mean value theorem

$$C_h(j) = -\frac{2}{j} \sum_{i=1}^k \int_{w_{i-1}}^{w_i} \sin(\omega j) h'(\omega) d\omega = -\frac{2}{h} \sum_{i=1}^k \left[h'(w_{i-1}^+) \int_{w_{i-1}}^{v_i} \sin(\omega j) d\omega + h'(w_i^-) \int_{v_i}^{w_i} \sin(\omega j) d\omega \right],$$

where v_i is some number in (w_{i-1}, w_i) . This implies that

$$|C_h(j)| \leq \frac{4}{j^2} \sum_{i=1}^k (|h'(w_{i-1}^+)| + |h'(w_i^-)|) \leq \frac{8kK}{j^2},$$

where K is a upper limit for $|h'(\omega)|$ in $[0, \pi]$, which implies, by the assumed conditions, that C_{f_δ} and C_g are both absolute summable.

For the second part of the proposition, we note that

$$\tilde{J}_n(\theta, \gamma) = \frac{1}{4\pi} \int_{-\pi}^\pi \left[\frac{\nabla^2 f_\delta(\omega)(\omega)[f_\delta(\omega) - I_n(\omega)]}{f_\delta(\omega)^2} - \frac{\nabla f_\delta(\omega) \nabla f_\delta(\omega)^t [f_\delta(\omega) - 2I_n(\omega)]}{f_\delta(\omega)^3} \right] d\omega,$$

which results in the correct limit by techniques similar to those already presented above; the details are therefore omitted. $\square$

If $\mu = \mu(\theta, \gamma)$ depends on the underlying model only through the model parameters and if the derivatives, with respect to (θ, γ) , exist and are continuous in a neighbourhood of (θ_0, γ_0) . Then with $\tilde{\mu}_S$ as the corresponding Whittle submodel estimator for submodel S , it follows by Theorem 7.1 that

these will result in the same limit distribution as in (3.10). This motivates, via application of (3.11), an alternative version for the FIC formula (3.12) by

$$\text{FIC}_w(S) = \tilde{\tau}_0^2 + \tilde{\nu}^t G_S \tilde{Q} G_S^t \tilde{\nu} + \tilde{\nu}^t (I_q - \tilde{G}_S) (\tilde{D}_n \tilde{D}_n^t - \tilde{Q}) (I_q - \tilde{G}_S) \tilde{\nu}, \quad (7.5)$$

where $\tilde{D}_n = \sqrt{n}(\tilde{\gamma}_n - \gamma_0)$ and ‘tilde’ indicates that Whittle approximation or related estimates are in use, see also the general comments in Section 3.3. Note that the Whittle version of the FIC can easily be shown to be valid for the extended focus functions of Sections 5 and 6. Similar arguments also motivate a Whittle based bias corrected FIC_w^* and the discussion on model averaging and AFIC of Section 4 can be seen to apply, *mutatis mutandem*.

Remark 7.2. Using the results of Theorem 7.1 we can now give an alternative proof for Proposition 3.2 under weaker conditions. This follows from the fact that P_0 and P_δ are mutually contiguous, see Davis (1973, Theorem 4.1), which essentially implies that convergence in probability means the same thing under both measures. In particular it implies that $\epsilon_n = |Z_n - \tilde{Z}_n| \rightarrow_{P_0} 0$ if and only if $\epsilon_n = |Z_n - \tilde{Z}_n| \rightarrow_{P_\delta} 0$. Then, since the convergence of ϵ_n under P_0 is known by Dahlhaus & Wefelmeyer (1996), it is possible to derive the limit distribution of Z_n under the model (3.3) from limit of $\tilde{Z}_n$, i.e. Theorem 7.1. The reason we can use such an argument here is that Z_n and $\tilde{Z}_n$ are functions of the same data set, in which case convergence in probability is enough to imply convergence in distribution. The argument actually works in both directions, meaning that Theorem 7.1 is in a sense surplus, since it can be shown to follow from Proposition 3.2. Nevertheless, we believe the alternative proof displays interesting use of techniques, which besides the relaxed model conditions justify its presence.

8. THE FIC FOR STATIONARY TIME SERIES MODELS WITH SMOOTH TREND

In this section we extend the stationary models by including a smooth trend component. This is an important class of models providing a coherent framework for joint selection among features with respect to both the trend part and the time series dependency part of an observed series. Let $y_1, \dots, y_n$ be realisations from the model

$$Y_{n,t} = m_{\text{true}}(t/n) + \epsilon_t \quad \text{for } t = 1, \dots, n, \quad (8.1)$$

where ϵ_t is a stationary Gaussian time series process with zero mean. The modelling framework is a local large-sample framework, with true spectral density and trend given by

$$f_{\text{true}} = f_\delta = f_{\theta_0, \gamma_0 + \delta_1 / \sqrt{n}} \quad \text{and} \quad m_{\text{true}} = m_\delta = m(u, \theta_0, \gamma_0 + \delta_2 / \sqrt{n}), \quad (8.2)$$

where $(\theta_0, \gamma_0) \in \Theta \times \Gamma \subset \mathbb{R}^{p+q}$ is an inner point of the compact space $\Theta \times \Gamma$.

The construction with triangular array in (8.1) 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 most standard models the set of parameters split, meaning that each element in the vector (θ, γ) is uniquely related to either trend or spectral density. To make the notation less cumbersome we essentially keep the notation above, but have in mind that $\theta = (\theta_f, \theta_m) \in \mathbb{R}^{p_1+p_2}$ and $\gamma = (\gamma_f, \gamma_m) \in \mathbb{R}^{q_1+q_2}$, where $p = p_1 + p_2$, $q = q_1 + q_2$ and $\gamma_f = \gamma_{0,f} + \delta_f/\sqrt{n}$ and $\gamma_m = \gamma_{0,m} + \delta_m/\sqrt{n}$ are vectors of dimension q_1 and q_2 , respectively.

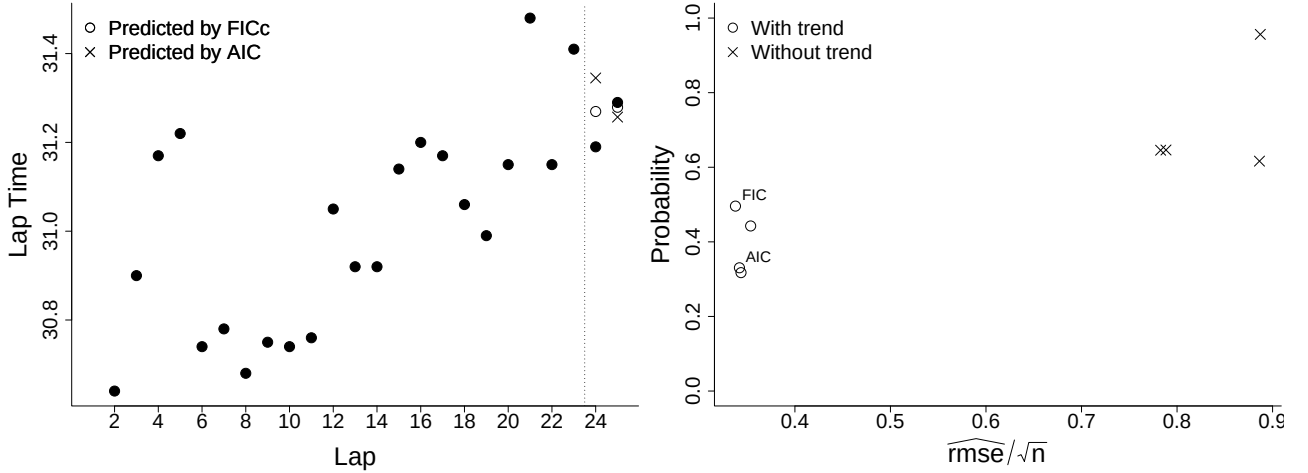


FIGURE 8.1. The lap times of Jochem Uytdehaage historical 10k race during the 2002 Winter Olympic Games in Salt Lake City. The foci are subsequent predictions of the final two laps (left panel) and the probability the total time, after observing the 23 first laps, will be less than 12:59.00.

Example 8.1. As an illustration we re-watch Jochem Uytdehaage's historical 12:58.92 run on the 10,000 meters during the 2002 Winter Olympic Games in Salt Lake City, the first time the 13-minute barrier was broken on this distance. In Figure 8.1 (left panel) the last 24 of the 25 lap times are plotted; we omit displaying the first lap time, as this is always larger than the others due to the time needed to get up to full speed. The task we set here relates to predicting aspects of the two last laps, after having observed the first 23 of the race. The candidate models are the nested submodels of an autoregressive model of order three, with a linear or a constant trend, i.e. eight different models. The two concrete foci we consider for this illustration are the sequential prediction of laps 24 and 25, and the probability that the end time will be below 12:59.00, again, conditional on having watched the race up to and including the 23rd lap. The reason we do not use the historical 13:00.00, is that for this particular threshold all candidate models estimates the probability to be approximately equal to one, and we have to refine the question to obtain interesting answers.

8.1. The limit experiment and the FIC. The arguments needed to obtain a FIC for the extended time series processes are similar to those already presented. Therefore, to avoid reintroduction of

notation and basic results, we only present the main results, modifications and extensions needed to complete the chain of arguments.

To simplify the first steps, we organise the parameters according to their ‘main’ association and group together (θ_f, γ_f) and (θ_m, γ_m) . Let

$$\begin{pmatrix} A_n \\ B_n \end{pmatrix} = \frac{1}{2n} \begin{pmatrix} (\mathbf{y}_n - \mathbf{m}_0)^\top \Sigma_0^{-1} [\nabla \Sigma_0] \Sigma_0^{-1} (\mathbf{y}_n - \mathbf{m}_0) - \text{tr}\{\Sigma_0^{-1} [\nabla \Sigma_0]\} \\ 2[\nabla \mathbf{m}_0]^\top \Sigma_0^{-1} (\mathbf{y}_n - \mathbf{m}_0) \end{pmatrix}$$

be the vector of scaled score functions, where $\mathbf{m}_0^\top = (m_0(1/n), \dots, m_0(1))$ and $m_0(u) = m(u, \theta_0, \gamma_0)$. Since we are only working with Gaussian models, the correlation between A_n and B_n is zero, meaning that

$$\lim_{n \rightarrow \infty} n \text{Var}_{P_0} \begin{pmatrix} A_n \\ B_n \end{pmatrix} = \lim_{n \rightarrow \infty} \begin{pmatrix} n \text{Var}_{P_0} A_n & 0 \\ 0 & n \text{Var}_{P_0} B_n \end{pmatrix} = \begin{pmatrix} J_f & 0 \\ 0 & J_m \end{pmatrix},$$

where under the conditions of Assumption 8.2 below

$$J_f = \frac{1}{4\pi} \int_{-\pi}^{\pi} \nabla \Psi_0(\omega) \nabla \Psi_0(\omega)^\top d\omega \quad \text{and} \quad J_m = \frac{1}{2\pi f_0(0)} \int_0^1 \nabla m_0(u) \nabla m_0(u)^\top du, \quad (8.3)$$

by Lemma 11.1 in the Appendix.

Assumption 8.2. The spectral density $f_{\theta, \gamma}$ is continuous and bounded away from zero and infinity, moreover, $f_{\theta, \gamma}$ and all components of $\nabla f_{\theta, \gamma}$ and $\nabla^2 f_{\theta, \gamma}$ are differentiable in ω with uniformly continuous derivatives. The components of $m(u, \theta, \gamma)$, $\nabla_\theta m(u, \theta, \gamma_0)$ and $\nabla_\theta^2 m(u, \theta, \gamma_0)$ are differentiable in u with uniformly continuous derivatives.

Let $Z_n = \sqrt{n}(U_n, V_n) = \nabla \ell_n(\theta_0, \gamma_0)/n$, be the score functions where the derivatives are with respect to $\theta = (\theta_f, \theta_m)$ and then $\gamma = (\gamma_f, \gamma_m)$. The elements of Z_n is then a permutation of the elements in (A_n, B_n) and J_{wide} is defined by

$$J_{\text{wide}} = \lim_{n \rightarrow \infty} n \text{Var}_{P_0} \begin{pmatrix} U_n \\ V_n \end{pmatrix} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} = \begin{pmatrix} J_{f00} & 0 & J_{f01} & 0 \\ 0 & J_{m00} & 0 & J_{m01} \\ J_{f10} & 0 & J_{f11} & 0 \\ 0 & J_{m10} & 0 & J_{m11} \end{pmatrix}. \quad (8.4)$$

The following corollary establishes the limit distribution of Z_n , since the current models can also be seen to be a special case of the processes in Section 9, the result follows as a corollary of Theorem 9.3 and is therefore stated without a separate proof.

Proposition 8.3. *Let $y_1, \dots, y_n$ be realisations from the model (8.2) and let $P_\delta = P_{\delta, n}$ be the associated probability measure. Then if $f_{\theta, \gamma}$ and $m(u, \theta, \gamma)$ satisfy conditions of Assumption 8.2*

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

with J_{wide} as defined in (8.4) and $\delta = (\delta_1, \delta_2)$. Also, $\widehat{J}_n(\bar{\theta}_n, \bar{\gamma}_n) = -\nabla^2 \ell_n(\bar{\theta}_n, \bar{\gamma}_n)/n \rightarrow_{P_\delta} J_{\text{wide}}$, provided $(\bar{\theta}_n, \bar{\gamma}_n) \rightarrow_{P_\delta} (\theta_0, \gamma_0)$.

Let μ be a focus function and assume that $\mu = \mu(\theta, \gamma)$ depends only on the underlying model through the model parameters and that the derivatives, with respect to (θ, γ) , exist and are continuous in a neighbourhood of (θ_0, γ_0) . Furthermore, let $\widehat{\mu}_S$, for $S \subset \{1, \dots, q\}$ and $q = q_1 + q_2$, represent an arbitrary submodel estimator. Then the result of Proposition 8.3, via the limit distribution of the normalised submodel estimators, leads to the FIC formula

$$\text{FIC}(S) = \widehat{\tau}_0^2 + \widehat{\nu}^t G_S \widehat{Q} G_S^t \widehat{\nu} + \widehat{\nu}^t (I_q - \widehat{G}_S) (D_n D_n^t - \widehat{Q}) (I_q - \widehat{G}_S) \widehat{\nu}, \quad (8.5)$$

where the estimators emerge using the standard plug-in principle, see also the general comments of Section 3.3. Note that (8.5) can easily be extended to the focus functions of Sections 5 and 6. This also motivates a bias corrected FIC* formula as in (3.13), and the topics of Section 4 are also easily seen to be applicable in the current framework.

In the case the model parameters splits, the matrices Q , Q_S and G_S all become block diagonal, which simplifies the structure and interpretation of the above FIC formula. To take full advantage of this, we need to introduce some new notion. Let S_f and S_m be model indexes referring to submodels of the spectral density and the trend function, respectively. Furthermore, define for $i \in \{f, m\}$ the quantities $\psi_{S_i} = \widehat{\nu}_i^t \widehat{G}_{S_i} D_{i,n}$ and $\widehat{\sigma}_{S_i}^2 = \widehat{\nu}_i^t G_S \widehat{Q} G_{S_i}^t \widehat{\nu}_i$, where $\nu_i = J_{i10} J_{i00}^{-1} \nabla_{\theta_i} \mu(f_{\theta_0, \gamma_0}) - \nabla_{\gamma_i} \mu(f_{\theta_0, \gamma_0})$, $G_{S_i} = \pi_{S_i}^t Q_{S_i} \pi_{S_i} Q_i^{-1}$, for the corresponding projection matrix π_{S_i} and where $Q_i = J_i^{11}$ is the lower right corner of J_i^{-1} . The FIC formula (8.5) can now be expressed as

$$\text{FIC}(S) = \widehat{\tau}_0^2 + 2(\widehat{\sigma}_{S_f}^2 + \widehat{\sigma}_{S_m}^2) + (\widehat{\psi}_{\text{wide}} - \widehat{\psi}_{S_f} - \widehat{\psi}_{S_m})^2. \quad (8.6)$$

Remark 8.4. In the arguments presented, both S_f or S_m are assumed to include at least one element. If either set is empty, meaning that we trust completely in the largest trend or dependency model, the framework above may not be the most appropriate, since it reduces the joint selection to a ‘one-dimensional’ selection problem, see the discussion on detrending in Section 3.3 for some related comments.

Remark 8.5. From the block diagonal structure of J_{wide} and the shape of ν used in the FIC formula (8.5) above, it follows that if a particular focus function is independent of either spectral density or trend, the FIC will not be able to ‘detect’ differences between submodels that are invariant with respect to the chosen focus. It is somewhat surprising, since it is clear that the estimates of $\mu = m(1) - m(0)$ or $\mu(1) = C(1) = \text{Cov}(Y_1, Y_2)$, will depend on how the dependency or trend is constructed. It is as consequence of the local large-sample framework, however, and it means that if the particular kind of trend or dependency is of no importance to us, we should not try to model this beyond the narrow model.

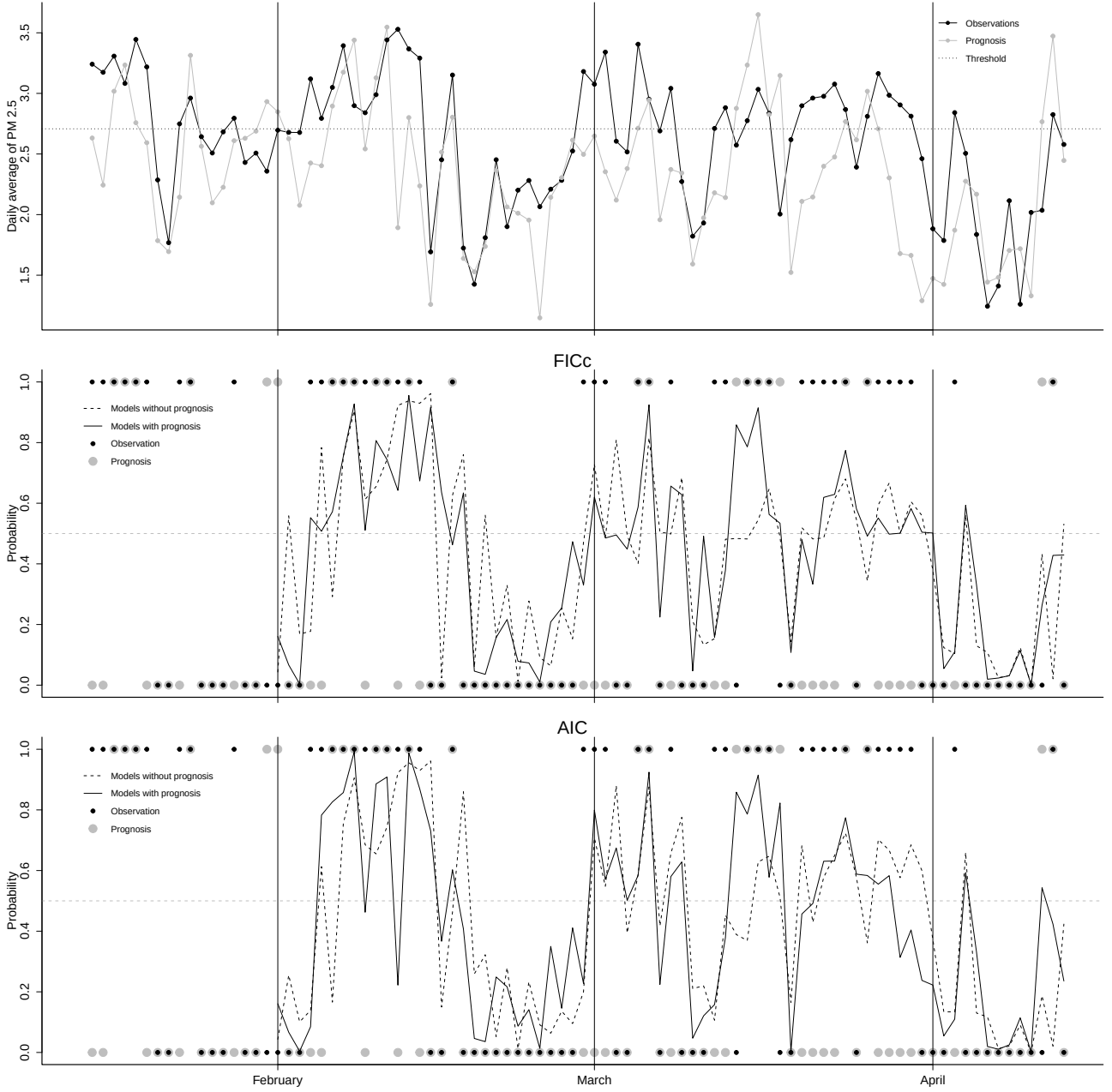


FIGURE 8.2. Daily average (log) measurements and prognosis of $\text{PM}_{2.5}$ at Alnabru, a neighbourhood in Oslo (Norway), measured from January 15 to April 13, 2012 (top panel) Sequential estimates for the probability that ‘tomorrows’ observation will exceed the critical $15 \mu\text{g}/\text{m}^3$ threshold. The effect of including the prognosis in the predictions are shown for both the cFIC and the AIC (centre and bottom panel).

Example 8.6. Consider the modification of (8.2) where

$$m_{\text{true}}(u) = m_{\delta}(u) = x(u)^{\text{t}}\beta + z(u)^{\text{t}}(\gamma_0 + \delta_2/\sqrt{n}),$$

where $x(u)^{\text{t}} = (x_1(u), \dots, x_{p_2}(u))$ are functions, or features, that are included in all models and $z(u)^{\text{t}} = (z_1(u), \dots, z_{q_2}(u))$ are potential extensions. This is a type of regression models with dependent errors and if all components of $x(u)$ and $z(u)$ satisfies the conditions of Assumption 8.2, such models

are easily seen to be covered by the framework above, see also Section 10 for additional comments on regression models with dependent errors.

As an illustration we consider daily average measurements of fine particulate matter (PM_{2.5}), a type of air pollutant referring to tiny particles, or droplets, in the air that are two and one half microns or less in width. Exposure to fine particles can cause short-term effects, like eye, nose, throat and lung irritation, long-term exposure can affect lung function and has also been shown to be related to asthma and heart disease. For health reasons, government regulations typically restrict the daily average emission to 25-35 $\mu\text{g}/\text{m}^3$, however, new directives in Norway suggest that levels can be set as low as 15 $\mu\text{g}/\text{m}^3$ with the possibility of 8 exceedances per year.

The Norwegian Institute for Air Research (NILU) in collaboration with the Norwegian Meteorological Institute (MET) makes forecasts for different kinds of air pollution; in particular they make prognosis for PM_{2.5}, see Figure 8.2. One possible objective is to use such forecast to detect periods or days with potential high emissions, enabling the legislature to take appropriate actions in advanced.

The sequential estimates of the probability that tomorrow's total emission will be above the 15 $\mu\text{g}/\text{m}^3$ threshold, given the current history, are shown in Figure 8.2 for the models selected by AIC and the cFIC. The candidate models are nested autoregressive models of order less than or equal to three, which can include the constant, the linear and/or the prognosis as trend. The idea is to compare how the two criteria assimilate the prognosis when this is included to the set of potential trend models. In this example the cFIC appears to be more robust and the AIC seems to select models that put too much confidence in the prognosis, making the predictions sensitive to this particular type of errors.

9. THE FIC FOR LOCALLY STATIONARY PROCESSES

In this section we introduce the locally stationary processes of Dahlhaus (1997) and derive a more general version of the FIC. These processes extend the classical stationary models by having spectra that change gradually as time evolves. In addition to their intrinsic interest, the models discussed in the previous sections can be seen to be included in this general class. The basic definitions and regularity constraints are as follows.

A sequence of Gaussian stochastic processes $\{Y_{n,t}\}_{t \leq n}$ for $n \geq 1$ is called locally stationary with transfer function A° and trend m if there exists a representation

$$Y_{n,t} = m(t/n) + \int_{-\pi}^{\pi} \exp(i\omega t) A_{t,n}^\circ(\omega) d\xi(\omega),$$

where $\xi(\omega)$ is a Gaussian complex-valued stochastic process on $[-\pi, \pi]$ with $\overline{\xi(\omega)} = \xi(-\omega)$, $E \xi(\omega) = 0$ and orthonormal increments, i.e.

$$E \{d\xi(\omega_1) \overline{d\xi(\omega_2)}\} = \eta(\omega_1 + \omega_2) d\omega_1 d\omega_2$$

where $\eta(\omega) = \sum_{j=-\infty}^{\infty} \delta(\omega + 2\pi j)$ is the period 2π extension of the Dirac delta function. Moreover, we assume that there exists a constant K and a function $A(u, \omega)$ on $[0, 1] \times \mathbb{R}$, which is 2π -periodic in ω , with $A(u, -\omega) = \overline{A(u, \omega)}$ and is such that

$$\sup_{t, \omega} |A_{t,n}^{\circ}(\omega) - A(t/n, \omega)| \leq \frac{K}{n}$$

for all n ; see Dahlhaus (1996a), Dahlhaus (1996b) and Van Bellegem & Dahlhaus (2006) for further discussion and results.

We will continue to work in local large-sample framework and assumes that the true data generating model has transfer function $A_{\text{true}}^{\circ} = A_{\delta}^{\circ} = A_{t,n}^{\circ}(\omega, \theta_0, \gamma_0 + \delta_1/\sqrt{n})$, with corresponding

$$A_{\delta}(u, \omega) = A(u, \omega, \theta_0, \gamma_0 + \delta_1/\sqrt{n}) \quad \text{and trend } m_{\delta}(u) = m(u, \theta_0, \gamma_0 + \delta_2/\sqrt{n}) \quad (9.1)$$

satisfying the conditions of Assumption 9.1. Also, since we usually do not distinguish between the parameters for the transfer and trend function in the large-sample arguments, we will use the more compact notation where $\theta = (\theta_A, \theta_m)$ and $\gamma = (\gamma_A, \gamma_m)$ represents all parameters.

As a simple illustration of such models, observe that stationary models with time varying amplitude are a subset of the locally stationary models. To see this, let Y_t is a stationary process with spectral representation

$$Y_t = \int_{-\pi}^{\pi} \exp(i\omega t) A(\omega) d\xi(\omega),$$

where μ and σ are continuous real values functions on the unit interval. Then $X_t = \mu(t/n) + \sigma(t/n)Y_t$ is locally stationary with $A_{t,n}^{\circ}(\omega) = A(t/n, \omega) = \sigma(t/n)A(\omega)$, see (Dahlhaus, 1996b, p. 142) for the complete discussion and other applications.

The following conditions are a modified version of Assumption 2.1 in Dahlhaus (1996a).

Assumption 9.1. (i) There exists a constant K with

$$\sup_{t, \omega} |\tilde{\nabla}\{A_{t,n}^{\circ}(\omega, \theta, \gamma) - A(t/n, \omega, \theta, \gamma)\}| \leq K/n,$$

where $\tilde{\nabla} = \nabla_i$ or $\tilde{\nabla} = \nabla_{i,j}$. The function $A(u, \omega, \theta, \gamma)$ is uniformly bounded from above and below. Also the components of $A(u, \omega, \theta, \gamma)$, $\nabla A(u, \omega, \theta, \gamma)$ and $\nabla^2 A(u, \omega, \theta, \gamma)$ are differentiable in u and ω with uniformly continuous derivatives. Since we work in a large sample framework, our models are in a sense ‘locally misspecified’, which means that $A(u, \omega, \theta, \gamma_0)$ has to be twice differentiable in u with uniformly bounded derivatives. (ii) The components of $m(u, \theta, \gamma)$, $\nabla m(u, \theta, \gamma_0)$ and $\nabla^2 m(u, \theta, \gamma_0)$ are all differentiable in u with uniformly continuous derivatives.

Remark 9.2. For a general transfer function A the time varying spectral density is defined as $f(u, \omega) = |A(u, \omega)|^2 = A(u, \omega)A(u, -\omega)$, which means that for standard stationary models with real-valued and strictly positive spectral densities we have $A(u, \omega) = f(\omega)^{1/2}$. This makes it straightforward

to translate Assumption 9.1 to fit with the notation of the stationary models, see the conditions of Assumption 3.1 and 8.2.

9.1. The limit experiment and the FIC. Let $Z_n = \sqrt{n}(U_n, V_n) = \nabla \ell_n(\theta_0, \gamma_0)/\sqrt{n}$ denote the vector of scaled score functions evaluated at the null-point (θ_0, γ_0) . Then

$$J_{\text{wide}} = \lim_{n \rightarrow \infty} \text{Var}_{P_0} Z_n = \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \Psi_0(u, \omega) \Psi_0(u, \omega)^t d\omega du + \frac{1}{2\pi} \int_0^1 \frac{\nabla m_0(u) \nabla m_0(u)^t}{f_0(u, 0)} du, \quad (9.2)$$

see Lemma 11.1 in the Appendix for details. The following theorem is an extended version of Theorem 2.4 in Dahlhaus (1996a) and contains the main arguments needed to derive a FIC for the locally stationary processes.

Theorem 9.3. *Let $y_1, \dots, y_n$ be realisations from the model (9.1) and let $P_\delta = P_{\delta, n}$ be the associated probability measure. Then if $A(u, \omega, \theta, \gamma)$ and $m(u, \theta, \gamma)$ satisfy conditions of Assumption 9.1*

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

where $\delta = (\delta_1, \delta_2)$ and J_{01} and J_{11} are block components of the matrix (9.2) similar to (3.8). Also, $\hat{J}_n(\bar{\theta}_n, \bar{\gamma}_n) = -\nabla^2 \ell_n(\bar{\theta}_n, \bar{\gamma}_n)/n \rightarrow_{P_\delta} J_{\text{wide}}$, provided $(\bar{\theta}_n, \bar{\gamma}_n) \rightarrow_{P_\delta} (\theta_0, \gamma_0)$.

Proof. The arguments needed are largely similar to those used in the proof of Theorem 2.4 in Dahlhaus (1996a). The main difference is the introduction of a local large-sample framework. Since most of the results needed, like Lemma 11.1, are uniform in the model parameters, this change of framework does not violate most of the original results, however.

The only part of the proof that we need to inspect with some more care is the expectation of Z_n in the local large-sample framework. For each $j = 1, \dots, p+q$ let

$$\begin{aligned} Z_{n,j} &= \frac{1}{\sqrt{n}} \nabla_j \ell_n(\theta_0, \gamma_0) = -\frac{\sqrt{n}}{2} \left\{ \frac{1}{n} \text{tr}\{\Sigma_0^{-1}[\nabla_j \Sigma_0]\} \right. \\ &\quad \left. - \frac{1}{n} (\mathbf{y}_n - \mathbf{m}_0)^t \Sigma_0^{-1}[\nabla_j \Sigma_0] \Sigma_0^{-1} (\mathbf{y}_n - \mathbf{m}_0) - \frac{1}{n} (\nabla_j \mathbf{m}_0)^t \Sigma_0^{-1} (\mathbf{y}_n - \mathbf{m}_0) \right\}. \end{aligned}$$

Next, in order to derive the expectation, we use that the true model is a local version of the narrow null model, meaning that $m_\delta = m_0 + \delta_2^t [\nabla_\gamma m_{\theta_0, \bar{\gamma}_n}]/\sqrt{n}$ and in addition that $A_\delta = A_0 + \delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}]/\sqrt{n}$, where $\|\bar{\gamma}_n - \gamma_0\| < \|\delta\|/\sqrt{n} < (\|\delta_1\| + \|\delta_2\|)/\sqrt{n}$. Since we have $\nabla_\gamma f_0(u, \omega) = \delta_1^t [\nabla_\gamma A_{\theta_0}(u, \omega)] A_{\theta_0}(u, -\omega) + \delta_1^t [\nabla_\gamma A_{\theta_0}(u, -\omega)] A_{\theta_0}(u, \omega)$ it follows

$$\begin{aligned} \Sigma_\delta &= \Sigma_n(A_\delta, A_\delta) = \Sigma_n(A_0, A_0) + \frac{1}{\sqrt{n}} \{ \Sigma_n(\delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}], A_0) + \Sigma_n(A_0, \delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}]) \} \\ &\quad + \frac{1}{n} \Sigma_n(\delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}], \delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}]) \\ &= \Sigma_0 + \frac{1}{\sqrt{n}} \Sigma_n(\delta_1^t \nabla_\gamma f_0) + \frac{1}{n} \Sigma_n(\delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}], \delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}]). \end{aligned}$$

With the current level of precision, the last term in the expansion of Σ_δ above can be neglected, and we obtain

$$\begin{aligned}
E_\delta Z_{n,j} &= -\frac{\sqrt{n}}{2} \left\{ \frac{1}{n} \text{tr}\{\Sigma_0^{-1}[\nabla_j \Sigma_0]\} - \frac{1}{n} \text{tr}\{\Sigma_0^{-1}[\nabla_j \Sigma_0] \Sigma_0^{-1} \Sigma_\delta\} \right. \\
&\quad \left. - \frac{1}{n} (\mathbf{m}_\delta - \mathbf{m}_0)^t \Sigma_0^{-1} [\nabla_j \Sigma_0] \Sigma_0^{-1} (\mathbf{m}_\delta - \mathbf{m}_0) - \frac{1}{n} [\nabla_j \mathbf{m}_0]^t \Sigma_0^{-1} (\mathbf{m}_\delta - \mathbf{m}_0) \right\} \\
&= \frac{1}{2} \left\{ \frac{1}{n} \text{tr}\{\Sigma_0^{-1}[\nabla_j \Sigma_0] \Sigma_0^{-1} (\Sigma_n(\delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}], A_0) + \Sigma_n(A_0, \delta_1^t [\nabla_\gamma A_{\theta_0, \bar{\gamma}_n}]))\} \right. \\
&\quad \left. + \frac{1}{n} [\nabla_j \mathbf{m}_0]^t \Sigma_0^{-1} (\delta_2^t \nabla_\gamma \mathbf{m}_{\theta_0, \bar{\gamma}_n}) \right\} + o(n^{-1/2}) \\
&= \frac{1}{n} \text{tr}\{\Sigma_0^{-1}[\nabla_j \Sigma_0] \Sigma_0^{-1} \Sigma_n(\delta_1^t [\nabla_\gamma f_{\theta_0, \bar{\gamma}_n}])\} + \frac{1}{n} (\nabla_j \mathbf{m}_0)^t \Sigma_0^{-1} (\delta_2^t [\nabla_\gamma \mathbf{m}_{\theta_0, \bar{\gamma}_n}]) + o(n^{-1/2}).
\end{aligned}$$

Since $f(u, \omega) = |A(u, \omega)|^2$ and we are working in a locally large-sample framework, it follows from Lemma 11.1 that

$$E_\delta Z_{n,j} \rightarrow \int_0^1 \int_{-\pi}^\pi \frac{\nabla_j f_0(u, \omega) [\delta_1^t [\nabla_\gamma f_0(u, \omega)]]}{f_0(u, \omega)^2} d\omega du + \int_0^1 \frac{\nabla_j m_0(u) [\delta_2^t [\nabla_\gamma m_0(u)]]}{f_0(u, 0)} du,$$

implying $E_\delta Z_n \rightarrow (J_{01} \delta_1, J_{11} \delta_2)^t$ as we needed to show. $\square$

Suppose $\mu = \mu(\theta, \gamma)$ depends on the underlying model only through the model parameters and that the derivatives, with respect to (θ, γ) , exist and are continuous in a neighbourhood of (θ_0, γ_0) . If $\hat{\mu}_S$, for $S \subset \{1, \dots, q\}$ and $q = q_1 + q_2$, represent an submodel estimator, it follows by similar arguments as used in Sections 3 and 8, that Theorem 9.3 provides the justification needed to establish a FIC for the locally stationary models by

$$\text{FIC}(S) = \hat{\tau}_0^2 + \hat{\nu}^t G_S \hat{Q} G_S^t \hat{\nu} + \hat{\nu}^t (I_q - \hat{G}_S) (D_n D_n^t - \hat{Q}) (I_q - \hat{G}_S) \hat{\nu}, \quad (9.3)$$

where the definitions and the corresponding estimates of ν , G , D_n and Q are analogues to that of Sections 3.2 and 8.1. Note that it is straightforward to extend the FIC for locally stationary processes, with respect to the focus functions of Sections 5 and 6. The results above also motivates a bias corrected FIC* formula, as in (3.13), and the model averaging and AFIC of Section 4 are also easily seen to be applicable in the current framework. Note that if the model parameters separate, we can obtain a similar representation of the FIC formula above to that in (8.6).

10. CONCLUDING REMARKS

A. Regression models with dependent errors. Consider the model

$$Y_t = m(x_t, \beta) + \epsilon_t, \quad \text{for } t = 1, \dots, n, \quad (10.1)$$

where x_t are d -dimensional covariates, $m(u, \beta)$ a function of known parametric structure and ϵ_t is a stationary Gaussian time series with zero mean. This is an important and general class of regression models with dependent errors, which includes the models of Section 8. The framework of this paper

does not cover models from this general class, however, and deriving the corresponding FIC will require new large-sample results, some of which are touched upon in Hermansen & Hjort (2014a). As discussed in Section 3.3, if we believe we know the structure of $m(u, \beta)$, we may instead detrend the observed series to make focused inference about the dependency, but in several cases it will be highly relevant to analyse the joint behaviour and interactions between trend and dependency.

B. Seasonality. Seasonal effects are commonly built into time series models, in that such phenomena occur naturally in the real world. A perhaps natural trend model capable of capturing such effects is $m_n(u, \theta, \gamma) = \theta_1 + \gamma_1 \cos(2\pi\gamma_2 un)$, but this model will lead to identifiability problems and result in a singular J_{wide} matrix. Therefore, consider instead

$$m_n(u, \theta, \gamma) = \theta_1 + \gamma_1 \cos(2\pi\alpha_1 un) + \gamma_2 \sin(2\pi\alpha_2 un) \quad (10.2)$$

where α_1 and α_2 are known constants. This model does not have the identifiability problems alluded to above and is still general enough to be useful. It is not clear whether these models will work in our framework, however, as the trend function will not have uniformly continuous derivatives in the limit.

To see how such sample size dependent trend functions m_n can be used in the current framework, note that the model conditions of Assumptions 8.2 (and 9.1) are stronger for the fitted than the underlying true model. Therefore, if m_n is independent of the sample size n at the null-point (θ_0, γ_0) , as in (10.2) with $\gamma = (\gamma_1, \gamma_2) = \gamma_0 = 0$, the fitted model used in the large sample arguments (i.e. the narrow model), will not cause any new problems.

Some additional work is needed to make the arguments work for the true $m_{\text{true}} = m_n(u, \theta_0, \gamma_0 + \delta/\sqrt{n})$. From the proof of Theorem 2.4 in Dahlhaus (1996a) (a central part of the FIC derivation of Section 9) it can be seen that there are two conditions that have to be worked with are; (i) $\|m_{\text{true}} - m_0\| = \|\mu_n(u)\|^2/n = o(n)$, which holds for most standard models, and (ii) the second result of Lemma 11.1 must hold with the sample size dependent trend function under the true model. In order to see what is needed to make proposition (ii) of Lemma 11.1 below to hold, we will write (with a slight misuse of the notation) $m_{\text{true}}(u) = m_0(u) + m_n(u)/\sqrt{n}$, where m_0 is assumed to satisfy the conditions of Assumption 8.2 and m_n is a function that depends on the sample size n , like the seasonal effect in (10.2). Then provided the maximal eigenvalue, say λ_n , of $A_n = \{\prod_{l \leq k} \Sigma_n(g_l)^{-1} \Sigma_n(f_l)\} \Sigma_n(g_k)^{-1}$ stays bounded for increasing n , it can be shown that $\mathbf{m}_{\text{true}}^t A_n \mathbf{m}_{\text{true}} = \mathbf{m}_0^t A_n \mathbf{m}_0 + O(n^{3/4})$, since $a_n^t A_n b_n \leq \|a_n\| \|b_n\| \lambda_n$, for all vectors $a_n, b_n \in \mathbb{R}^n$, making proposition (ii) of Lemma 11.1 in the Appendix essentially unchanged.

C. Change points. For the models of Sections 8 and 9 consider the prototype trend function $m(u) = \theta + \gamma I(u < u_0)$, for $0 \leq u_0 \leq 1$, representing a time series processes with change points or discontinuities. A focused selection strategy for such models would certainly be of interest, however, joint estimation of parameters that include a finite number of unknown discontinuities is rather

comprehensive and will require additional work, see for example Grønneberg et al. (2014) for change points models with i.i.d. observations.

In some situations the potential change points occur at known locations. In such cases it might be especially important to determine, with high precision, the effect of the change, as with γ above. Such models does not satisfy the conditions of Assumption 8.2, however, if the number of discontinuity points is finite, the argument leading up to the FIC formula will still be true.

To see this, we need to work through similar arguments as in the discussion of seasonality above, the only part requiring additional work is proposition (ii) of Lemma 11.1. This will be seen to hold since most of the underlying ‘action’ happens close to the main diagonal, meaning that essentially all off-diagonal behaviour will become negligible for large n . Let $m(u) = m_0(u) + m_1(u)I(u \leq u_1)$ for $0 \leq u_1 \leq 1$, and define again $A_n = \{\prod_{l \leq k} \Sigma_n(g_l)^{-1} \Sigma_n(f_l)\} \Sigma_n(g_k)^{-1}$, then by dividing A_n into blocks, corresponding to the discontinuity point u_1 , a updated version of proposition (ii) of Lemma 11.1 can be shown to hold by working trough all the different blocked elements and

$$(ii') \quad \frac{1}{n} \mathbf{m}_n^t A_n \mathbf{m}_n = \frac{1}{2\pi g_k(0)} \left\{ \prod_{l=1}^{k-1} \frac{f_l(0)}{g_l(0)} \right\} \int_0^{u_1} m(u)^2 du + O(n^{-2/3} \log^{2k+2} n).$$

The case with multiple and finite number of change points and/or different trend functions m_1 and m_2 can be handled very similarly.

D. The Whittle Approximation for time series models with trend. In Section 7 we worked with an alternative derivation and formulation of the FIC for the stationary processes. A similar derivation is also possible for the models of Section 8. The natural starting point is to work with the mean correct observations in the Whittle approximation formula (7.1), which then leads to

$$\tilde{\ell}_n(\theta, \gamma) = -\frac{n}{2} \left\{ \log 2\pi + \frac{1}{2\pi} \int_{-\pi}^{\pi} \log 2\pi f_{\theta, \gamma}(\omega) d\omega + \frac{1}{n} (\mathbf{y}_n - \mathbf{m}_{\theta, \gamma})^t \Sigma_n(\{(2\pi)^2 f_{\theta, \gamma}\}^{-1}) (\mathbf{y}_n - \mathbf{m}_{\theta, \gamma}) \right\}.$$

The corresponding score functions, evaluated at the null-point (θ_0, γ_0) , are then given by

$$\begin{pmatrix} \tilde{A}_n \\ \tilde{B}_n \end{pmatrix} = -\frac{1}{2n} \begin{pmatrix} (\mathbf{y}_n - \mathbf{m}_0)^t [\nabla \Sigma_n(\{(2\pi)^2 f_0\}^{-1})] (\mathbf{y}_n - \mathbf{m}_0) + \text{tr}\{\nabla \Sigma_n(\log f_0)\} \\ -2[\nabla \mathbf{m}_0]^t \Sigma_0^{-1} (\mathbf{y}_n - \mathbf{m}_0) \end{pmatrix},$$

and from this it is then quite straightforward to check that this will result in equivalent large-sample results as obtained in Proposition 8.3.

E. Parametric vs. nonparametric. The main concern of this paper has been with comparison of nested parametric models, however, nonparametric estimation of the spectral density using the periodogram in (7.1) or one of its smoothed cousins are quite common, see e.g. Brillinger (1975). An extension of the work presented here is to make focused inference where the goal is to selection among parametric and nonparametric candidates. This can be solved inside the framework of Jullum & Hjort (2014). An alternative approach is to build on the current local large-sample framework where the observations are assumed to be realisations from the model with true spectral density

$f_\delta(\omega) = f_0(\omega) + r(\omega)/\sqrt{n}$, where $f_0 = f_{\theta_0}$ is a standard type of parametric model. Such models are already extensively studied by others, e.g. Dzhangaridze (1986), making the extension potentially less cumbersome.

F. Introducing an additional, independent, set of realisations. In the discussion of the data-dependent focus functions, we assumed that these only depended on the last few elements of the complete series. If this is not the case, using the law of iterated expectations, or conditional mean squared error, will not necessarily lead to useful results. A solution that is often seen in the literature, is to sneak around this problem by introducing a second, independent, series of realisations from the same model. The general idea is to use one series for estimation, while the other is reserved for the testing the model, deriving the model selection criterion or constructing the focus function. This strategy is used in the derivation of FPE in Akaike (1969), it is also part of the toolkit in the discussion of order selection in Section 9.3 in Brockwell & Davis (1991) and is a key component of the work in Claeskens et al. (2007). Moreover, this strategy can easily be seen to be applicable in the current framework and we will not go into details. In practice, note that there exist some situations where we might have two independent realisations from the same process, see p. 361 in Claeskens et al. (2007), but it is mainly an abstract construction used to simplify the mathematics. Therefore, a common solution is to ‘cheat’ and use the same set of observations twice.

G. Spectral density with jump discontinuities. The spectral densities discussed in Sections 3–8 are all assumed to be without discontinuities, moreover, all of them are assumed to satisfy quite strong smoothness assumptions. The reason for this is partly that we wish to use the general results of Section 9 as a unified theory. A potential extension is therefore to work with the Whittle approximations of Section 7 and extend the current derivation to include spectral densities that have either known, or unknown, jump discontinuities, the latter one being the real challenge, see Taniguchi (2008) for an attempt to develop a systematic asymptotic theory and references therein for some real life applications.

11. APPENDIX

The following result is Lemma A.5 from Dahlhaus (1996a) and as an important remark, we point out that both results below holds uniformly in the model parameters.

Lemma 11.1. *Let $k \in \mathbb{N}$ and suppose A_l, B_l, C_l , for all $l = 1, \dots, k$, are transfer functions as defined in Section 9 that all fulfil condition (i) of Assumption 9.1, then*

$$\begin{aligned} (i) \quad & \frac{1}{n} \operatorname{tr} \left\{ \prod_{l=1}^k \Sigma_n(C_l, C_l)^{-1} \Sigma_n(A_l, B_l) \right\} \\ &= \frac{1}{2\pi} \int_0^1 \int_{-\pi}^{\pi} \left[\prod_{l=1}^k \frac{A_l(u, \omega) B_l(u, \omega)}{|C_l(u, \omega)|^2} \right] d\omega du + O(n^{-2/3} \log^{2k+2} n). \end{aligned}$$

Moreover, if m_1, m_2 are real valued functions defined on $[0, 1]$ that satisfies condition (ii) of Assumption 9.1, then

$$(ii) \frac{1}{n} \mathbf{m}_{1n}^t \left\{ \prod_{l=1}^{k-1} \Sigma_n(C_l, C_l)^{-1} \Sigma_n(A_l, B_l) \right\} \Sigma_n(C_k, C_k)^{-1} \mathbf{m}_{2n} \\ = \frac{1}{2\pi} \int_0^1 \left[\prod_{l=1}^{k-1} \frac{A_l(u, 0) B_l(u, 0)}{|C_l(u, 0)|^2} \right] |C_k(u, 0)|^{-2} m_1(u) m_2(u) du + O(n^{-2/3} \log^{2k+2} n),$$

where $\mathbf{m}_{i,n}^t = (m_i(1/n), \dots, m_i(1))$, for $i = 1, 2$.

It is straightforward to translate the results of Lemma 11.1 into a version that applies to standard stationary time series models. For $k \in \mathbb{N}$ let f_l and g_l , for $l = 1, \dots, k$, be sequences of spectral densities and let $A_l = f_l$, $B_l = 1$ and $C_l = \sqrt{g_l}$. Then under the assumptions of Lemma 11.1 (see Remark 9.2) it follows that

$$(i) \frac{1}{n} \text{tr} \left\{ \prod_{l=1}^k \Sigma_n(g_l)^{-1} \Sigma_n(f_l) \right\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \prod_{l=1}^k \frac{f_l(\omega)}{g_l(\omega)} \right\} d\omega + O(n^{-2/3} \log^{2k+2} n)$$

and

$$(ii) \frac{1}{n} \mathbf{m}_{1n}^t \left\{ \prod_{l=1}^{k-1} \Sigma_n(g_l)^{-1} \Sigma_n(f_l) \right\} \Sigma_n(g_k)^{-1} \mathbf{m}_{2n} \\ = \frac{1}{2\pi g_k(0)} \left\{ \prod_{l=1}^{k-1} \frac{f_l(0)}{g_l(0)} \right\} \int_0^1 m_1(u) m_2(u) du + O(n^{-2/3} \log^{2k+2} n).$$

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. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, vol. 1. 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.
- CLAESKENS, G., CROUX, C. & VAN KERCKHOVEN, J. (2007). Prediction-focused model selection for autoregressive models. *Australian & New Zealand Journal of Statistics* **49**, 359–379.
- 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.

- 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.
- DZHAPARIDZE, K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer.
- GRAY, R. (2006). *Toeplitz and Circulant Matrices: A review*. now publishers.
- 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.
- 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.
- HJORT, N. L. (2008). Focused information criteria for the linear hazard regression model. In *Statistical Models and Methods for Biomedical and Technical Systems*. Springer, pp. 487–502.
- HJORT, N. L. & CLAESKENS, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.
- HJORT, N. L. & CLAESKENS, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association* **101**, 1449–1464.
- JULLUM, M. & HJORT, N. L. (2014). Parametric or nonparametric: The fic approach. Tech. rep., University of Oslo and Norwegian Computing Centre.
- 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. D. R. & TSAI, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific.
- PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press.

- ROHAN, N. & RAMANATHAN, T. (2011). Order selection in ARMA models using the focused information criterion. *Australian & New Zealand Journal of Statistics* **53**, 217–231.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- TANIGUCHI, M. (2008). Non-regular estimation theory for piecewise continuous spectral densities. *Stochastic Processes and their Applications* **118**, 153–170.
- VAN BELLEGEM, S. & DAHLHAUS, R. (2006). Semiparametric estimation by model selection for locally stationary processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 721–746.
- 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`