Statistica Sinica **27** (2017), 951-981
doi:https://doi.org/10.5705/ss.202015.0364

PARAMETRIC OR NONPARAMETRIC: THE FIC APPROACH

Martin Jullum and Nils Lid Hjort

University of Oslo

Abstract: Should one rely on a parametric or nonparametric model when analysing a given data set? This classic question cannot be answered by traditional model selection criteria like AIC and BIC, since a nonparametric model has no likelihood. The purpose of the present paper is to develop a focused information criterion (FIC) for comparing general non-nested parametric models with a nonparametric alternative. It relies in part on the notion of a focus parameter, a population quantity of particular interest in the statistical analysis. The FIC compares and ranks candidate models based on estimated precision of the different model-based estimators for the focus parameter. It has earlier been developed for several classes of problems, but mainly involving parametric models. The new FIC, including also nonparametrics, is novel also in the mathematical context, being derived without the local neighbourhood asymptotics underlying previous versions of FIC. Certain average-weighted versions, called AFIC, allowing several focus parameters to be considered simultaneously, are also developed. We concentrate on the standard i.i.d. setting and certain direct extensions thereof, but also sketch further generalisations to other types of data. Theoretical and simulation-based results demonstrate desirable properties and satisfactory performance.

Key words and phrases: Asymptotic theory, focused information criterion, model selection.

1. Introduction

Statistical model selection is the task of choosing a good model for one's data. There is of course a vast literature on this broad topic, see e.g. methods surveyed in Claeskens and Hjort (2008). Methods are particularly plentiful when it comes to comparison and ranking of competing parametric models, e.g. for regression analyses, where AIC (the Akaike information criterion), BIC (the Bayesian information criterion), DIC (the deviance information criterion), TIC (the Takeuchi information criterion or model-robust AIC), and other information criteria are in frequent use. The idea of the focused information criterion (FIC) is to specifically address the quality of the final outcomes of a fitted model. This differs

from the ideas underlying the other information criteria pointed to, as these directly or indirectly assess general overall issues and goodness of fit aspects. The FIC starts by defining a population quantity of interest for the particular data analysis at hand, termed the *focus parameter*, as judged by context and scientific relevance. One then proceeds by estimating the mean squared error (or other risk aspects) of the different model estimates of this particular quantity. Thus, one is not merely saying 'this model is good for these data', but rather 'this model is good for these data when it comes to inference for this particular aspect of the distribution'. Examples of focus parameters are a quantile, the standard deviation, the interquartile range, the kurtosis, the probability that a data point will exceed a certain threshold value, or indeed most other statistically meaningful functionals mapping a distribution to a scalar. To avoid confusion, note in particular that the focus parameter need not be a specified parameter of the parametric distributions we handle.

The FIC and similar focused methods have been developed for nested parametric and regression models (Claeskens and Hjort (2003, 2008)), generalised additive partial linear models (Zhang and Liang (2011)), Cox's proportional hazards semiparametric regression model (Hjort and Claeskens (2006)), Aalen's linear hazards risk regression model (Hjort (2008)), certain semiparametric versions of the Aalen model (Gandy and Hjort (2013)), autoregressive time series model (Claeskens, Croux and Van Kerckhoven (2007)), as well as for certain applications in economics (Behl et al. (2012)), finance (Brownlees and Gallo (2008)) and fisheries science (Hermansen, Hjort and Kjesbu (2016)).

The above constructions and further developments of the FIC have mainly relied on estimating risk functions via large-sample results derived under a certain local misspecification framework (Claeskens and Hjort (2003, 2008)). In this framework the assumed true density or probability mass function gradually shrinks with the sample size (at rate $1/\sqrt{n}$) from a biggest 'wide' model, hitting a simplest 'narrow' model in the limit. This has proven to lead to useful risk approximations and then to model selection and model averaging methods, but carries certain inherent limitations. One of these is that the candidate models need to lie between these narrow and wide models. Another is that the resulting FIC apparatus, while effectively comparing nested parametric candidate models, cannot easily be used for comparison with nonparametric alternatives or other model formulations lying outside the widest of the parametric candidate models.

In the present paper we leave the local misspecification idea and work with a fixed true distribution of unknown form. Our purpose is to develop new FIC

methodology for comparing and ranking a set of (not necessarily nested) parametric models, jointly with a nonparametric alternative. This different framework causes the development of the FIC to pan out somewhat differently from that in the various papers pointed to above. We shall again reach large-sample results and exploit the ensuing risk approximations, but although we borrow ideas from the previous developments, we need new mathematics and derivations. The resulting methods hence involve different risk approximations and, in particular, lead to new FIC formulae. Note that when referring to 'the FIC' below, we mean the new FIC methodology developed in the present paper, unless otherwise stated.

1.1. The present FIC idea

To set the idea straight, consider i.i.d. observations $Y_1, \ldots, Y_n$ from some unknown distribution G. The task is to estimate a focus parameter $\mu = T(G)$, where T is a suitable functional which maps a distribution to a scalar. The most natural estimator class for μ uses $\widehat{\mu} = T(\widehat{G})$, with an estimator $\widehat{G}$ of the unknown distribution G. The question is however which of the several reasonable distribution estimators $\widehat{G}$ we should insert? As explained above we shall consider both parametric and nonparametric solutions.

The nonparametric solution is to use data directly without any structural assumption for the unknown distribution. In the i.i.d. setting this leads to $\widehat{\mu}_{\rm np} = T(\widehat{G}_n)$, with $\widehat{G}_n$ the empirical distribution function. Under weak regularity conditions (see Section 2) this estimator is asymptotically unbiased and has limiting distribution

$$\sqrt{n}(\widehat{\mu}_{\rm np} - \mu) \xrightarrow{d} {\rm N}(0, v_{\rm np}), \tag{1.1}$$

with the limiting variance $v_{\rm np}$ identified in Section 2.

Next consider parametric solutions. Working generically, let $f(\cdot\,;\theta)$ be a parametric family of density or probability mass functions, with θ belonging to some p-dimensional parameter space having $F(y;\theta) = F_\theta(y)$ as its corresponding distribution. Let $\widehat{\theta} = {\rm argmax}_\theta\, \ell_n(\theta)$ be the maximum likelihood estimator, with $\ell_n(\theta) = \sum_{i=1}^n \log f(Y_i;\theta)$ being the log-likelihood function of the data. We need to consider the behaviour outside model conditions, where generally no true parametric value $\theta_{\rm true}$ exists. For g the density or probability mass function associated with G, assume there is a unique minimiser θ_0 of the Kullback–Leibler divergence ${\rm KL}(g, f(\cdot\,;\theta)) = \int \log\{g(y)/f(y;\theta)\}\,{\rm d}G(y)$ from true model to parametric family. White (1982) was among the first to demonstrate that under

mild regularity conditions, the maximum likelihood estimator is consistent for this least false parameter value and $\sqrt{n}(\widehat{\theta}-\theta_0) \xrightarrow{d} \mathrm{N}_p(0,\Sigma)$ for an appropriate Σ. Note that if the model happens to be fully correct, with $G = F_{\theta_0}$, then $\theta_0 = \theta_{\mathrm{true}}$, and Σ simplifies to the inverse Fisher information matrix formula usually found in standard textbooks. Write now $s(\theta)$ for the more cumbersome $T(F_\theta) = T(F(\cdot\,;\theta))$, the focus parameter under the parametric model written as a function of θ. The parametric focus parameter estimator is $\widehat{\mu}_{\mathrm{pm}} = s(\widehat{\theta})$, aiming at the least false μ: $\mu_0 = s(\theta_0)$. A delta method argument yields

$$\sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_0) = \sqrt{n}\{s(\widehat{\theta}) - s(\theta_0)\} \xrightarrow{d} \mathrm{N}(0, v_{\mathrm{pm}}), \tag{1.2}$$

with an expression for the limiting variance v_{pm} given in Section 2.

Results (1.1) and (1.2) may now be rewritten as

$$\begin{aligned} \widehat{\mu}_{\mathrm{np}} &= \mu + \frac{v_{\mathrm{np}}^{1/2} Z_n}{\sqrt{n}} + o_{\mathrm{pr}}(n^{-1/2}), \\ \widehat{\mu}_{\mathrm{pm}} &= \mu + b + \frac{v_{\mathrm{pm}}^{1/2} Z_n^*}{\sqrt{n}} + o_{\mathrm{pr}}(n^{-1/2}), \end{aligned} \tag{1.3}$$

where $b = \mu_0 - \mu = s(\theta_0) - T(G)$ is the bias caused by using the best parametric approximation rather than the true distribution. Here Z_n and Z_n^* are variables tending to the standard normal in distribution. Relying on a squared error loss gives risks for the two estimators which decompose nicely into squared bias and variance. The expressions in (1.3) then yield first-order approximations of the risks of the nonparametric and parametric estimators:

$$\mathrm{mse}_{\mathrm{np}} = 0^2 + \frac{v_{\mathrm{np}}}{n} \quad \text{and} \quad \mathrm{mse}_{\mathrm{pm}} = b^2 + \frac{v_{\mathrm{pm}}}{n}. \tag{1.4}$$

Pedantically speaking, we have demonstrated that our nonparametric and parametric estimators are close in distribution to variables having these expressions as their exact mean squared errors, as opposed to showing that the exact mean squared errors of $\widehat{\mu}_{\mathrm{np}}$ and $\widehat{\mu}_{\mathrm{pm}}$ themselves converge. Our FIC strategy consists of estimating quantities like those in (1.4) – i.e. to obtain estimates for each of the possibly several different $\mathrm{mse}_{\mathrm{pm}}$, each corresponding to a different parametric model and family F_θ, least false parameter θ_0, and so on – in addition to the estimate of $\mathrm{mse}_{\mathrm{np}}$. The models are then ranked according to these estimates, coined the FIC scores, and the model (and corresponding estimator) with the smallest FIC score is selected.

As seen from (1.4), the selection between parametric and nonparametric then comes down to balancing squared bias and variance. Viewing the nonparametric model as an infinite-dimensional parametric model, it is natural that this estima-

tor has zero squared bias, but the largest variance. Indeed, including this candidate has the benefit that the winning model should always be reasonable: the nonparametric is selected when all parametric biases are large, while a parametric model is preferred when its bias is small enough. Despite this behaviour being typical for comparison of parametrics and nonparametrics, the literature on specifically comparing parametrics to nonparametrics is slim. Note that such comparison is e.g. not possible with the traditional AIC and the BIC methods as these rely on likelihoods. One recent approach is however Liu and Yang (2011). With Ω_M the set of candidate models, the authors attempt to classify model selection situations as either parametric ($G \in \Omega_M$) or nonparametric ($G \notin \Omega_M$), as they put it. The results are however used to attempt to choose between the AIC- and BIC-selector, and not to choose between parametric or nonparametric models as such.

When several focus parameters are of interest, a possible strategy is to apply the FIC repeatedly, for one focus parameter at a time, perhaps producing different rankings of the models and estimators. In many cases it might be more attractive to find *one* model appropriate for estimating all these focus parameters simultaneously; say not only the 0.9 quantile, but all quantiles from 0.85 to 0.95. Such an averaged focused information criterion (AFIC), where the focus parameters may be given different weights to reflect their relative importance, emerges by properly generalising the FIC idea (see Section 4).

1.2. Motivating illustration

Consider for motivation the data set of Roman era Egyptian life-lengths, a century before Christ, taken from Pearson (1902). This data set, from the very first volume of *Biometrika*, contains the age at death of 82 males (M) and 59 females (F). The numbers range from 1 to 96 and Pearson argues that they may be taken as a random sample from the better-living classes of the society at that time. Modelling these as two separate distributions, we define $\mu = \text{med(M)} - \text{med(F)}$, the difference in median life-lengths, as the focus parameter. We put up five different parametric model pairs (same parametric class for each gender) to compete with the nonparametric which uses the difference in the sample medians directly; see Table 1 for estimates and model selection results. The 'FIC plot' of Figure 1 visualises the results from applying the FIC scheme. Using the root of the FIC score does not alter the ranking of the candidate models, but lends itself better to interpretation as it brings the FIC score back to the scale of the observations. Perhaps surprisingly, the best model for estimating

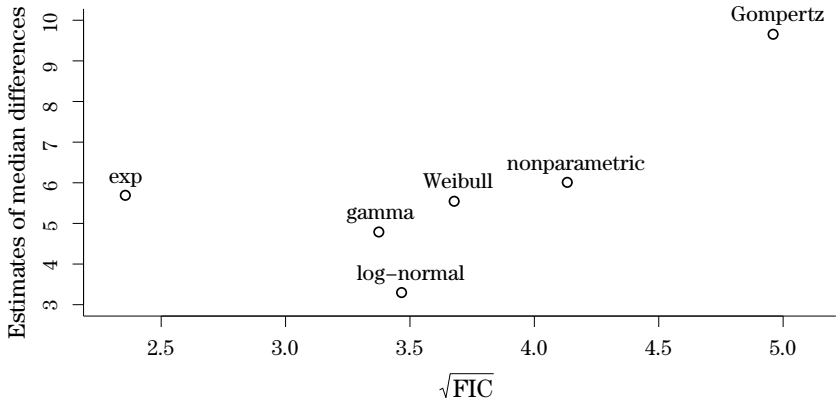

Figure 1. FIC plot for life lengths in Roman era Egypt: Six different estimates of $\mu = \text{med(M)} - \text{med(F)}$ are given, corresponding to five parametric models along with the direct nonparametric estimate. A small $\sqrt{\text{FIC}}$ indicates better precision.

μ turns out to be the simplest, using exponential distributions for both groups. This model does rather badly in terms of the overall performance measured by $\text{AIC(model)} = 2\{\ell_n(\widehat{\theta}) - p_{\text{model}}\}$, where p_{model} is the dimension of the model, but succeeds better than the others for the particular task of estimating μ, according to the FIC. This is possibly caused by two wrongs indeed making a right in this particular case. The exponential model seems to underestimate the median mortality by about the same amount for both males and females, implying that the difference has a small bias. Being the simplest model, this model also has small variance. For various other focus parameters, like apparently any difference in quantiles above 0.6, the exponential model will not be the winner. This is typical FIC behaviour; it is seldom that the same model is judged best for a wide range of focus parameters.

Supposing we are interested in right tail differences between the distributions, it is appropriate to apply the AFIC strategy mentioned above. One may for instance put up a focus parameter set consisting of differences in quantiles from 0.90 upwards to say 0.99 or 0.999. Almost no matter how the quantile differences are weighted, the Gompertz model is the clear winner, hence being deemed the overall best model for estimation of upper quantile differences. The exception is if a small neighbourhood around the 0.95 quantiles is deemed extremely important relative to the others, in which case the nonparametric is deemed better.

1.3. The present paper

In Section 2 we present precise conditions and derive asymptotic results used

Table 1. Life lengths in Roman era Egypt: The table displays estimates of median life length for male, female, and the difference, based on six different models; see also Figure 1. Also listed are the number of parameters estimated for each model, the AIC, the $\sqrt{\text{FIC}}$ scores, and the ranking based on FIC.

Model	$\widehat{\text{med}}(\text{M})$	$\widehat{\text{med}}(\text{F})$	$\widehat{\mu}$	dim	AIC	$\sqrt{\text{FIC}}$	FIC rank
nonparametric	28.000	22.000	6.000	Inf	NA	4.128	5
exponential	23.650	17.969	5.681	2	−1249.009	2.351	1
log−normal	23.237	19.958	3.279	4	−1262.809	3.447	3
gamma	26.655	21.877	4.778	4	−1232.129	3.370	2
Gompertz	31.783	22.139	9.644	4	−1224.776	4.955	6
Weibull	28.263	22.728	5.534	4	−1227.909	3.673	4

to build and illuminate the new FIC scheme for i.i.d. data. Section 3 discusses practical uses of the FIC, concentrating on focus parameters that are smooth functions of means and quantiles, but also discussing others. Section 4 then extends the framework to the case of addressing a set of focus parameters through the weighted or averaged focus information criterion (AFIC). Section 5 considers performance aspects of our methods. This covers analysis of limiting model selection probabilities (including asymptotic significance level of the FIC when viewed as an implicit focused goodness-of-model test procedure), summaries of simulation results, and asymptotic comparisons of the new and original FIC. Section 6 considers model averaging aspects and Section 7 discusses extensions of the present FIC approach to other data types and frameworks, including that of density estimation and regression. Finally, some concluding remarks are offered in Section 8. Supplementary material containing proofs of results in the main paper, another illustration, details of simulations studies, details on FIC and AFIC for categorical data, and some local asymptotics results, are given in Jullum and Hjort (2017a).

2. Derivation of the FIC

To develop the estimators of v_{np}/n and $b^2 + v_{\text{pm}}/n$ in (1.4), it is fruitful to start from a general result pertaining to the joint limit distribution of the right hand side of (1.1) and (1.2). We first introduce some helpful quantities and a set of working conditions for the parametric models. For convenience we shall write $\partial h(x_0)/\partial x$ for the derivative of a function h with respect to x, evaluated at $x = x_0$. Let $u(y;\theta)$, $I(y;\theta)$ and $\dot{I}(y;\theta)$ denote, respectively, the first, second, and third derivatives of $\log f(y;\theta)$ w.r.t. θ. Let also

$$J = -\text{E}_G\{I(Y_i;\theta_0)\} \quad \text{and} \quad K = \text{E}_G\{u(Y_i;\theta_0)u(Y_i;\theta_0)^{\text{t}}\}. \tag{2.1}$$

Though weaker regularity conditions are possible, see e.g. Hjort and Pollard (1993) and van der Vaart (2000, Chap. 5), we work under reasonably traditional maximum likelihood conditions:

(C0) The support of the model is independent of θ; the minimiser θ_0 of $\mathrm{KL}(g, f(\cdot\,; \theta))$ is unique and lies in an open subset Θ of Euclidean space; $u(y;\theta)$, $I(y;\theta)$, and $\dot{I}(y;\theta)$ exist and the latter is continuous for every y; J, J^{-1}, and K are finite; and all elements of $\dot{I}(y;\theta)$ are bounded by an integrable function in a neighbourhood of $\theta = \theta_0$.

These conditions will be assumed without further discussion. We turn to the focus parameter. For a statistical functional $T(\cdot)$ taking distributions H as its input, the influence function (Huber (1981); Hampel et al. (1986)) is defined as

$$\mathrm{IF}(y; H) = \frac{\lim_{t\to 0}\{T_{H,y}(t) - T(H)\}}{t} = \frac{\partial T_{H,y}(0)}{\partial t}, \tag{2.2}$$

where $T_{H,y}(t) = T((1-t)H + t\delta_y)$ and δ_y is the distribution with unit point mass at y.

We need the general notion of Hadamard differentiability. For general normed spaces $\mathbb{D}$ and $\mathbb{E}$, a map $\phi\colon \mathbb{D}_\phi \mapsto \mathbb{E}$, defined on a subset $\mathbb{D}_\phi \subseteq \mathbb{D}$ that contains β, is called Hadamard differentiable at β if there exists a continuous, linear map $\dot{\phi}\colon \mathbb{D} \mapsto \mathbb{E}$ (called the derivative of ϕ at β) such that $\|\{\phi(\beta + th_t) - \phi(\beta)\}/t - \dot{\phi}_\beta(h)\|_{\mathbb{E}} \to 0$ as $t \searrow 0$ for every $h_t \to h$ such that $\beta + th_t$ is contained in $\mathbb{D}_\phi$. (We have avoided introducing the notation of Hadamard differentiability tangentially to a subset of $\mathbb{D}$, as such are better stated explicitly in our practical cases.) With this definition and by recalling that $s(\theta) = T(F_\theta)$, we define the regularity conditions:

(C1) The focus functional T is Hadamard differentiable at G with respect to the uniform norm $\|\cdot\|_\infty = \sup_y |\cdot|$;
(C2) $c = \partial s(\theta_0)/\partial\theta$ is finite;
(C3) $\mathrm{E}_G\{\mathrm{IF}(Y_i; G)\} = 0$;
(C4) $\mathrm{E}_G\{\mathrm{IF}(Y_i; G)^2\}$ is finite.

The three first conditions here can typically be checked. The last one can usually be narrowed down to a simpler assumption for G which then needs to be assumed. It should be noted that (C2), along with a few of the conditions in (C0), follows by instead assuming that the parametric focus functional $S(\cdot)$ with $S(G) = T(F(\cdot\,; R(G))) = \mu_0$, where $R(G) = \mathrm{argmin}_\theta\, \mathrm{KL}(g, f_\theta)$ and $S(\widehat{G}_n) = \widehat{\mu}_{\mathrm{pm}}$, is Hadamard differentiable as well. We will however stick to the more natural and direct conditions above.

Proposition 1. *Under* (C1–C4)*, we have that, as* $n \to \infty$*,*

$$\begin{pmatrix} \sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu) \\ \sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_0) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c^{\mathrm{t}} J^{-1} U \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} v_{\mathrm{np}} & v_{\mathrm{c}} \\ v_{\mathrm{c}} & v_{\mathrm{pm}} \end{pmatrix} \right). \tag{2.3}$$

Here (X, U) *is jointly zero-mean normal with dimensions 1 and p, having variances* $v_{\mathrm{np}} = \mathrm{E}_G\{\mathrm{IF}(Y_i; G)^2\}$ *and* K*, and covariance* $d = \mathrm{E}_G(XU) = \int \mathrm{IF}(y; G)\, u(y; \theta_0)\, \mathrm{d}G(y)$*. In particular,* $v_{\mathrm{pm}} = c^{\mathrm{t}} J^{-1} K J^{-1} c$ *and* $v_c = c^{\mathrm{t}} J^{-1} d$*.*

Proof. Under the assumed conditions it follows from, respectively, Shao (2003, Thm. 5.5) and van der Vaart (2000, Thms 5.41 and 5.42) that

$$\widehat{\mu}_{\mathrm{np}} - \mu = \frac{1}{n} \sum_{i=1}^{n} \mathrm{IF}(Y_i; G) + o_{\mathrm{pr}}(n^{-1/2}), \tag{2.4}$$

$$\widehat{\theta} - \theta_0 = \frac{1}{n} \sum_{i=1}^{n} J^{-1} u(Y_i; \theta_0) + o_{\mathrm{pr}}(n^{-1/2}). \tag{2.5}$$

Applying the standard central limit theorem to the summands, followed by application of the delta method using $h(x, y) = \{x, s(y)\}^{\mathrm{t}}$ as transformation function, yields the limit distribution. Slutsky's theorem takes care of the remainders and completes the proof.

Since the variance matrix of (X, U) is nonnegative definite, we always have $v_{\mathrm{np}} \geq d^{\mathrm{t}} K^{-1} d$. When a parametric model is true, with $G = F_{\theta_0}$ (and a few other mild regularity conditions hold, cf. Corollary 1) we get $J = K$ and $c = d$, implying $v_{\mathrm{np}} \geq c^{\mathrm{t}} J^{-1} c = v_{\mathrm{pm}}$. In this case we also have $v_c = v_{\mathrm{pm}}$ and the limiting correlation $(v_{\mathrm{pm}}/v_{\mathrm{np}})^{1/2}$ between the two estimators.

Sometimes we do not need (C1–C4) directly, only that (2.3) holds. This is beneficial as the above result is sometimes as easily shown on a 'case by case' basis without going through the sets of conditions. The parametric smoothness of (C0) can indeed be relaxed with (2.5) still holding; see e.g. van der Vaart (2000, Thm. 5.39). There are also situations where T is not Hadamard differentiable, but where (2.5) still holds. Some of these situations may be handled by the concept of Fréchet differentiability. By Shao (2003, Thm. 5.5(ii)), (2.5) still holds if we replace the Hadamard differentiability condition with Fréchet differentiability equipped with a norm $\|\cdot\|_0$ possessing the property that $\|\widehat{G}_n - G\|_0 = O_{\mathrm{pr}}(n^{-1/2})$. As Fréchet differentiability is stronger than Hadamard differentiability, this is however only useful with other norms than the uniform. Note also that Proposition 1, through the conditions (C1–C4), is restricted to $\sqrt{n}$-consistent estimators. Hence, estimators with other convergence rates need to be handled separately. Density estimation and regression, as examples of this, are discussed in Sections

7.4 and 7.5.

Consider now estimation of the quantities in (1.3). For $\text{mse}_{\text{np}} = v_{\text{np}}/n$, we use $\widehat{\text{mse}}_{\text{np}} = \widehat{v}_{\text{np}}/n$, with $\widehat{v}_{\text{np}}$ some appropriate and consistent estimator of v_{np}. For a large class of functionals, a good general recipe is to use the empirical analogue $\widehat{v}_{\text{np}} = n^{-1}\sum_{i=1}^n \text{IF}(Y_i;\widehat{G}_n)^2$. Consider now $\text{mse}_{\text{pm}} = b^2 + v_{\text{pm}}/n$, where estimation of $v_{\text{pm}} = c^{\text{t}}J^{-1}KJ^{-1}c$ is obtained by taking the empirical analogues and inserting $\widehat{\theta}$ for θ_0. For J and K of (2.1) we use $\widehat{J} = -n^{-1}\sum_{i=1}^n I(Y_i;\widehat{\theta})$ and $\widehat{K} = n^{-1}\sum_{i=1}^n u(Y_i;\widehat{\theta})u(Y_i;\widehat{\theta})^{\text{t}}$. The first matrix here is n^{-1} times the Hessian matrix typically computed in connection with numerically finding the maximum likelihood estimates, and the second matrix is also easy to compute. For $c = \partial s(\theta_0)/\partial\theta$ we also employ plug-in and use $\widehat{c} = \partial s(\widehat{\theta})/\partial\theta$, which is easy to compute numerically in cases where there is no closed form expression for the derivatives. We thus get $\widehat{v}_{\text{pm}} = \widehat{c}^{\text{t}}\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}\widehat{c}$. These plug-in variance estimators are the canonical choices. They are of the usual $\sqrt{n}$ precision order, in the sense that they under mild regularity assumptions have the property that both $\sqrt{n}(\widehat{v}_{\text{np}}-v_{\text{np}})$ and $\sqrt{n}(\widehat{v}_{\text{pm}}-v_{\text{pm}})$ converge to zero-mean normal limit distributions. We do not need such results or these limits in what follows.

For the square of the bias $b = \mu_0 - \mu = s(\theta_0) - \mu(G)$ we start from $\widehat{b} = s(\widehat{\theta}) - \mu(\widehat{G}_n) = \widehat{\mu}_{\text{pm}} - \widehat{\mu}_{\text{np}}$. When the conclusion of Proposition 1 holds, we get

$$\sqrt{n}(\widehat{b}-b) \xrightarrow{d} c^{\text{t}}J^{-1}U - X \sim \text{N}(0,\kappa), \tag{2.6}$$

where $\kappa = v_{\text{pm}} + v_{\text{np}} - 2v_{\text{c}}$. Here $v_{\text{c}} = c^{\text{t}}J^{-1}d$ is estimated by plugging in $\widehat{c}$ and $\widehat{J}$ as already given. Analogously, we estimate d by $\widehat{d} = n^{-1}\sum_{i=1}^n \text{IF}(Y_i;\widehat{G}_n)u(Y_i;\widehat{\theta})$. Though $\widehat{b}$ is approximately unbiased for b, its square will typically tend to overestimate b^2, since $\text{E}_G(\widehat{b}^2) = b^2+\kappa/n+o(n^{-1})$. Hence, we are led to the modification $\widehat{\text{bsq}}_0 = \widehat{b}^2 - \widehat{\kappa}/n$, where $\widehat{\kappa} = \widehat{v}_{\text{pm}} + \widehat{v}_{\text{np}} - 2\widehat{v}_{\text{c}}$. The estimator is unbiased up to order n^{-1} for the squared bias. Accuracy to this order is important to capture, in that the other half of the mse coin is the variance term, which is precisely of this size. An appropriate further modification is

$$\widehat{\text{bsq}} = \max(0,\widehat{\text{bsq}}_0) = \max(0,\widehat{b}^2 - \frac{\widehat{\kappa}}{n}), \tag{2.7}$$

truncating negative estimates of the obviously nonnegative b^2 to zero.

Having established estimates for the nonparametric and parametric first order mean squared error approximations in (1.4), we define the FIC scores:

$$\begin{aligned}\text{FIC}_{\text{np}} &= \widehat{\text{mse}}_{\text{pm}} = \frac{\widehat{v}_{\text{np}}}{n},\\ \text{FIC}_{\text{pm}} &= \widehat{\text{mse}}_{\text{pm}} = \widehat{\text{bsq}} + \frac{\widehat{v}_{\text{pm}}}{n} = \max(0,\widehat{b}^2 - \frac{\widehat{\kappa}}{n}) + \frac{\widehat{v}_{\text{pm}}}{n}.\end{aligned} \tag{2.8}$$

Even though the above formula for $\mathrm{FIC}_{\mathrm{pm}}$ is our canonical choice, the non-truncated version $\mathrm{FIC}^*_{\mathrm{pm}} = \widehat{\mathrm{bsq}}_0 + \widehat{v}_{\mathrm{pm}}/n$ may be used on occasion. At any rate, $\mathrm{FIC}_{\mathrm{pm}}$ and $\mathrm{FIC}^*_{\mathrm{pm}}$ agree when $\widehat{b}^2 \geq \widehat{\kappa}/n$, which happens with a probability tending to one for the realistic case of $b \neq 0$. When they agree we can also express them as $\mathrm{FIC}_{\mathrm{pm}} = \mathrm{FIC}^*_{\mathrm{pm}} = \widehat{b}^2 - \widehat{v}_{\mathrm{np}}/n + 2\widehat{v}_{\mathrm{c}}/n$.

The focused model selection strategy is now clear. Given a set consisting of k parametric candidate models, in addition to the nonparametric one, one computes the focus parameter estimates along with the FIC scores based on (2.8). The same formula (with different estimates and quantities for different parametric families) can be used for all parametric candidates as $\mathrm{FIC}_{\mathrm{pm}}$ is independent of any other competing parametric models. One selects the model and estimator associated with the smallest FIC score. In particular this setup can also be used when only parametric models are under consideration, and these parametric models can be non-nested. In such a case, one is still required to compute the nonparametric estimator and its variance, since $\mathrm{FIC}_{\mathrm{pm}}$ depends on them.

As the sample size increases, the variance part of (2.8) typically becomes negligible compared to the squared bias. Hence, consistency of the FIC scores as MSE estimators is not very informative. It is more illuminating to consider consistency of the scaled variance and squared bias involved in the FIC scores separately. Under (C2), the parametric focus functional $S(\cdot)$ does indeed have influence function $c^{\mathrm{t}}J^{-1}u(y;\theta_0)$, which appears in (2.5) and is being estimated (using $\widehat{G}_n$ for G) by $\widehat{c}^{\mathrm{t}}\widehat{J}^{-1}u(y;\widehat{\theta})$. By working directly on the focus functionals T and S, general conditions for consistency of plug-in based variance and bias estimators can be worked out. This involves Gâteaux differentiability, a weaker form of functional differentiability where the convergence in the definition of Hadamard differentiability only needs to hold for every fixed $h_t = h$, see e.g. van der Vaart (2000, Chap. 20.2).

Proposition 2. *Suppose T and S are Gâteaux differentiable at G and $\widehat{G}_n$. Suppose $\sup_{|y|\leq k} |\mathrm{IF}(y;\widehat{G}_n) - \mathrm{IF}(y;G)| = o_{\mathrm{pr}}(1)$ for any $k > 0$, that there exist a constant $k_0 > 0$ and a function $r(y) \geq 0$ such that $r(Y_i)$ has finite mean and $\Pr_G(\mathrm{IF}(y;\widehat{G}_n)^2 \leq r(y)$ for all $y \geq k_0) \to 1$, and that an equivalent assumption holds for the influence function of S. Then the variance and covariance estimators $\widehat{v}_{\mathrm{np}}$, $\widehat{v}_{\mathrm{pm}}$, $\widehat{\kappa}$, $\widehat{v}_{\mathrm{c}}$ are all consistent when being based solely on plugging in $\widehat{G}_n$ for G and $\widehat{\theta}$ for θ_0 in their asymptotic analogues. If in addition the conclusion of Proposition 1 holds, then the bias terms $\widehat{b}$ (and $\widehat{b}^2$) are consistent.*

The proofs of Proposition 2 and subsequent results are given in the supple-

mentary material (Jullum and Hjort (2017a)). Most natural estimators of the variance etc. are based on direct plug-in estimation, but there are exceptions. Making minor adjustments to them do not change the asymptotics, so using them in FIC formulae is generally unproblematic.

For estimating the uncertainty of an estimator, other options are also available. The bootstrap and the jackknife (Efron and Tibshirani (1993)) can be applied to estimate the variance of most statistics dealt with here. The squared bias is also fairly easy to estimate for the nonparametric candidate. A possible approach to estimate the squared bias is to use the jackknife or bootstrap only to correct the initial natural squared bias estimate $\widehat{b}^2 = (\widehat{\mu}_{\rm pm} - \widehat{\mu}_{\rm np})^2$. This, combined with jackknife or bootstrap estimates of the variance, would essentially lead to the same FIC scores as those in (2.8). From this perspective some of the explicit formulae from our asymptotic theory in this section could be by-passed. It is however a strength of our approach to have explicit formulae, for computation and performance analyses. In Section 5 we derive informative properties and performance based on this theoretical framework. Even though similar results may perhaps be obtained also for specific types of bootstrap or jackknife frameworks under certain conditions (Shao and Tu (1996)), they would not be as readily and generally available. The usual jackknife estimate of the variance of a quantile, for example, is known to be inconsistent and needs to be corrected (Efron (1979)). Finally, since the (Monte Carlo) bootstrap possesses sampling variability, models with similar FIC scores may require an inconveniently large number of bootstrap samples to have their FIC scores separated with satisfactory certainty. Thus, we leave the jackknife and bootstrap idea as alternative estimation procedures, and in what follows we stick to the FIC formulae we have derived, based on large-sample theory and the plug-in principle.

3. FIC in Practice

In this section we introduce a wide class of natural focus parameters and show that they fit our scheme. We also show that other types of focus parameters may be used, widening the applicability with a few direct extensions.

3.1. Smooth functions of means and quantiles

Most of the functionals that interest us belong to the class we refer to as 'smooth functions of means and quantiles'. This class has the functional form

$$T(G) = A(\xi, \zeta) = A(\xi_1, \ldots, \xi_k, \zeta_1, \ldots, \zeta_m), \tag{3.1}$$

where $\xi_j = \mathrm{E}_G\{h_j(Y_i)\} = \int h_j(y)\,\mathrm{d}G(y)$ and $\zeta_l = G^{-1}(p_l)$ for one-dimensional functions h_j and $p_l \in (0,1)$. Here $A\colon \mathbb{R}^{k+m} \mapsto \mathbb{R}$ is a smooth function, i.e. continuously differentiable at the evaluation points. Some interesting functionals are smooth functions of means only. The standard deviation, skewness, and kurtosis functionals are of this type, for example, with $k = 2, 3, 4$, respectively. Another example is the probability of observing a variable in a specified interval: Let A be the identity function, $k = 1$, and h_1 be the indicator function for this interval. Functionals based solely on quantiles are also of interest. Any quantile function $G^{-1}(p)$ is within this class, along with functions of quantiles like the interquartile and interdecile range, and the midhinge (the average of the first and third quartile). The nonparametric skew (mean − median)/sd (Hotelling and Solomons (1932)) involves both means and a quantile, and may be handled by our scheme.

For this full class, the nonparametric estimator is $\widehat{\mu}_{\mathrm{np}} = A(\overline{h}, \widehat{\zeta})$ where $\overline{h}$ and $\widehat{\zeta}$ have elements $\overline{h}_j = n^{-1}\sum_{i=1}^n h_j(Y_i)$ and $\widehat{\zeta}_l = \widehat{G}_n^{-1}(p_l)$, for $j = 1, \ldots, k$ and $l = 1, \ldots, m$. Similarly, the parametric estimators are of the form $\widehat{\mu}_{\mathrm{pm}} = A(\xi(\widehat{\theta}), \zeta(\widehat{\theta}))$, where $\xi(\theta)$ and $\zeta(\theta)$ have elements $\xi_j(\theta) = \int h_j(y)\,\mathrm{d}F_\theta(y)$ and $\zeta_l(\theta) = F_\theta^{-1}(p_l)$ for $j = 1, \ldots, k$ and $l = 1, \ldots, m$. For this class, the influence function is given by

$$\mathrm{IF}(y; G) = \sum_{j=1}^{k} a_{0,j}\{h_j(y) - \xi_j\} + \sum_{l=1}^{m} a_{1,l}\frac{p_l - \mathbf{1}_{\{y \le G^{-1}(p_l)\}}}{g(G^{-1}(p_l))}, \tag{3.2}$$

where $a_{0,j} = \partial A(\xi, \zeta)/\partial \xi_j, a_{1,l} = \partial A(\xi, \zeta)/\partial \zeta_l$ and $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. While the part of the influence function related to the means is easily estimated by plugging in $\widehat{G}_n$ for G and hence replacing ξ_j and ξ by $\overline{h}_j$ and $\overline{h}$, the part relating to the quantiles is more delicate. By Proposition 2, we need consistent estimators for quantities of the form $g(G^{-1}(p))$. Such can be most easily constructed using $\widehat{g}_n(\widetilde{G}_n^{-1}(p))$, say, involving a kernel density estimator for g and a possibly smoothed version $\widetilde{G}_n$ of the empirical distribution function $\widehat{G}_n$ for G. Details securing consistency with the right sequence of bandwidths are found in e.g. Sheather and Marron (1990).

The following proposition shows that the class of smooth functions of means and quantiles are applicable to our scheme.

Proposition 3. *Let $\mu = T(G)$ be of the form* (3.1)*, with all partial derivatives of $A(\xi, \zeta)$ finite and g positive and continuous in all $G^{-1}(p_l), l = 1, \ldots, m$. Then* (C1) *and* (C3) *hold. If all partial derivatives of $\xi(\theta_0)$ and $\zeta(\theta_0)$ are finite, and $\mathrm{E}_G\{h_j(Y_i)^2\}$ is finite for all $j = 1, \ldots, k$, then also* (C2) *and* (C4) *hold, and Proposition 1 is in force.*

As an illustration, suppose data $Y_1, \ldots, Y_n$ are observed on the positive half-line, and the skewness $\gamma = \mathrm{E}_G\{(Y_i - M_1)^3/\sigma^3\}$ of the underlying distribution needs to be estimated; here $M_j = \mathrm{E}_G(Y_i^j)$ and σ are the j-th moment and the standard deviation, respectively. This is a smooth function of means only, as

$$\gamma = h(M_1, M_2, M_3) = \frac{(M_3 - 3M_1M_2^2 + 2M_1^3)}{(M_2 - M_1^2)^{3/2}}.$$

The nonparametric estimate is $\widehat{\gamma}_{\mathrm{np}} = h(\widehat{M}_1, \widehat{M}_2, \widehat{M}_3)$, involving the averages of Y_i, Y_i^2, Y_i^3, and has the FIC score

$$\mathrm{FIC}_{\mathrm{np}} = \frac{\widehat{v}_{\mathrm{np}}}{n}, \quad \widehat{v}_{\mathrm{np}} = \frac{1}{n}\sum_{i=1}^{n}\{\widehat{k}_1(Y_i - \overline{M}_1) + \widehat{k}_2(Y_i^2 - \overline{M}_2) + \widehat{k}_3(Y_i^3 - \overline{M}_3)\}^2,$$

in terms of certain coefficient estimates $\widehat{k}_1, \widehat{k}_2, \widehat{k}_3$. A simple parametric alternative fits the Gamma distribution with density $\{\beta^\alpha/\Gamma(\alpha)\}y^{\alpha-1}\exp(-\beta y)$, for which the skewness is $2/\alpha^{1/2}$. With FIC it is easy to determine, for a given data set, whether the best estimate of the skewness is the nonparametric one or the simpler and less variable parametric $2/\widehat{\alpha}^{1/2}$. Thus, one is not necessarily interested in how well the gamma model fits the data overall, but concentrates on judging whether $2/\widehat{\alpha}^{1/2}$ is a good estimator or not for the skewness, completely ignoring $\widehat{\beta}$.

To learn which models are good for modelling different aspects of a distribution, one approach consults the FIC scores obtained when the FIC is sequentially applied to all focus parameters in a suitable set. One may for instance run the FIC through the c.d.f. by sequentially considering each focus parameter of the form $\mu(y) = G(y)$ for $y \in \mathbb{R}$, or the quantile function $\mu(p) = G^{-1}(p)$ for each $p \in (0,1)$. One may similarly run the FIC through a set of moments or cumulants or, say, the moment-generating function $\mu(t) = \mathrm{E}_G\{\exp(tY_i)\}$ for t close to zero. The supplementary material (Jullum and Hjort (2017a)) provides an illustration of this concept by running through the c.d.f. for a data set with SAT writing scores.

3.2. Other types of focus parameters and data frameworks

The class of smooth functions of means and quantiles is as mentioned a widely applicable class, but there are also focus parameters outside this class that are Hadamard differentiable and thus fit our scheme. In this regard, the chain rule for Hadamard differentiability (van der Vaart (2000, Thm. 20.9)) is helpful. The median absolute deviation (MAD) is the median of the absolute deviation from the

median of the data: $\text{med}(|\,\text{med}(G) - G|)$. This can be written in functional form as $T(G) = H_G^{-1}(1/2)$, where $H_G(x) = G(\nu + x) - G((\nu - x)-)$ and $\nu = G^{-1}(1/2)$. This functional has a precise influence function and, under some conditions, Hadamard differentiability is ensured by van der Vaart (2000, Thm. 21.9). The trimmed mean, in functional form $T(G) = (1-2\alpha)^{-1}\int_\alpha^{1-\alpha} G^{-1}(y)\,\mathrm{d}y$ is, under similar conditions, also Hadamard differentiable (see van der Vaart (2000, Exmp. 22.11)).

We concentrate on univariate observations, but the derivations of Section 2, and especially the joint limiting distribution in (2.3) and the FIC formulae in (2.8), hold also in the more general case of multivariate i.i.d. data.

The Egyptian life-time data in the introduction were not of the standard i.i.d. type that has been dealt with here. There were *two* samples or populations, and the focus parameter was $\mu = \text{med(M)} - \text{med(F)}$. This is handled by a simple extension of the criterion. Consider more generally a focus parameter of the form $\mu = T_1(G_1) - T_2(G_2)$ for individual functionals T_1, T_2 defined for different samples or populations G_1 and G_2. If there is no interaction between the two distributions, μ is naturally estimated by $\widehat{\mu} = \widehat{\mu}_1 - \widehat{\mu}_2 = T_1(\widehat{G}_1) - T_2(\widehat{G}_2)$ and has mean squared error $\text{mse}(\widehat{\mu}) = \text{E}_G[\{(\widehat{\mu}_1 - \widehat{\mu}_2) - (\mu_1 - \mu_2)\}^2]$. Here this is estimated by $\text{FIC}_{\text{pm}} = \max\{0, (\widehat{b}_1 - \widehat{b}_2)^2 - \widehat{\kappa}_1/n_1 - \widehat{\kappa}_2/n_2\} + \widehat{v}_{\text{pm},1}/n_1 + \widehat{v}_{\text{pm},2}/n_2$ for parametric models. When one or both of the models are nonparametric, the formula is the same, except $\widehat{b}_j$ and $\widehat{\kappa}_j$ are set to zero for j nonparametric, and $\widehat{v}_{\text{pm},j}$ is replaced by $\widehat{v}_{\text{np},j}$. One can model G_1 and G_2 differently and mix parametrics and nonparametrics. When data are of a similar type it is however natural to only consider pairs of the same model type. Similar schemes can be established for comparisons of more than two samples, products of focus parameters, etc.

4. Weighted FIC

The above apparatus is geared towards optimal selection for a single μ. One is often interested in more than one parameter simultaneously, however, say not merely the median but also other quantiles, or several probabilities $\Pr(Y_i \in A)$ for an ensemble of A sets. We develop a suitable average weighted focused information criterion AFIC that selects *one* model aimed at estimating the whole set of weighted focus parameters with lowest risk (cf. Claeskens and Hjort (2003, Sec. 7) and Claeskens and Hjort (2008, Chap. 6)). Suppose these focus parameters are listed as $\mu(t)$, for t in some index set. Thus, there is a nonparametric

estimator $\widehat{\mu}_{\rm np}(t)$ and one or more parametric estimators $\widehat{\mu}_{\rm pm}(t)$ for each $\mu(t)$. As an overall loss measure when estimating the $\mu(t)$ with $\widehat{\mu}(t)$, we use

$$L = \int \{\widehat{\mu}(t) - \mu(t)\}^2 \, {\rm d}W(t),$$

with W some cumulative weight function, chosen to reflect the relative importance of the different $\mu(t)$. The risk or expected loss may hence be expressed as

$$\text{risk} = {\rm E}_G(L) = \int \text{mse}(t) \, {\rm d}W(t),$$

with $\text{mse}(t) = {\rm E}_G\{\widehat{\mu}(t) - \mu(t)\}^2$. Our previous results leading to both the joint limit (2.3) and our FIC scores (2.8) hold for each such $\mu(t)$ with the same quantities, indexed by t, under the same set of conditions. To estimate the risk we can plug-in FIC scores as estimates of $\text{mse}(t)$, but choose to truncate the squared bias generalisation to zero after integration as we are no longer seeking natural estimates for the individual mses, but for the integrated risk. This leads us to the weighted or averaged FIC (AFIC) scores

$$\begin{aligned} \text{AFIC}_{\rm np} &= \frac{1}{n} \int \widehat{v}_{\rm np}(t) \, {\rm d}W(t), \\ \text{AFIC}_{\rm pm} &= \max\left[0, \int \left\{ \widehat{b}(t)^2 - \frac{\widehat{\kappa}(t)}{n} \right\} {\rm d}W(t) \right] + \frac{1}{n} \int \widehat{v}_{\rm pm}(t) \, {\rm d}W(t). \end{aligned} \tag{4.1}$$

The quantities $\widehat{b}(t)$, $\widehat{v}_{\rm np}(t)$, $\widehat{v}_{\rm pm}(t)$, and $\widehat{\kappa}(t)$ are estimators of $b(t)$, $v_{\rm np}(t)$, $v_{\rm pm}(t)$, and $\kappa(t)$, the t-indexed modifications of the corresponding (unindexed) quantities introduced in Section 2.

In the above reasoning, we have essentially worked with known, non-stochastic weight functions. When the weight function W depends on one or more unknown quantities, the natural solution is to simply insert the empirical analogues of these. Replacing W by some estimate $\widehat{W}$ in (4.1) is perfectly valid in the sense that one is still estimating the same risk. A special case of this is touched in Section 5.2. If W itself is stochastic, the risk function changes and new derivations, which ought to result in different AFIC formulae, are required. A practical illustration of the AFIC in action is given in the supplementary material (Jullum and Hjort (2017a)).

5. Properties, performance, and relation to goodness-of-fit

In this section we investigate the behaviour of the developed FIC and AFIC schemes in a few model frameworks. This sheds light on certain implied goodness-of-model tests associated with the FIC. We also briefly discuss simulation results

from the supplementary material (Jullum and Hjort (2017a)) comparing the performance of the FIC and AFIC to competing selection criteria.

5.1. Behaviour of FIC

When to use a given parametric model rather than the nonparametric option involves the quantity $Z_n = n\widehat{b}^2$, where $\widehat{b} = \widehat{\mu}_{\rm pm} - \widehat{\mu}_{\rm np}$, cf. (2.6), along with $\widehat{\eta} = \widehat{v}_{\rm np} - \widehat{v}_{\rm c}$. Recalling that $\widehat{\kappa} = \widehat{v}_{\rm pm} + \widehat{v}_{\rm np} - 2\widehat{v}_{\rm c}$, and re-arranging terms in the inequality $\mathrm{FIC}_{\rm pm} \leq \mathrm{FIC}_{\rm np}$, it is seen that this is equivalent to $\max(\widehat{\kappa}, Z_n) \leq 2\widehat{\eta}$. For the untruncated version which uses $\widehat{\mathrm{bsq}}_0$ instead of $\widehat{\mathrm{bsq}}$ to estimate the squared bias, we have $\mathrm{FIC}^*_{\rm pm} \leq \mathrm{FIC}_{\rm np}$ if and only if $Z_n \leq 2\widehat{\eta}$. Assuming that the estimated variance of the nonparametric estimator is no smaller than that of the simpler parametric alternative, which typically is true, we have $\widehat{\kappa} \leq 2\widehat{\eta}$. Hence, for both the truncated and non-truncated versions, a parametric model is chosen over the nonparametric when $Z_n \leq 2\widehat{\eta}$. The ranking among different parametric models may change depending on which version of the scheme is used; it is only the individual comparison with the nonparametric candidate that is identical for the two versions.

Assume the nonparametric candidate competes against k different parametric models pm_j with limiting bias b_j, $j = 1, \ldots, k$, defined as before but now for the different competing models. Let the quantities η_j and κ_j be the natural nonzero limiting quantities of $\widehat{\eta}_j$ and $\widehat{\kappa}_j$ for these candidate models. To gain further insight one investigates the selection probability for a specific parametric model, say pm_j,

$$\alpha_n(G, j) = \Pr_G(\text{FIC selects } \mathrm{pm}_j) \tag{5.1}$$

under different conditions.

Lemma 1. *Under the conclusions of Propositions* 1 *and* 2*, if* $b_j \neq 0$*, then* $\alpha_n(G, j) \to 0$*, and if* $b_j = 0$*,* $v_{\mathrm{pm},j} \leq v_{\rm np}$*, and* $b_l \neq 0$ *for* $l \neq j$*, then* $\alpha_n(G, j) \to \Pr(\chi^2_1 \leq 2\eta_j/\kappa_j)$*, both for the truncated and the untruncated version of the FIC.*

Corollary 1. *Assume that the conclusions of Propositions* 1 *and* 2 *hold, that the* j*-th model is in fact fully correct, and that* $b_l \neq 0$ *for* $l \neq j$*, and let* $\Theta^{(j)}$ *be some neighbourhood of the least false parameter value of the* j*-th model. If (C1–C4) hold for this* j*-th model and* $\sup_{\theta \in \Theta^{(j)}} \|u(Y_i; \theta)\|$*,* $\sup_{\theta \in \Theta^{(j)}} \|I(Y_i; G) + u(Y_i; \theta)u(Y_i; \theta)^{\rm t}\|$ *and* $|\mathrm{IF}(Y_i; G)| \sup_{\theta \in \Theta^{(j)}} \|u(Y_i; \theta)\|$ *have finite means, then* $c_j = d_j, \eta_j = \kappa_j$*, and* $\alpha_n(G, j) \to \Pr(\chi^2_1 \leq 2) \doteq 0.843$*.*

The limiting behaviour of the FIC scheme is not surprising. As seen from (2.8), when the parametric models are biased (i.e. having $b \neq 0$), the nonpara-

metric, by its nature being correct in the limit, is eventually chosen by the FIC as the sample size grows. However, when a parametric model produces asymptotically unbiased estimates (i.e. when $b = 0$), the FIC selects this parametric model rather than the nonparametric with a positive probability. The precise probability depends on the model structure and focus through η_j/κ_j, except when the unbiasedness is caused by the parametric model being exact – then the probability is 0.843. It is no paradox that the probability of choosing an unbiased parametric model is smaller than 1. In that situation, the nonparametric model is correct in the limit and is thereby selected with a positive probability.

When several parametric models have the above unbiasedness property, the limiting selection probabilities are generally smaller for all candidate models. The precise probabilities depend on the focus parameter, whether the truncated or untruncated version of the FIC is used, and how these models are nested or otherwise related to each other. Consider for instance the case in which an exponential model is correct, the focus parameter is the median, and the Weibull and exponential models are considered as parametric options. The asymptotic probabilitites for FIC selecting, respectively, the nonparametric, Weibull, and exponential models are then 0.085, 0.125 and 0.789 for the truncated version and 0.085, 0.183 and 0.731 for the untruncated version. Thus, the probability of selecting the nonparametric is the same for the two versions, while the probabilities for the parametric candidates are different. These probabilities were obtained by simulating in the limit experiment as discussed in Remark 1 of Jullum and Hjort (2017a).

Remark 1. *The implied FIC test has level* 0.157. It is worth noting clearly that the FIC is a selection criterion, constructed to compare and rank candidate models based on estimated precision, and not a test procedure per se. One can nevertheless choose to view FIC with two candidate models, the nonparametric and *one* parametric model, as an 'implicit testing procedure'. FIC is then a procedure for checking the null hypothesis that the parametric model is adequate for the purpose of estimating the focus parameter with sufficient precision. When testing the hypothesis, $\beta_n(G) = 1 - \alpha_n(G)$ is the power function of the test, tending to 1 for each G with non-zero bias b. For $G = F_{\theta_0}$, the probability $\beta_n(F_{\theta_0})$ is the probability of rejecting the parametric model when it is correct, i.e. the significance level of the test. This implied level is close to $1-0.843 = 0.157$ for large n. Our view is that the FIC approach has a conceptual advantage over the goodness-of-fit testing approach, as the risk assessment starting point delivers an implicit test with a certain significance level, found to be 0.157, as opposed

to fixing an artificially pre-set significance level, like 0.05.

5.2. Behaviour of AFIC

Based on arguments similar to those used for the FIC, the behaviour of the AFIC of Section 4 is related to the goodness-of-fit type statistic $Z_n^* = n\int \widehat{b}^2(t)\,\mathrm{d}W(t)$. In particular, as long as $\int \widehat{v}_{\rm pm}(t)\,\mathrm{d}W(t) \le \int \widehat{v}_{\rm np}(t)\,\mathrm{d}W(t)$ (which is typically the case), the parametric model is preferred over the nonparametric model when

$$Z_n^* = n\int \{\widehat{\mu}_{\rm pm}(t) - \widehat{\mu}_{\rm np}(t)\}^2\,\mathrm{d}W(t) \le 2\widehat{\eta}^*, \tag{5.2}$$

for $\widehat{\eta}^* = \int\{\widehat{v}_{\rm np}(t) - \widehat{v}_{\rm c}(t)\}\,\mathrm{d}W(t)$. From the first part of Lemma 1, when $b(t) \neq 0$ for some set of t values with positive measure (with respect to W), then $\Pr_G(Z_n^* \le 2\widehat{\eta}^*) \to 0$ as $n \to \infty$, i.e. the nonparametric is chosen with probability tending to 1. This result holds independently of whether truncation of the squared bias is done after integration as in (4.1), before integration, or not at all.

We investigate the limiting behaviour of AFIC when a parametric model is fully correct. Even if the decisive terms appear similar to those for the FIC, such an investigation is more complicated for AFIC and depends both on the type of focus parameters and the weight function W. We shall therefore limit ourselves to the rather unfocused case where the focus parameter set is the complete c.d.f. and we consider weight functions W_1 and W_2 specified through $\mathrm{d}W_1(y) = \mathrm{d}F(y;\theta_0)$ and $\mathrm{d}W_2(y) = 1\,\mathrm{d}y$. The former is estimated by inserting $\widehat{\theta}$ for the unknown θ_0. In these cases Z_n^* equals, respectively, $C_{1,n} = \int B_n(y)^2\,\mathrm{d}F(y;\widehat{\theta})$ and $C_{2,n} = \int B_n(y)^2\,\mathrm{d}y$, for $B_n(y) = \sqrt{n}\{\widehat{G}_n(y) - F(y;\widehat{\theta})\}$. These have corresponding η^* estimates given by $\widehat{\eta}_1^* = \int\{\widehat{v}_{\rm np}(y) - \widehat{v}_{\rm c}(y)\}\,\mathrm{d}F(y;\widehat{\theta})$ and $\widehat{\eta}_2^* = \int\{\widehat{v}_{\rm np}(y) - \widehat{v}_{\rm c}(y)\}\,\mathrm{d}y$. Here $C_{1,n}$ corresponds to the classic Cramér–von Mises goodness-of-fit test statistic with estimated parameters, see e.g. Durbin, Knott and Taylor (1975).

Durbin (1973) studied the limiting behaviour of the process $B_n^0(u) = \sqrt{n}\{\widehat{G}_n(F^{-1}(u;\widehat{\theta})) - F^{-1}(u;\widehat{\theta})\}$ for $u \in [0,1]$. Results achieved there may be extended to deal with the B_n process and consequently also the convergence of $C_{1,n}$ and $C_{2,n}$.

Lemma 2. *Take $G = F_{\theta_0}$, with $\int G(y)\{1-G(y)\}\,\mathrm{d}y$ finite. Suppose that for some neighbourhood Θ^* around θ_0, $F(y;\theta)$ has a density $f(y;\theta)$ with $c(y;\theta) = \partial F(y;\theta)/\partial\theta = \int_{-\infty}^y f(x;\theta)u(x;\theta)\,\mathrm{d}x$ continuous in θ. If $\mathrm{E}_G\{\sup_{\theta\in\Theta^*}\|u(Y_i;\theta)\|\}$ is finite, then $B_n = \sqrt{n}\{\widehat{G}_n - F(\cdot\,;\widehat{\theta})\}$ converges in distribution to B, a Gaussian zero-mean process with covariance function*

$$\mathrm{Cov}\{B(y_1),B(y_2)\} = F(\min(y_1,y_2);\theta_0) - F(y_1;\theta_0)F(y_2;\theta_0) - c(y_1;\theta_0)^{\rm t}J^{-1}c(y_2;\theta_0).$$

Also,

$$C_{1,n} \xrightarrow{d} C_1 = \int B(y)^2 \,\mathrm{d}G(y) = \int_0^1 B(G^{-1}(r))^2 \,\mathrm{d}r,$$
$$C_{2,n} \xrightarrow{d} C_2 = \int B(y)^2 \,\mathrm{d}y = \int_0^1 \frac{B(G^{-1}(r))^2}{g(G^{-1}(r))} \,\mathrm{d}r.$$

This result takes care of the left side of (5.2) for W_1 and W_2. The corresponding right sides concern $\widehat{\eta}_1^*$ and $\widehat{\eta}_2^*$, as estimates of, respectively, $\eta_1^* = \int\{v_{\rm np}(y) - v_{\rm c}(y)\}\,\mathrm{d}F(y;\theta_0) = \int\{v_{\rm np}(y) - v_{\rm pm}(y)\}\,\mathrm{d}F(y;\theta_0)$ and $\eta_2^* = \int\{v_{\rm np}(y) - v_{\rm c}(y)\}\,\mathrm{d}y = \int\{v_{\rm np}(y) - v_{\rm pm}(y)\}\,\mathrm{d}y$. These estimates need to be consistent with η_1^* and η_2^* finite. However, since $v_{\rm pm}(y) \le v_{\rm np}(y)$ when $G = F_{\theta_0}$ and the additional conditions of Corollary 1 hold, then $\eta_1^* \le 2\int G(y)\{1-G(y)\}\,\mathrm{d}G(y) \le 1/2$ and $\eta_2^* \le 2\int G(y)\{1-G(y)\}\,\mathrm{d}y$. The latter integral is assumed to be finite in the above lemma. Thus, under these conditions, and with $\widehat{v}_{\rm np}(y)$ and $\widehat{v}_{\rm pm}(y)$ consistent for each single y (as per the conditions of Proposition 2), we then have $\widehat{\eta}_1^* \to_{\rm pr} \eta_1^*$ and $\widehat{\eta}_2^* \to_{\rm pr} \eta_2^*$, with both limits finite.

Under specific model assumptions and setups one can compute the limiting probability of selecting a correct parametric model over the nonparametric when using AFIC with the above weight functions. The distribution of C_1 and C_2 may be approximated by sampling the Gaussian B processes on a dense grid of $u \in [0,1]$ followed by Riemann integration, while η_1^* and η_2^* can be computed by numerical integration.

By sampling 10^7 B processes on a grid of size 2,000, we approximated the limiting probability for the event that AFIC selects $\mathrm{N}(\xi,\sigma^2)$ (when both ξ and σ are unknown) over the nonparametric alternative. We obtained probabilities of 0.938 and 0.951 for, respectively, W_1 and W_2, corresponding to implied test levels of respectively 0.062 and 0.049. Such tests are parameter independent, but different families of distributions give different probabilities. For example, repeating the simulations, holding first ξ and then σ fixed while the other is being estimated, gives limiting test levels equal to 0.071 and 0.116 for W_1, and 0.062 and 0.106 for W_2.

Remark 2. *A new interpretation of the Pearson chi-squared test.* Consider counts $N = (N_1,\ldots,N_k)$ from a multinomial model with probability vector $(p_1,\ldots,p_k)$, where $\sum_{j=1}^k p_j = 1$ and $\sum_{j=1}^k N_j = n$. Pearson (1900) introduced the test statistic $\sum_{j=1}^k (N_j - np_j)^2/(np_j) = n\sum_{j=1}^k (\bar{p}_j - p_j)^2/p_j$ and showed its convergence to the χ^2_{k-1}, where $\bar{p}_j = N_j/n$. If the null distribution has a parametric form, say $p_j = f_j(\theta)$ in terms of a parameter θ of dimension say $q \le k-2$, then the modified Pearson test statistic is $X_n = n\sum_{j=1}^k \{\bar{p}_j - f_j(\widehat{\theta})\}^2/f_j(\widehat{\theta})$, tend-

ing under that model to a $\chi^2_{\rm df}$, with $\rm df = k-1-q$. This holds for both maximum likelihood and minimum-chi-squared estimators.

Since categorical i.i.d. data are just a special case of the general i.i.d. theory, all of the developed FIC and AFIC theory holds here. This is dealt with in the supplementary material (Jullum and Hjort (2017a)), with particular attention to the AFIC case where interest is in all probabilities $p_1, \ldots, p_k$, with loss function weights $1/p_1, \ldots, 1/p_k$, so that the loss function is $\sum_{j=1}^k (\widehat{p}_j - p_j)^2/p_j$. In the AFIC scheme, one learns that a parametric model and its estimates $f_j(\widehat{\theta})$ are preferred to the nonparametric with its $\bar{p}_j$ when $X_n \leq 2\{k-1-{\rm Tr}(\widehat{J}^{-1}\widehat{K}^*)\}$, with $\widehat{K}^*$ a $q \times q$-dimensional matrix defined in the supplementary material (Jullum and Hjort (2017a)). This makes AFIC directly related to the Pearson chi-squared, but derived via assessment of risk. If the parametric model is correct, AFIC selects that model with probability tending to $\Pr(\chi^2_{\rm df} \leq 2\,{\rm df})$. This generalises the implied test of Remark 1 when we have categorical data, and sheds new light on the Pearson chi-squared test, both regarding interpretation and the implied significance levels. The test level decreases with increasing df and is, for instance, $0.157, 0.135, 0.112, 0.092, 0.075$ for $\rm df = 1, \ldots, 5$.

For assessing independence in an $r \times s$ table, one finds the following AFIC recipe: Accept independence when $X_n \leq 2(r-1)(s-1)$, where $X_n = \sum_{i,j}(N_{i,j} - n\widehat{\alpha}_i\widehat{\beta}_j)^2/N_{i,j}$ is the chi-squared test, with $\widehat{\alpha}_i = N_{i,\cdot}/n$ and $\widehat{\beta}_j = N_{\cdot,j}/n$. Here $N_{i,\cdot} = \sum_j N_{i,j}$ and similarly with $N_{\cdot,j}$.

5.3. The original vs. the new FIC

Local misspecification frameworks are often used to study test power, see e.g. Lehmann (1998, Chap. 3.3). Their frequent use in such studies is due to the fact that they bring variance and squared biases on the same asymptotic scale. Although we left the local misspecification framework when deriving the FIC, such a framework may be useful for studying limiting properties and especially comparing the FIC methodology in the present paper with the original FIC scheme of Claeskens and Hjort (2003). Taking the true density or probability mass function to be $g_n(y) = f(y; \theta_0, \gamma_0 + \delta/\sqrt{n})$, the comparison is restricted to nested parametric models ranging from an unbiased 'wide' model with large variance to a locally biased 'narrow' model with minimal variance, where the wide model plays the role of the nonparametric model. Under suitable regularity conditions this framework deems the two FIC regimes asymptotically equivalent when $v_{\rm np} = v_{\rm wide}$ (with the nonparametric and wide models being equivalent in the new FIC scheme). The more typical case of $v_{\rm wide} < v_{\rm np}$ reflects in some

local asymptotics sense that the new FIC includes an additional uncertainty level outside the wide model, and replaces wide model variances with nonparametric model variances. In this regard, the new FIC scheme may be thought of as the most model robust. Further details with precise formulae, regularity conditions, and proofs are given in the supplementary material (Jullum and Hjort (2017a)).

5.4. Summary of simulation experiments

The supplementary material (Jullum and Hjort (2017a)) describes some simulation studies investigating the performance of various versions of our FIC and AFIC schemes for the case in which none of the parametric models are fully correct.

We checked the performance of estimators which used the models ranked the best by various FIC and AFIC schemes. When concentrating on $\mu = G(y)$ for a wide range of y-values, the truncated and untruncated squared bias versions of the FIC performed similarly. Their performance was clearly better than the BIC and comparably or slightly better than the AIC. Versions of the AFIC covering the whole distribution performed comparably with the AIC and BIC.

For various other focus parameters, the full version of the FIC (with the nonparametric candidate included) performed better than the AIC, the BIC, a constructed Kolmogorov–Smirnov criterion, and the nonparametric estimator itself, for moderate to large samples. For small samples, the AIC and BIC typically performed somewhat better.

6. Model Averaging

An alternative to relying on a specific model for estimation of the focus parameter is to use an average over several candidate models. Model averaging uses as a final estimator a weighted average across all candidate models, say $\widehat{\mu}_{\text{final}} = \sum_j a_j \widehat{\mu}_j$, where the weights a_j sum to 1 and typically depend on aspects of the data. Mixing parametrics and nonparametrics in such a setting is conceptually appealing as it mixes nonzero bias and low variance with zero bias and higher variance. Motivated by the form of model averaging schemes for AIC, BIC, and similar (see e.g. Claeskens and Hjort (2008, Chap. 7)), we suggest using

$$a_j = \frac{\exp(-\lambda \,\text{FIC}_j/\text{FIC}_{\text{np}})}{\sum_k \exp(-\lambda \,\text{FIC}_k/\text{FIC}_{\text{np}})}, \tag{6.1}$$

for some specified tuning-parameter λ. This weight function has the property of producing the same weights independently of the scale of the focus parameter.

The size of λ relates to the emphasis on relative differences in mean squared errors and may be set based on cross-validation or similar ideas. As $\lambda \to 0$ in (6.1), all estimators are weighted equally. When $\lambda \to \infty$ all weight is concentrated on the model and estimator with the smallest FIC score, bringing the scheme back to regular model selection based on FIC. A corresponding model averaging scheme can be created based on the AFIC in Section 4.

A local misspecification framework is useful when working with model averaging. When the parametric models are nested, one can derive local limit distributions for model average estimators with weight functions like (6.1), analogous to Hjort and Claeskens (2003, Thm. 4.1). Such limiting distributions may be used to address post-selection uncertainties of the model average estimators similarly to Hjort and Claeskens (2003). Details and proof for the limiting distribution of the model averaging scheme are given in the supplementary material (Jullum and Hjort (2017a)). See also Hjort and Claeskens (2003) and Claeskens and Hjort (2008, Chap. 7) for further remarks on model averaging based on AIC, BIC, and the original FIC.

7. Other Data Frameworks

Proposition 1 and the joint limiting distribution structure of (2.3) form the basis for deriving the FIC and AFIC methods presented in this paper. Versions of (2.3) hold also for cases outside the i.i.d. regime we have been working within. In particular, below we touch on FIC and AFIC methods for hazard rate and time series models. In other situations, structures more complicated than (2.3) arises, for example when comparing nonparametric and parametric density estimation and regression, leading in their turn to certain necessary refinements.

7.1. Hazard rate models

The Kaplan–Meier and Nelson–Aalen nonparametric estimators are extensively used in practical applications involving censored data for, respectively, survival curves and cumulative hazard rates, even in cases when a parametric approach would have been better; see e.g. Miller (1981), who asks "what price Kaplan–Meier?". The cumulative hazard $A(t)$ at a point t may be estimated by the nonparametric Nelson–Aalen estimator $\widehat{A}_{\rm naa}(t)$ or a parametric candidate $\widehat{A}_{\rm pm}(t) = A(t;\widehat{\theta})$. Here $A(t;\theta)$ is θt for the constant hazard rated exponential model, $(\theta_1 t)^{\theta_2}$ for the Weibull, and $(\theta_1/\theta_2)\{\exp(\theta_2 t) - 1\}$ for the Gompertz, etc. Under suitable regularity conditions, mainly those in Andersen et al. (1993, Thm. IV.1.2) and Hjort (1992, Thm. 2.1), in addition to $A(t;\theta)$

being smooth w.r.t. θ at a certain least false θ_0, we have joint convergence of $\sqrt{n}\{\widehat{A}_{\rm naa}(t)-A(t)\}$ and $\sqrt{n}\{\widehat{A}_{\rm pm}(t)-A(t;\theta_0)\}$ to a zero-mean Gaussian distribution with a certain covariance matrix. A delta method argument for $\exp\{-A(t)\}$ reveals the analogue for a survival probability which also holds when using the nonparametric Kaplan–Meier estimator rather than the asymptotically equivalent $\exp\{-\widehat{A}_{\rm naa}(t)\}$. Hence, with consistent estimation of that covariance matrix, FIC and AFIC schemes may be put up both for cumulative hazard and for survival probability estimation. The survival probability schemes may then be used to give a possibly more fine-tuned and focused answer to Miller's rhetorical question than those given by Miller (1981) and Meier et al. (2004).

7.2. Proportional hazard regression

Suppose covariate vectors x_i are recorded along with (possibly censored) survival times Y_i. One may then ask 'what price semiparametric Cox regression?'. This question is more complicated. With $\alpha_i(s)$ the hazard rate for individual i, the traditional proportionality assumption is that $\alpha_i(s)=\alpha_0(s)\exp(x_i^{\rm t}\beta)$, with $\alpha_0(s)$ and β unknown. The most appropriate modelling schemes are (i) the semiparametric Cox method, which estimates β by maximising the partial likelihood, say $\widehat{\beta}_{\rm cox}$, and the baseline hazard $A_0(t)=\int_0^{\rm t}\alpha_0(s)\,{\rm d}s$ by the Breslow estimator $\widehat{A}_{\rm br}(t)$ (Breslow (1972)); and (ii) fully parametric candidates, which with a suitable $\alpha_0(s;\theta)\exp(x_i^{\rm t}\beta)$ use the full maximum likelihood estimators $(\widehat{\theta},\widehat{\beta})$ in consequent inference. These approaches give rise to the semiparametric $\widehat{A}_{\rm br}(t)\exp(x^{\rm t}\widehat{\beta}_{\rm cox})$ and the parametric $A(t;\widehat{\theta})\exp(x^{\rm t}\widehat{\beta})$, for estimating the cumulative hazard rate $A(t\,|\,x)$ for a given individual, and similarly also for the survival probabilities $S(t\,|\,x)=\exp\{-A(t\,|\,x)\}$. Under regularity conditions, including those for the standard hazard rate models above, one can establish joint convergence for

$$\big(\sqrt{n}\{\widehat{A}_{\rm br}(t)\exp(x^{\rm t}\widehat{\beta}_{\rm cox})-A(t\,|\,x)\},\sqrt{n}\{A(t;\widehat{\theta})\exp(x^{\rm t}\widehat{\beta})-A(t;\theta_0)\exp(x^{\rm t}\beta_0)\}\big),$$

involving certain least false parameters (θ_0,β_0). With appropriate efforts this leads to FIC and AFIC formulae for choosing between semiparametric and parametric hazard models, in terms of precision of estimators for either cumulative hazards or survival curves. This is indeed the main theme in Jullum and Hjort (2017b).

7.3. Stationary time series

In the time series culture, there seems to be more or less two separate

schools, when it comes to modelling, estimation and prediction: the parametric and the nonparametric. For the following very brief explanation of how the FIC might be put to work also here, consider a zero-mean stationary Gaussian time series process with spectral distribution function G on $[-\pi,\pi]$. A class of focus parameters take the form $\mu(G) = A(\int h(\omega)\,\mathrm{d}\omega)$, where A is smooth and $h = (h_1,\ldots,h_k)^{\mathrm{t}}$ is a vector of univariate bounded functions h_j on $[-\pi,\pi]$, each having at most a finite number of discontinuities. This class includes all covariances, correlations, natural predictors, and threshold probabilities, and have natural parametric and nonparametric estimators $\widehat{\mu}_{\mathrm{pm}} = A(\int_{-\pi}^{\pi} h(\omega) f(\omega;\widehat{\theta})\,\mathrm{d}\omega)$ and $\widehat{\mu}_{\mathrm{np}} = A(\int_{-\pi}^{\pi} h(\omega) I_n(\omega)\,\mathrm{d}\omega)$. Here $I_n(\omega)$ is the classical periodogram, $f(\cdot\,;\theta)$ is a spectral density function parametrised by θ, and $\widehat{\theta}$ is the maximiser of the Gaussian log-likelihood (or of its Whittle approximation). Then, under mild regularity conditions, we have joint convergence in distribution for $\sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu)$ and $\sqrt{n}(\widehat{\mu}_{\mathrm{pm}} - \mu_0)$ to certain zero-mean Gaussian distributions. Once again μ_0 is a parametric least false variant of μ. This may be used to establish FIC and AFIC formulae also for this framework, leading to selection and averaging methods when comparing e.g. autoregressive models of different order along with the nonparametric.

7.4. Parametric or nonparametric density estimation

Although a joint limiting distribution like that of (2.3) holds in a wide range of situations, frameworks involving nonparametric smoothing are typically different. One such is density estimation for i.i.d. data. In this situation, nonparametric estimators typically converge no faster than $n^{-2/5}$ (see e.g. Brown and Farrell (1990)), being slower than the usual $n^{-1/2}$ that still works for the parametric candidates; hence, there is no direct analogue of (2.3) of Proposition 1. This complicates constructions of FIC formulae, both for nonparametrics and parametrics, but such can nevertheless be reached. We provide a brief investigation into these matters when interest is in estimation of $\mu = g(y)$ for a particular y.

The traditional nonparametric density estimator with i.i.d. observations Y_1, $\ldots$, Y_n is the kernel-based

$$\widehat{g}_n(y) = n^{-1}\sum_{i=1}^{n} h_n^{-1} M(h_n^{-1}(y - Y_i)), \tag{7.1}$$

with bandwidth h_n and kernel function M. Let $k_2 = \int x^2 M(x)\,\mathrm{d}x, R_M = \int M(x)^2\,\mathrm{d}x$, with $g''(y)$ the second order derivative of $g(y)$. With bandwidth of the optimal size $h_n = an^{-1/5}$, for some constant a, one finds that

$$\begin{pmatrix} n^{2/5}\{\widehat{g}_n(y) - g(y)\} \\ \sqrt{n}\{f(y;\widehat{\theta}) - f(y;\theta_0)\} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

with $X_1 \sim \mathrm{N}(\frac{1}{2}k_2 a^2 g''(y), R_M g(y)/a)$ and $X_2 \sim \mathrm{N}(0, v_{\mathrm{pm}}(y))$, where the latter has variance $v_{\mathrm{pm}}(y) = f(y;\theta_0)^2 u(y;\theta_0)^{\mathrm{t}} J^{-1} K J^{-1} u(y;\theta_0)$. The covariance between the two variables on the left hand side is of size $O(n^{-1/10})$, but in the large-sample limit this disappears, rendering X_1 and X_2 independent. We concentrate on leading terms only, discarding terms of lower order, like the negative $g^2(y)/n$ term often included in the nonparametric variance. The more complicated parallels to (1.4) are then

$$\begin{aligned} \mathrm{mse}_{\mathrm{np}} &= \frac{1}{4}k_2^2 h_n^4 g''(y)^2 + \frac{R_M g(y)}{nh_n}, \\ \mathrm{mse}_{\mathrm{pm}} &= b(y)^2 + n^{-1} v_{\mathrm{pm}}(y), \end{aligned}$$

with $b(y) = f(y;\theta_0) - g(y)$. These quantities can be estimated from data, though with certain complications and perhaps separate fine-tuning for $g''(y)^2$ and the variance of $\widehat{b}(y) = f(y;\widehat{\theta}) - \widehat{g}_n(y)$, where the latter is used for correcting $\widehat{b}(y)^2$ for overshooting bias when estimating $b(y)^2$.

Matters simplify somewhat when we employ suitable bias-reduction methods for estimating $g(y)$ in the first place, e.g. along the lines of Jones, Linton and Nielsen (1995); Hjort and Jones (1996). Some of these methods lead to density estimators, say $\widetilde{g}_n(y)$, with the ability to reduce the bias significantly without making the variance much larger. Specifically, $n^{2/5}\{\widetilde{g}_n(y) - g(y)\} \xrightarrow{d} \mathrm{N}(0, S_M g(y)/a)$ holds, with S_M a constant depending only on M and equal to or a bit bigger than R_M above. A variation of these methods, due to Hengartner and Matzner-Løber (2009), involving a pilot estimator with a separate bandwidth, indeed satisfies the above with $S_M = R_M$. With any of these methods,

$$\mathrm{mse}_{\mathrm{np}} = \frac{S_M g(y)}{nh_n} \quad \text{and} \quad \mathrm{mse}_{\mathrm{pm}} = b(y)^2 + n^{-1} v_{\mathrm{pm}}(y).$$

With $\widetilde{b}(y) = f(y;\widehat{\theta}) - \widetilde{g}_n(y)$, we find $n^{2/5}\{\widetilde{b}(y) - b(y)\} \xrightarrow{d} \mathrm{N}(0, S_M g(y)/a)$, basically since $\widetilde{g}_n(y)$ is more variable than $f(y;\widehat{\theta})$. With $h = h_n$, this leads to the FIC formulae

$$\begin{aligned} \mathrm{FIC}_{\mathrm{np}} &= \frac{S_M \widetilde{g}_n(y)}{nh}, \\ \mathrm{FIC}_{\mathrm{pm}} &= \max\{0, \widetilde{b}(y)^2 - \frac{S_M \widetilde{g}_n(y)}{nh}\} + n^{-1}\widehat{v}_{\mathrm{pm}}. \end{aligned}$$

As opposed to the cases $g(y)$ and $g''(y)$, the integrated squared density $\int g(y)^2 \,\mathrm{d}y$ and the density roughness $\int g''(y)^2 \,\mathrm{d}y$ turn out to be estimable at

a $\sqrt{n}$ rate, under suitable smoothness conditions, see Fan and Marron (1992). Hence, these cases fall under our framework associated with (2.3). The latter quantity is involved in the penalisation terms for smoothing methods, etc., and the first appears e.g. when assessing the accuracy of the Hodges–Lehmann estimator for location. In both of these cases nonparametric methods are associated with significant estimation variability. It is hence tempting to use the FIC to decide whether we should bypass these and go for a parametric option instead. In particular, such a FIC step could determine whether one should follow a complicated nonparametric recipe for choosing the bandwidth in (7.1) or a simple normal-based rule-of-thumb bandwidth. The FIC methods above may also be extended to suitable AFIC versions, determining from data whether one should use a kernel estimator for g or one from a list of parametric candidate models, over some interval $[a, b]$.

7.5. Parametric or nonparametric regression

For the classical regression framework, the situation, as per construction of FIC formulae for comparing nonparametrics and parametrics, is partly similar to the case for density estimation. Again, the nonparametric method typically converges at a slower speed than parametric options, but to the correct quantity. For this situation, assume pairs (x_i, Y_i) are observed from the model $Y_i = m(x_i) + \varepsilon_i$, with the ε_i being i.i.d. $\mathrm{N}(0, \sigma^2)$. For simplicity of presentation we take the x_i to all be on the unit interval, stemming from a design density $g_X(x)$ there. Of interest is to select among different parametric models for the regression line $\mathrm{E}(Y \,|\, x)$, say $m(x; \beta)$, and the nonparametric model which simply takes $m(x)$ to be unknown and smooth. These questions can be put up for a single position x, or, with AFIC, for an interval of x values.

Making this operational demands choosing one of many available nonparametric smoothers. Here we take the local linear $\widehat{m}(x) = \widehat{a}_x + \widehat{b}_x x$, with $(\widehat{a}_x, \widehat{b}_x)$ chosen to minimise $Q_x(a, b) = \sum_{i=1}^n M_h(x_i - x)(y_i - a - bx_i)^2$, for $M_h(u) = h^{-1} M(h^{-1}u)$, with a kernel M (say the standard normal) and bandwidth h. Then, with $h = h_n$ tending to zero and nh_n growing to infinity,

$$\mathrm{E}\,\{\widehat{m}(x)\} \doteq m(x) + \frac{1}{2} k_2 h^2 m''(x) \quad \text{and} \quad \mathrm{Var}\,\{\widehat{m}(x)\} \doteq \frac{\sigma^2 R_M}{nh g_X(x)},$$

with k_2 and R_M constants depending on the kernel, and $m''(x)$ the second derivative of $m(x)$. Thus

$$\mathrm{mse}_{\mathrm{np}}(x) \doteq \frac{1}{4} k_2^2 h^4 m''(x)^2 + \frac{\sigma^2 R_M}{nh g_X(x)}.$$

This may be estimated from data, via (i) a kernel density estimator for $g_X(x)$; (ii) a separate smoother with a separate bandwidth for $m''(x)$; and (iii) an estimator for σ^2 which does not involve any parametric models, but uses $n^{-1}\sum_{i=1}^n\{y_i - \widehat{m}(x_i)\}^2$ or similar. The corresponding mse for a parametric $m(x;\beta)$ takes the form

$$\mathrm{mse}_{\mathrm{pm}}(x) \doteq b(x)^2 + \dot{m}(x;\beta_{0,n})^{\mathrm{t}}\Sigma_n\dot{m}(x;\beta_{0,n}).$$

Here $b(x) = m(x;\beta_{0,n}) - m(x)$, involving the least false $\beta_{0,n}$, minimising $\sum_{i=1}^n\{m(x_i;\beta) - m(x_i)\}^2$; Σ_n is the variance matrix of the maximum likelihood estimator $\widehat{\beta}$, defined as the minimiser of $Q_n(\beta) = \sum_{i=1}^n\{y_i - m(x_i;\beta)\}^2$; and $\dot{m}(x;\beta) = \partial m(x;\beta)/\partial\beta$.

Spelling out details for estimating $\mathrm{mse}_{\mathrm{pm}}(x)$ from data takes a bit of care, with ingredients rather similar to those for density estimation. It can be accomplished, but with additional fine-tuning questions to tend to.

8. Concluding Remarks

We have concentrated on studying the simplest data frameworks and classes of models, and have chosen to give detailed analysis regarding those, rather than broadening the concepts and ideas further. Some extensions follow along the same lines as those carried out here, only with more involved theory, while others require adjustments in our proposed construction strategy. An example of the former is the extension to relying on M-estimators (van der Vaart (2000, Chap. 5)) for estimating the parameters in the parametric models, typically for robustness reasons, rather than using maximum likelihood. Under regularity conditions essentially analogous to those for maximum likelihood theory, (2.3) still holds with appropriately modified quantities. In particular, such an extended FIC apparatus makes it possible to rank different estimates of the same focus parameter, derived from the same parametric model (say the ML and some competing M-estimators).

Our theory typically also works for other parametric estimation routines, provided they have the same convergence rate, such as the estimator gotten by any finite number of steps of the EM algorithm. It is also possible to use the same procedure for other types of data, as described in Section 7. Similarly, other nonparametric estimation methods may be worked with. We also point out that other nonparametric estimation methods might be first-order large-sample equivalent to using $\widehat{\mu}_{\mathrm{np}} = T(\widehat{G}_n)$, for certain classes of problems, such as when using the empirical likelihood (Hjort, McKeague and Van Keilegom (2009)).

A nice property of our model selection criteria is that they are all invariant under parametrisation: two parametric classes having identical model forms and flexibility, but with differing parametric distributional expressions, produce identical estimates of μ *and* can be treated identically by our criteria.

Our focused approach is built on the concept that changing focus may change the ranking of candidate models. However, when a focus parameter may be expressed as a linear function of another focus parameter, all the new FIC scores are proportional to the corresponding FIC scores before transformation. This results in identical ranking for the FIC and AFIC. Another consequence is that the rankings often are the same when a transformation of the focus parameter is almost linear.

When constructing our FIC we have used the squared error loss function, where the risk is expressed as squared bias plus variance. Other loss functions may be worked with, though these would often lead to risks being harder to estimate unbiasedly. In particular, a FIC can be put up for the so-called linex loss function, where loss is measured as $\{\exp(a\Delta) - 1 - a\Delta\}/a$ with $\Delta = \widehat{\mu} - \mu$ and a some tuning parameter.

`R`-scripts and functions presenting FIC and AFIC tables and plots upon specification of data, parametric competitors and one or more focus parameters, are prepared for the cases handled in the paper. The programs are currently available on request from the authors, but are planned to be included in an `R`-package.

Supplementary Materials

The supplementary material (Jullum and Hjort (2017a)) contains proofs of results in the main paper, another illustration, details of simulation studies, some details on FIC and AFIC for categorical data, and some local asymptotics results.

References

Andersen, P. K., Borgan, Ø., Gill, R. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, Heidelberg.

Behl, P., Dette, H., Frondel, M. and Tauchmann, H. (2012). Choice is suffering: A focused information criterion for model selection. *Economic Modelling* **29**, 817–822.

Billingsley, P. (1999). *Convergence of Probability Measures [2nd edition]*. Wiley, New York.

Breslow, N. (1972). Contribution to the discussion of the paper by D.R. Cox. *Journal of the Royal Statistical Society B* **34**, 216–217.

Brown, L. D. and Farrell, R. H. (1990). A lower bound for the risk in estimating the value of a

probability density. *Journal of the American Statistical Association* **85**, 1147–1153.

Brownlees, C. T. and Gallo, G. M. (2008). On variable selection for volatility forecasting: The role of focused selection criteria. *Journal of Financial Econometrics* **6**, 513–539.

Claeskens, G., Croux, C. and Van Kerckhoven, J. (2007). Prediction focussed model selection for autoregressive models. *The Australian and New Zealand Journal of Statistics* **49**, 359–379.

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

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

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics* **1**, 279–290.

Durbin, J., Knott, M. and Taylor, C. C. (1975). Components of Cramer-von Mises statistics. ii. *Journal of the Royal Statistical Society B* **37**, 216–237.

Durrett, R. (2010). *Probability: Theory and Examples [4th edition]*. Cambridge University Press, Cambridge.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistcs* **7**, 1–26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Fan, J. and Marron, J. S. (1992). Best possible constant for bandwidth selection. *Annals of Statistics* **20**, 2057–2070.

Gandy, A. and Hjort, N. L. (2013). Focused information criteria for semiparametric linear hazard regression. Technical report, Department of Mathematics, University of Oslo.

Hampel, F. R., Ronchetti, E., Rousseuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hengartner, N. W. and Matzner-Løber, É. (2009). Asymptotic unbiased density estimators. *ESAIM: Probability and Statistics* **13**, 1–14.

Hermansen, G. H., Hjort, N. L. and Kjesbu, O. S. (2016). Recent advances in statistical methodology applied to the Hjort liver index time series (1859−2012) and associated influential factors. *Canadian Journal of Fisheries and Aquatic Sciences* **73**, 279–295.

Hjort, N. L. (1992). On inference in parametric survival data models. *International Statistical Review* **60**, 355–387.

Hjort, N. L. (2008). Focused information criteria for the linear hazard regression model. In Vonta, F., Nikulin, M., Limnios, N. and Huber-Carol, C., editors, *Statistical Models and Methods for Biomedical and Technical Systems* 487–502. Birkhäuser, Boston.

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

Hjort, N. L. and Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *Journal of the American Statistical Association* **101**, 1449–1464.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics* **24**, 1619–1647.

Hjort, N. L., McKeague, I. W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Annals of Statistics* **37**, 1079–1111.

Hjort, N. L. and Pollard, D. B. (1993). Asymptotics for minimisers of convex processes. Technical report, Department of Mathematics, University of Oslo.

Hotelling, H. and Solomons, L. M. (1932). The limits of a measure of skewness. *Annals of Mathematical Statistics* **3**, 141–142.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Jones, M. C., Linton, O. and Nielsen, J. P. (1995). A simple bias reduction method for density estimation. *Biometrika* **82**, 327–338.

Jullum, M. and Hjort, N. L. (2017a). Supplement to "Parametric or nonparametric: The FIC approach". *Statistica Sinica*. online `http://www3.stat.sinica.edu.tw/statistica/`

Jullum, M. and Hjort, N. L. (2017b). What price semiparametric Cox regression? *Submitted for publication*.

Lehmann, E. L. (1998). *Elements of Large-Sample Theory*. Springer-Verlag, Berlin.

Liu, W. and Yang, Y. (2011). Parametric or nonparametric? A parametricness index for model selection. *Annals of Statistics* **39**, 2074–2102.

Meier, P., Karrison, T., Chappell, R. and Xie, H. (2004). The price of Kaplan-Meier. *Journal of the American Statistical Association* **99**, 890–896.

Miller, R. G. (1981). What price Kaplan–Meier? *Biometrics* **39**, 1077–1081.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5* **50**, 157–176.

Pearson, K. (1902). On the change in expectation of life in man during a period of circa 2000 years. *Biometrika* **1**, 261–264.

Shao, J. (2003). *Mathematical Statistics [2nd edition]*. Springer-Verlag, Berlin.

Shao, J. and Tu, D. (1996). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

Sheather, S. J. and Marron, J. S. (1990). Kernel quantile estimation. *Journal of the American Statistical Association* **85**, 410–416.

van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Annals of Statistics* **39**, 174–200.

Department of Mathematics, University of Oslo, 0316, Norway

E-mail: jullum@nr.no

Department of Mathematics, University of Oslo, 0316, Norway

E-mail: nils@math.uio.no

(Received October 2015; accepted July 2016)