PARAMETRIC OR NONPARAMETRIC: THE FIC APPROACH

Martin Jullum and Nils Lid Hjort

Department of Mathematics, University of Oslo

Supplementary Material

This is a supplement to the paper Jullum and Hjort (2016). Section S1 provides proofs of the stand-alone results provided in the main paper. Section S2 describes some simulation studies investigating the performance of the FIC and AFIC schemes developed in the main paper. Section S3 gives some details on their application to categorical data, and Section S4 presents an illustration of the developed FIC scheme applied to sequential c.d.f. estimation of Pennsylvanian SAT scores. Finally, Section S5 studies the developed FIC strategy under the local misspecification framework of the original FIC (Claeskens and Hjort (2003)).

S1. Proofs of main paper results

Below we provide proofs of all stand-alone results given in the main paper. All equation numbers and other numbering refer to those in the main paper.

Proof of Proposition 2

The combined assumptions for each of the two functionals T and S and their influence functions, imply the analogues conditions for the two-dimensional functional (T, S) and its two-dimensional influence function. This is precisely the condition required by Shao (2003, Theorem 5.15) to establish consistency for the covariance matrix of $(T(\widehat{G}_n), S(\widehat{G}_n)) = (\widehat{\mu}_{\rm np}, \widehat{\mu}_{\rm pm})$. Hence, the individual variance and covariance estimators $\widehat{v}_{\rm np}$, $\widehat{v}_{\rm pm}$, $\widehat{\kappa}$ and $\widehat{v}_{\rm c}$ are all consistent. When also the conclusion of Proposition 1 holds, $\sqrt{n}(\widehat{\mu}_{\rm np} - \mu)$ and $\sqrt{n}(\widehat{\mu}_{\rm pm} - \mu_0)$ are bounded in probability. This implies consistency of $\widehat{\mu}_{\rm np}$ and $\widehat{\mu}_{\rm pm}$ for respectively μ and μ_0, and consequently also for $\widehat{b}$ and $\widehat{b}^2$ (by the continuous mapping theorem).

Proof of Proposition 3

Let us first show that (C1) holds, i.e. that T is Hadamard differentiable at G w.r.t. the uniform norm (hereafter referred to as just Hadamard). The mean functional $T_1(G) = \xi = \int h(y)\,{\rm d}G(y)$ is linear in G and hence obviously Hadamard (for any proper norm). By van der Vaart (2000, Lemma 21.3), the quantile functional $T_2(G) = G^{-1}(p)$ is also Hadamard when g is positive and continuous in $G^{-1}(p)$. The chain rule for Hadamard differentiability (van der Vaart (2000, Theorem 20.9)) ensures that smooth (continuously differentiable) functions of finitely many functionals on the form of T_1 and T_2 also are Hadamard. The functional T is thus Hadamard since A is smooth with finite partial derivatives.

Since ${\rm E}_G\{h_j(y)\} = \xi_j$ and ${\rm E}_G(\mathbf{1}_{\{y \le G^{-1}(p_l)\}}) = p_l$ for all j, l, it is easy to see that

$\mathrm{E}_G\{\mathrm{IF}(Y_i;G)\}=0$ in (3.2) whenever all partial derivatives of A are finite and all $g(G^{-1}(p_l))$ are positive – hence (C3) holds. For (C4) to hold it suffices to show that all the individual influence functions have finite variance. Since the partial derivatives are finite, this reduces to the means of $h_j(y)^2$ and $\mathbf{1}^2_{\{y\le G^{-1}(p_l)\}}=\mathbf{1}_{\{y\le G^{-1}(p_l)\}}$ needing to be finite – but the former here is assumed to be finite and the latter equals $p_l\in[0,1]$. Finally (C2) holds since

$$\partial s(\theta_0)/\partial\theta=[\{\partial A(\xi,\zeta)/\xi\}\{\partial\xi(\theta)/\partial\theta\},\{\partial A(\xi,\zeta)/\zeta\}\{\partial\zeta(\theta)/\partial\theta\}]^{\mathrm{t}},$$

which by assumption has all elements finite. ∎

Proof of Lemma 1

Let $Z_{n,j}$ denote the Z_n corresponding to the j-th parametric model. When $b_j\neq 0$ it follows from the consistency of $\widehat{b}_j$ that $Z_{n,j}$ tends to infinity in probability, while $2\widehat{\eta}_j\to_{\mathrm{pr}}2\eta_j$. Hence, $\alpha_n(G,j)=\Pr\left(Z_{n,j}\le 2\widehat{\eta}_j\right)\to 0$ for both the truncated and untruncated version of the FIC, proving the first claim. If for some j, $b_j=0$ and $b_l\neq 0$ for $l\neq j$, then we have by (2.6) (which follows from the conclusion of Proposition 1) that $\sqrt{n}\widehat{b}_j\to_d\mathrm{N}(0,\kappa_j)$. Hence $Z_{n,j}=n\widehat{b}_j^2\to_d\kappa_j\chi^2_1$. The second result then follows by consistency of $\widehat{\eta}_j$. ∎

Proof of Corollary 1

We give this proof for continuous g. The proof is analogous for discrete g or combinations of the two by replacing the appropriate integrals by sums. As we shall deal with the correct j-th model exclusively, we omit the sub- and superscript j to simplify the notation. We shall need interchangeability of integration w.r.t. y and differentiation w.r.t. θ at θ_0 a few times below; this will be justified at the end of the proof. Let us denote by $\dot{f}$ and $\ddot{f}$, respectively, the first and second derivative of f w.r.t. θ.

That the two matrices J and K of (2.1) are equal under model conditions is simply the well-known Bartlett identity (requiring $\int\ddot{f}(y;\theta_0)\,\mathrm{d}y=0$). We shall now show that $c=d$ when $G=F_{\theta_0}$. Let us for notational convenience write $H_t=(F_{\theta_0+t}-F_{\theta_0})/t$ with $F_\theta=F(\cdot\,;\theta)$. The sequence of equalities below, leading to $c=d$, is based on the definition of Hadamard differentiability and its linearity, in addition to the representation of H_t as the integral $\int(1-\delta_x(y))\,\mathrm{d}H_t(y)=\int\delta_y(x)\,\mathrm{d}H_t(y)$, with $\delta_y(x)=\mathbf{1}_{\{x\ge y\}}$. We have that

$$\begin{aligned}
c&=\lim_{t\to0}\frac{T(F_{\theta_0+t})-T(F_{\theta_0})}{t}=\lim_{t\to0}\frac{T(F_{\theta_0}+tH_t)-T(F_{\theta_0})}{t}=\lim_{t\to0}\dot{T}_G(H_t)\\
&=\lim_{t\to0}\dot{T}_G(\int\delta_y(\cdot)\,\mathrm{d}H_t(y))=\lim_{t\to0}\int\dot{T}_G(\delta_y)\,\mathrm{d}H_t(y)=\lim_{t\to0}\int\mathrm{IF}(y;G)\,\mathrm{d}H_t(y)\\
&=\lim_{t\to0}\frac{1}{t}\left\{\int\mathrm{IF}(y;G)\,\mathrm{d}F_{\theta_0+t}(y)-\int\mathrm{IF}(y;G)\,\mathrm{d}F_{\theta_0}(y)\right\}\\
&=\int\mathrm{IF}(y;G)\frac{\partial}{\partial t}f(y;\theta_0+t)\Big|_{t=0}\mathrm{d}y=\int\mathrm{IF}(y;G)u(y;\theta_0)f(y;\theta_0)\,\mathrm{d}y=d,
\end{aligned}$$

where we have interchanged differentiation and integration (twice) in the second to last equality. The conclusion then follows since $c = d$ implies $v_{\rm pm} = v_c$, which again implies $\kappa = \eta$.

What remains is to justify the interchange of differentiation and integration in $\int \ddot{f}(y;\theta_0)\,\mathrm{d}y$ and $\int \mathrm{IF}(y;G)\dot{f}(y;\theta_0)\,\mathrm{d}y$, where the former requires that this may be done in $\int \dot{f}(y;\theta_0)\,\mathrm{d}y$ as well. From the initial regularity assumptions we have that $f(y;\theta_0)$ is integrable and that $\dot{f}(y;\theta_0)$ is finite and continuous in θ_0. Observe further that $\int \sup_{\theta\in\Theta^{(j)}} \|\dot{f}(y;\theta)\|\,\mathrm{d}y = \mathrm{E}_G\left(\sup_{\theta\in\Theta^{(j)}} \|u(Y_i;\theta)\|\right)$ which by assumption is finite. Thus, Durrett (2010, Theorem A.5.2) ensures that $\dot{f}(y;\theta)$ is integrable with $\int \dot{f}(y;\theta_0)\,\mathrm{d}y = (\partial/\partial\theta)\int f(y;\theta_0)\,\mathrm{d}y$ (which is zero). These arguments may be repeated for $I\dot{f}$ and $\ddot{f}$ with proper replacements of the supremum quantities (and noting that $\ddot{f}/f = I + uu^{\rm t}$ for the latter) to show that the interchange is valid also for those quantities. The proof is hence complete. ∎

Proof of Lemma 2

Note first that $\mathrm{E}_G(\sup_{\theta\in\Theta^*} \|u(Y_i;\theta)\|)$ being finite ensures that $J = K$ by arguments in the proof of Corollary 1. Condition (C0) (which in particular implies (2.5)) and the explicitly assumed conditions then match those of Durbin (1973, Theorem 2). That theorem establishes process convergence for $B_n^0(r) = \sqrt{n}\{\widehat{G}_n(F^{-1}(r;\widehat{\theta})) - F^{-1}(r;\widehat{\theta})\}$ for $r \in [0,1]$ towards $B^0(r)$, a zero-mean Gaussian process with covariance function

$$\mathrm{Cov}\{B^0(r_1), B^0(r_2)\} = \min(r_1, r_2) - r_1 r_2 - c(F^{-1}(r_1;\theta_0);\theta_0)^{\rm t} J^{-1} c(F^{-1}(r_2;\theta_0);\theta_0).$$

Since the $F(\cdot\,;\widehat{\theta})$ process converges to $F(\cdot\,;\theta_0)$ jointly with $B_n^0(r)$, Billingsley (1999, Lemma p. 151) applied to $B_n^0(F(y;\widehat{\theta}))$ gives $B_n \xrightarrow{d} B$ as stated in the lemma. The same argument, in addition to the continuity of the integral of a square function, shows that $C_{1,n} \xrightarrow{d} C_1$, and when $\int G(y)\{1-G(y)\}\,\mathrm{d}y$ is finite also $C_{2,n} \xrightarrow{d} C_2$. ∎

S2. Illustration: Running through the c.d.f. of Pennsylvanian SAT scores

Consider a sample of $n = 100$ school averaged grades from the 'writing' series of the SAT standardized test in Pennsylvania schools in 2009. For learning and illustrational purposes we will here *sequentially* consider the focus parameters $\mu(y) = G(y)$ for different y-values. As competing models we put up the nonparametric, the Gaussian and the skewed Gaussian model, where the skewed Gaussian model has cumulative distribution function $\Phi((y-\xi)/\sigma)^\lambda$, with $\Phi(\cdot)$ the cumulative distribution of the standard normal distribution. In particular, $\widehat{\lambda} = 0.290$ for these data, reflecting a small skewness to the right compared to the Gaussian model. Summary plots are provided in Figure S1. As the figure shows, the parametric distributions are deemed better than the nonparametric alternative mainly close to the tails of the distribution and some intervals on each side of the median. Also, the skewed Gaussian distribution is mostly better than the regular Gaussian one, except in the longest tails where

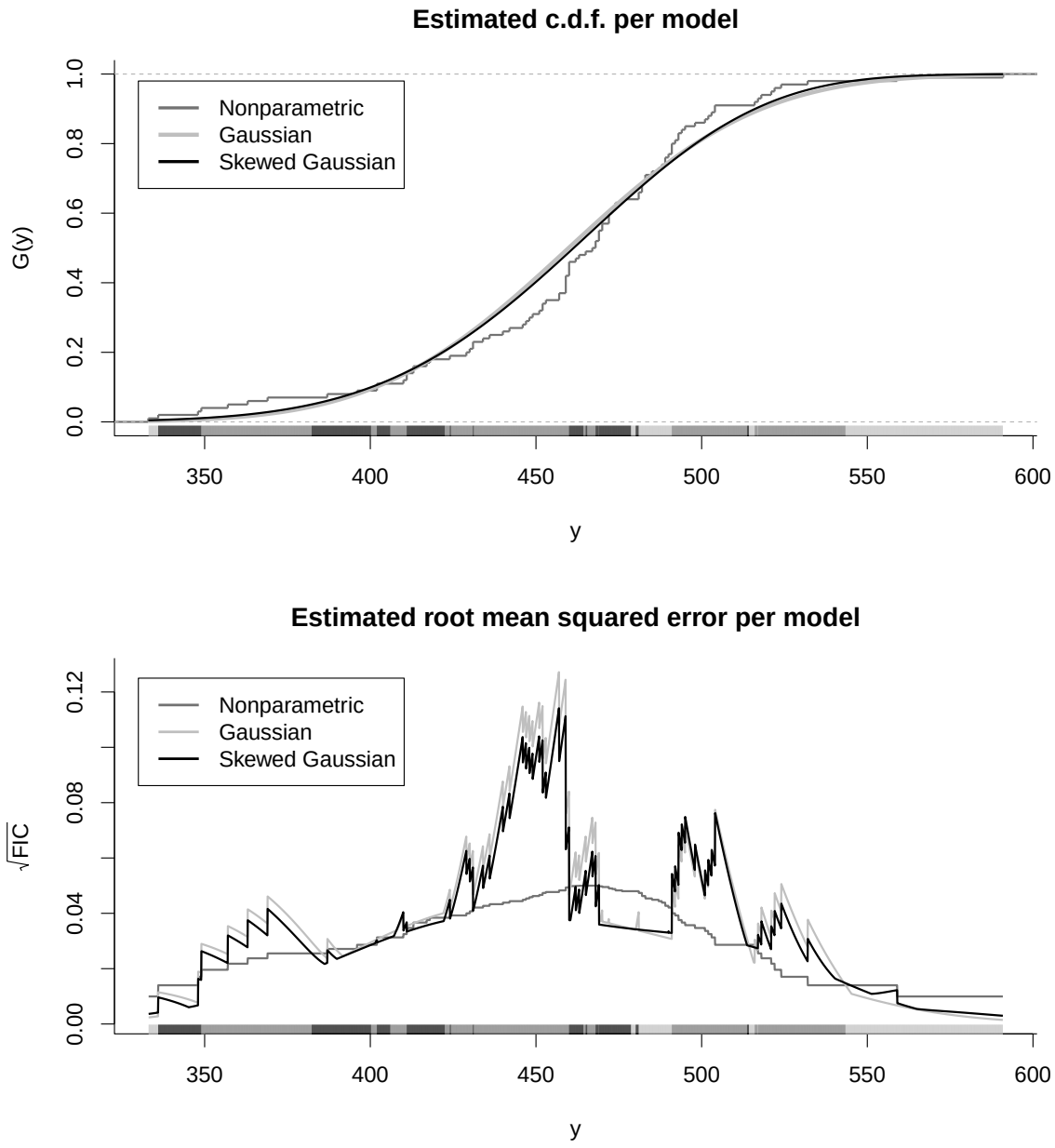

Figure S1: Summary plots for the FIC applied to data on writing skills in college for different y-values of the focus parameter $\mu(y) = G(y)$. For the three candidate models the upper plot shows the estimated c.d.f's per model, while the lower plot shows the root of the FIC score. The greyscale bar in the bottom indicates which model is deemed the best for each value of y.

the estimates are essentially identical and the Gaussian model is slightly better due to lower variance. The rooted FIC score for the parametric distributions are rather noisy. This is caused by the nature of the parametric FIC formula shifting in the jumps of $\widehat{\mu}_{\rm np}(y)$, i.e. at the data points.

Instead of focusing on $G(y)$ for one y at a time, one could of course use the AFIC of Section 4 in the main paper to find the overall best model for say some interval of y values. Let us take the neutral position here and provide equal weight 1 to all positions on the positive half line. The AFIC then decides that the nonparametric model is the relatively clear winner with a rooted AFIC score of 0.495. The skewed Gaussian model and usual Gaussian model are respectively the runner-up and losing models in terms of overall performance, having rooted AFIC scores of 0.648 and 0.712. This result is not surprising considering the smaller rooted FIC scores of the nonparametric model shown in the lower plot of Figure S1.

S3. Simulation experiments

Practical (finite-sample) performance analysis is somewhat more complicated with the FIC and AFIC than with other model selection criteria, since one also needs to decide upon the focus parameter. In addition, the inclusion of a nonparametric alternative makes it difficult to compare our criteria with frequently used parametric model selection criteria like the AIC and the BIC under fair and realistic comparison terms. Due to these circumstances, we will consider two different scenarios for simulation based performance analysis; one focusing solely on the performance of the best ranked *parametric* candidate models, and one allowing the nonparametric candidate to be selected as well.

In these simulation studies we shall be working with *one* fixed true model. To mimic a realistic situation this data-generating model will not be among the parametric candidate models. Different scenarios are then simulated by varying the sample size and considering different focus parameters. We shall consider three different parametric models: the exponential, the gamma, and the Weibull, and take the true data-generating distribution to be a mixture of the gamma and the Weibull, having density $g(y) = \frac{1}{2} f_{\rm gam}(y; 1.5, 1) + \frac{1}{2} f_{\rm wei}(y; 1.5, 1.6)$, where $f_{\rm gam}(y; \alpha, \beta) = \{\beta^\alpha/\Gamma(\alpha)\} y^{\alpha-1} \exp(-\beta y)$ and $f_{\rm wei}(y; \alpha, \beta) = \alpha\beta(y\beta)^{\alpha-1} \exp\{-(\beta y)^\alpha\}$. The density g and cumulative distribution G is plotted in Figure S2 along with least false versions of the parametric candidate models. The minimum Kullback–Leibler divergence from the true distribution to the exponential, gamma and Weibull distribution classes are, respectively, 0.045, 0.017 and 0.025. This reflects that in terms of the Kullback–Leibler divergence, the gamma distribution is closest to the true distribution, while the exponential distribution is the most distant. None of the parametric models are however very far from the truth, at least for most of the sample space. This is indeed intended, as parametric models far from

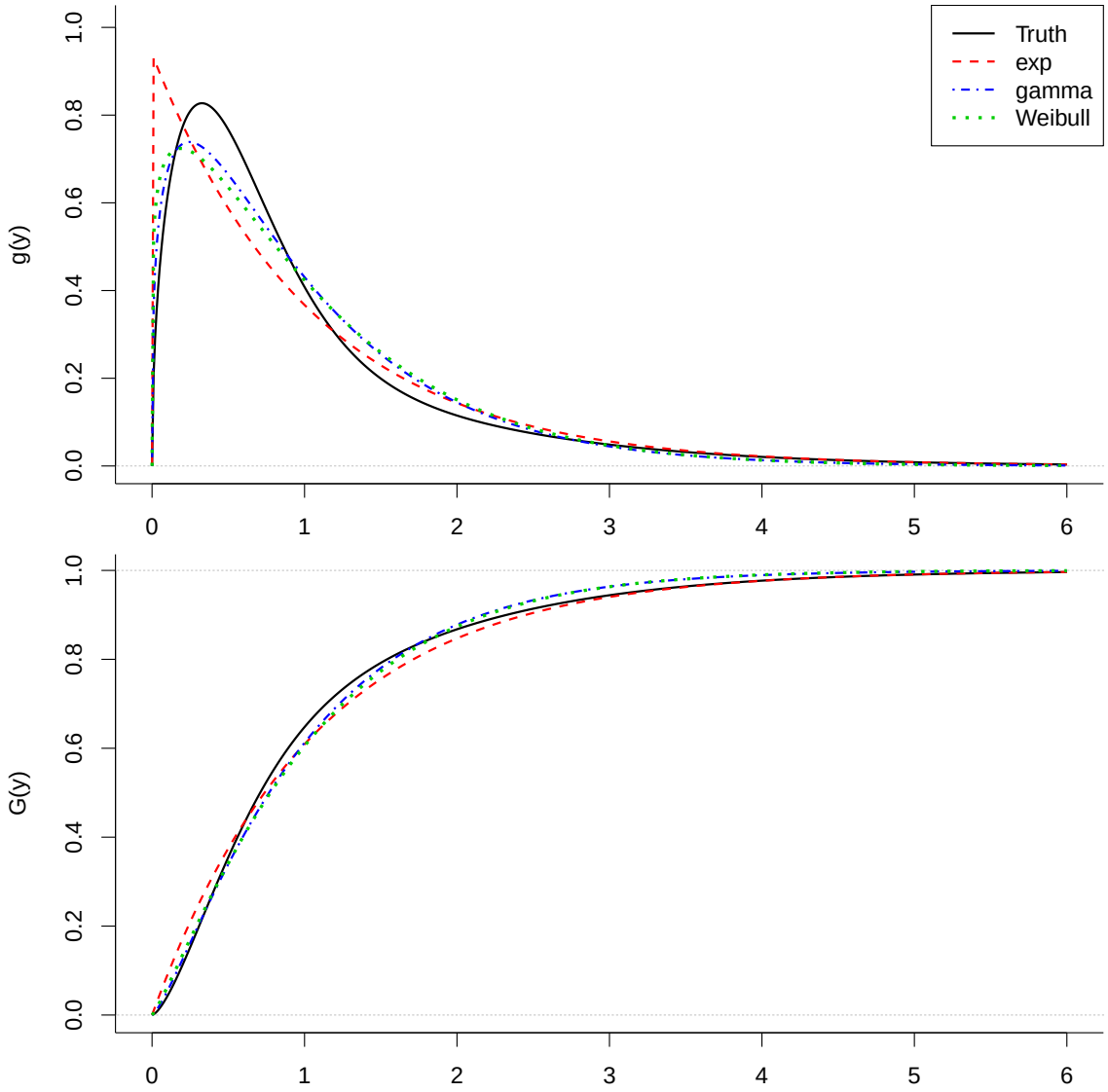

Figure S2: The density and cumulative distribution for the true data generating distribution used in the simulation experiment, are plotted along with the least false versions of the parametric candidate models.

the truth are not of much interest from a practical point of view, and would anyhow seldom be selected.

S3.1. Parametric selection performance

Let us first compare the performance of the estimators based on the FIC and AFIC rankings to those selected by the AIC and the BIC. To make this comparison completely fair, we shall, as mentioned, exclude the nonparametric candidate model from the final ranking of

the FIC and AFIC. This causes all criteria to have the same set of candidate models to choose from. To span a broad range of scenarios, we study performance when the focus parameter is $\mu(y) = G(y)$ for y-values ranging from 0 to 6. For completeness we include both the truncated version of the FIC scheme (cf. (2.8) in the main paper) and the untruncated analogue FIC*. We also include two different weight functions W in the comparison. AFIC no. 1 puts equal weight on y from 0 to 6 (i.e. W is the Lebesgue measure on that interval), and AFIC no. 2 uses $W(y) = G(y)$, which is estimated via $\widehat{G}_n(y)$ for each simulated data set. Figure S3 shows the performance of the different criteria in terms of the empirical root mean squared error of the focus parameters for two different sample sizes, $n = 75$ and $n = 250$. As seen from the plots, the truncated and untruncated versions of the FIC perform similarly. They both clearly outperform the AIC and the BIC for μ between 0.40 and 0.75 for both sample sizes. AIC performs somewhat better for some smaller values of μ. BIC, however, is significantly better only in a tiny interval around 0.35 for $n = 250$. Overall, the performance of both versions of the FIC are better than the BIC, and comparable or perhaps slightly better than the AIC. As expected, the two fairly unfocused AFIC versions varies more in terms of performance. Their overall performance is not as good as the FIC, but comparable with the AIC and BIC.

S3.2. Full performance comparison

Even though our criteria may very well be used solely to choose the best parametric model and estimator, we would in practice typically trust the nonparametric candidate if it actually has lower estimated risk. We shall now compare the performance of estimators based on the full version of the FIC with those based on other model selection criteria. We compare the FIC with the AIC and BIC (which only select among the parametric models), in addition to a heuristically defined Kolmogorov–Smirnov (KS) criterion, and the model-selection-ignorant estimator always trusting $\widehat{\mu}_{\rm np}$. The KS criterion is included to have the FIC compared with 'some' other criterion selecting among parametrics and nonparametrics. It chooses the parametric model with smallest KS statistic $\Delta(F) = \sup_y |\widehat{G}_n(y) - F(y;\widehat{\theta})|$ if it is smaller than a specified threshold value. To follow tradition we set this threshold value to $1.3581/\sqrt{n}$ – corresponding to an asymptotic significance level of 0.05 in a Kolmogorov–Smirnov test of *one* parametric model with fixed parameters. Figure S4 shows empirical root mean squared error comparisons for four different focus parameters: the median, the 0.95-quantile, the standard deviation and the skewness.

Although not illustrated here, the untruncated FIC* performs comparably with the FIC for all these cases. As seen from the plots, the FIC performs well compared to the other criteria. For small sample sizes, other criteria perform somewhat better for two of the focus parameters. The FIC generally performs better than using the nonparametric estimator

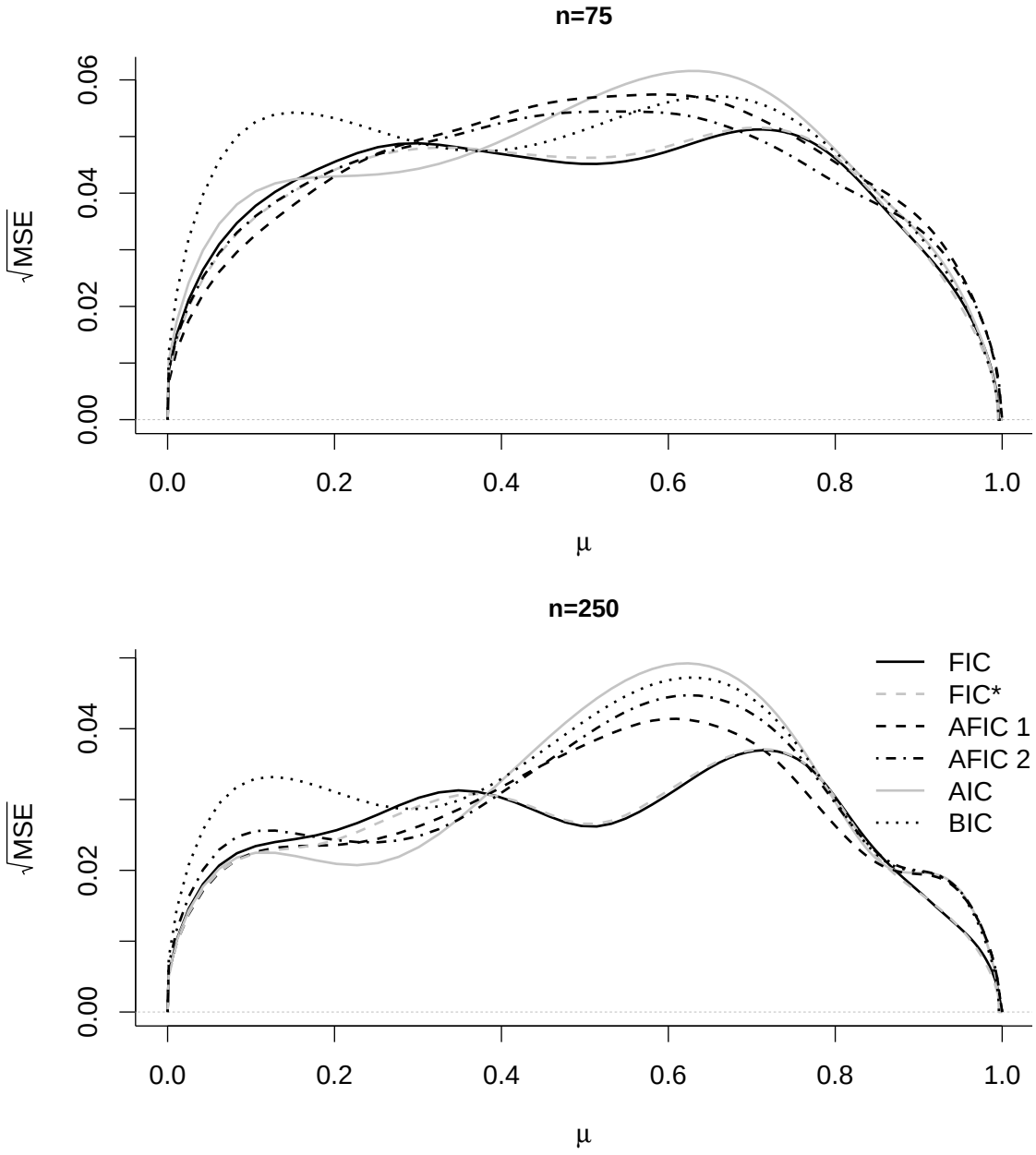

Figure S3: Empirical root mean squared errors for $\widehat{\mu}$ chosen by the FIC, FIC*, AFIC 1, AFIC 2, AIC and BIC when $\mu(y) = G(y)$. The empirical root mean squared errors are smoothed with spline functions and plotted as functions of actual values of $\mu(y) = G(y)$, rather than y directly, in order to show the differences more clearly. For both $n = 75$ and $n = 250$, 10^3 samples are used to compute the empirical error on a grid from $y = 0.01$ to $y = 6$ with 200 equally spaced points.

directly, except when a large sample is used to estimate the skewness. Note also that since none of these models have zero bias, Lemma 1 in the main paper ensures that increasing the

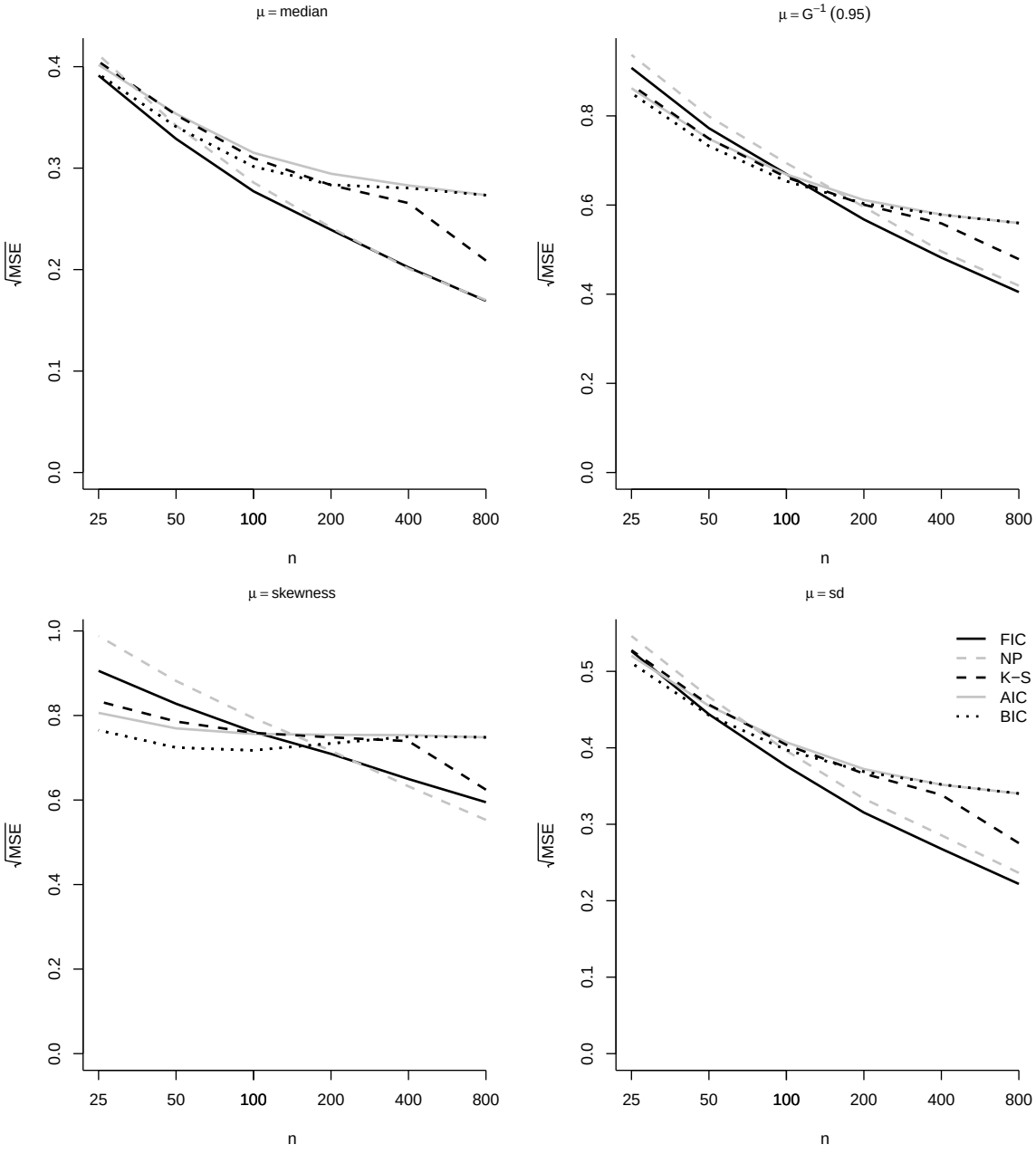

Figure S4: Empirical root mean squared errors for $\widehat{\mu}$ are shown for estimators chosen by the FIC, the KS criterion, AIC, BIC and using $\widehat{\mu}_{\rm np}$ directly, for four different focus parameters. For $n = 25$, a total of 2×10^4 simulations are used. As n increases, the empirical mean squared error becomes more stable and the number of simulations are reduced (as the computational cost increases), reaching 10^3 simulations for $n = 800$.

sample size further would make the FIC-based estimator and the nonparametric estimator coincide as the best estimator. The KS-based estimator also has this property.

S4. Details on FIC and AFIC for categorical data

Models for categorical data abound in statistical literature, both regarding probability distributions on integers, for cross-tables of various orders, and more general multinomial setups; see e.g. Agresti (2002) and various review articles mentioned in that book. We shall now see how the FIC and especially a certain version of the AFIC scheme work out for this framework.

Consider counts $N = (N_1, \ldots, N_k)$ from a multinomial model with probability vector $(p_1, \ldots, p_k)$, with $\sum_{j=1}^k p_j = 1$ and $\sum_{j=1}^k N_j = n$. Since N is a sum of i.i.d. variables, this situation may be handled by the FIC scheme developed in Section 2 of the main paper. Focus parameters for this model take the form $\mu = A(p_1, \ldots, p_k)$ for suitable smooth functions A. These are functions of means and therefore fit our scheme (cf. Proposition 3 in the main paper) with nonparametric and parametric plug-in estimators of the p_j. The natural nonparametric estimator of p_j is $\bar{p}_j = N_j/n$. The parametric alternatives are of the form $\widehat{p}_j = f_j(\widehat{\theta})$, for some postulated $f_j(\theta)$ with θ of dimension say $q \le k-2$, where $\sum_{j=1}^k f_j(\theta) = 1$. In particular, the FIC method may be used to check if a parametric model leads to a better estimate of p_j itself than the direct $\bar{p}_j$. Writing $\psi_j(\theta) = \partial \log f_j(\theta)/\partial\theta$, we have

$$\mathrm{mse}_{\mathrm{np}} = p_j(1-p_j)/n \quad \text{and} \quad \mathrm{mse}_{\mathrm{pm}} = b_j^2 + v_{\mathrm{pm},j}/n,$$

with $b_j = f_j(\theta_0) - p_j$ and $v_{\mathrm{pm},j} = f_j(\theta_0)^2 \psi_j(\theta_0)^{\mathrm{t}} J^{-1} K J^{-1} \psi_j(\theta_0)$, in terms of the least false parameter associated with the model, characterised by $\sum_{j=1}^k p_j \psi_j(\theta_0) = 0$. Following the recipe of Section 2 of the main paper for these multinomial models (see (2.1) there) one also finds

$$J = -\sum_{j=1}^k p_j \frac{\partial^2 \log f_j(\theta_0)}{\partial\theta\partial\theta^{\mathrm{t}}} \quad \text{and} \quad K = \sum_{j=1}^k p_j \psi_j(\theta_0)\psi_j(\theta_0)^{\mathrm{t}}.$$

The methods of Section 2 then yield FIC scores, i.e. estimates of the mse, for deciding which model is best for estimating any given p_j.

Let us next turn to the AFIC, with focus on the full probability vector $(p_1, \ldots, p_k)$ and with weights equal to their inverses. This specific setting turns out to be closely connected to the classical Pearson chi-squared tests, as indicated in the main paper's Remark 2. Per the AFIC strategy, the aim is to compare models by assessing their associated risk functions

$$\mathrm{risk} = \mathrm{risk}(p_1, \ldots, p_k) = \mathrm{E}_G\Big\{ \sum_{j=1}^k (\widehat{p}_j - p_j)^2/p_j \Big\}.$$

Note that since $\mathrm{E}_G\{(\bar{p}_j - p_j)^2\} = p_j(1-p_j)/n$, the nonparametric risk is exactly $(k-1)/n$

(with no further need for risk estimation). A parametric model needs to have estimated risk below this threshold in order to be judged better than the default nonparametric one. Applying the standard AFIC strategy of Section 4 in the main paper leads to

$$\mathrm{AFIC}_{\mathrm{pm}} = \max\left\{0, \sum_{j=1}^k \frac{1}{\bar p_j}\left(\widehat b_j^2 - \frac{\widehat v_{\mathrm{np},j} + \widehat v_{\mathrm{pm},j} - 2\widehat v_{\mathrm{c},j}}{n}\right)\right\} + \sum_{j=1}^k \frac{\widehat v_{\mathrm{pm},j}}{n\bar p_j},$$

where once again subscript j indicates affiliation with p_j. Here $v_{\mathrm{c},j} = f_j(\theta_0)p_j\psi_j(\theta_0)^{\mathrm{t}}J^{-1}\psi_j(\theta_0)$. Introducing the new quantity $K^* = \sum_{j=1}^k f_j(\theta_0)\psi_j(\theta_0)\psi_j(\theta_0)^{\mathrm{t}}$, some re-arranging shows that $\sum_{j=1}^k \widehat v_{\mathrm{c},j}/\bar p_j = \mathrm{Tr}(\widehat J^{-1}\widehat{K^*})/n$. Here $\widehat J$ is the usual plug-in version of J, and $\widehat{K^*}$ likewise the empirical analogue of K^* where $\widehat\theta$ is plugged in for θ_0. Note next that the classical Pearson chi-squared type statistics traditionally come in two forms, namely

$$\sum_{j=1}^k \frac{\{N_j - nf_j(\widehat\theta)\}^2}{nf_j(\widehat\theta)} = n\sum_{j=1}^k \frac{\widehat b_j^2}{f_j(\widehat\theta)} \quad \text{and} \quad \sum_{j=1}^k \frac{\{N_j - nf_j(\widehat\theta)\}^2}{N_j} = n\sum_{j=1}^k \frac{\widehat b_j^2}{\bar p_j}.$$

These are large-sample equivalent under model conditions, then both tending to χ^2_{df} with $\mathrm{df} = k-1-q$, but have different distributions outside the model. Our AFIC approach, aiming at consistent risk estimation to determine whether a parametric model is good enough, is seen to be closely related to the second of these two, say X_n.

Arguments similar to those used to derive the limit behaviour in Section 5 in the main paper show that (both versions of) the AFIC prefer the parametric model over the nonparametric whenever $X_n \le 2\{(k-1) - \mathrm{Tr}(\widehat J^{-1}\widehat{K^*})\}$. When the parametric model is fully correct, both $\widehat{K^*}$ and $\widehat J$ tend to $J = K$ in probability, such that $\mathrm{Tr}(\widehat J^{-1}\widehat{K^*}) \to_{\mathrm{pr}} q = \dim(\theta)$. This implies that when a certain parametric model is true, the AFIC criterion selects this parametric model over the nonparametric with a probability tending to $\Pr(\chi^2_{\mathrm{df}} \le 2\,\mathrm{df})$. Hence, this version of the AFIC sheds new light on the classical Pearson chi-squared test, as discussed in Remark 2 of the main paper.

A particularly enlightening special case is that of assessing independence in an $r \times s$ table, i.e. the hypothesis that the cell probabilities $p_{i,j}$ can be expressed as $\alpha_i\beta_j$ for all (i,j). Some of the matrix calculations above simplify for this case. In particular, both the J and the K^* matrix are found to be equal to $\mathrm{diag}(A,B)$, say, with blocks A and B of sizes $(r-1)\times(r-1)$ and $(s-1)\times(s-1)$ respectively, and with elements $a_{i,j} = \alpha_i^{-1}\mathbf{1}_{\{i=j\}} + \alpha_r^{-1}$ and $b_{i,j} = \beta_i^{-1}\mathbf{1}_{\{i=j\}} + \beta_s^{-1}$. In particular, the trace of $J^{-1}K^*$ is $r+s-2$, regardless of whether the independence model is correct or not. This leads to the following AFIC recipe: Accept independence precisely when $X_n \le 2(r-1)(s-1)$, where $X_n = \sum_{i,j}(N_{i,j} - n\widehat\alpha_i\widehat\beta_j)^2/N_{i,j}$ is the classical chi-squared test for independence, with $\widehat\alpha_i = N_{i,\cdot}/n$ and $\widehat\beta_j = N_{\cdot,j}/n$. Again, this criterion has been established via risk assessments only, and the usual goodness-of-fit thinking involving the null distribution of X_n etc. do not enter the argument.

S5. FIC in a local asymptotics framework

Although we left the local misspecification framework for deriving the FIC in the main paper, such a framework may be useful for studying limiting properties and especially for comparing the new version of the FIC with the original FIC of Claeskens and Hjort (2003). To ease representation for readers familiar with the original FIC, and also since we shall re-use results and theory from its development, we will to a large extent adopt the notation used in Claeskens and Hjort (2008).

S5.1. Limiting distributions for the original and new FIC

Consider the local misspecification framework of Claeskens and Hjort (2003); Hjort and Claeskens (2003) where the true distribution $G = G_n$ has density or probability mass function

$$g_n(y) = f(y; \theta_0, \gamma_0 + \delta/\sqrt{n}), \tag{S5.1}$$

where the θ and γ parameters are of dimension p and q respectively. The γ_0 here is the 'null value' of γ for which $f(y;\theta,\gamma_0) = f(y;\theta)$, i.e. where the model reduces to the simplest 'narrow' model with p parameters described by θ. The δ parameter denotes the $O(1/\sqrt{n})$ distance from the truth to the narrow model in the direction of the γ-parameter. Assume that all parametric candidate models under consideration belong to this class. Thus, the biggest (wide) model uses all the $p+q$ parameters, the narrow model uses only the p first parameters, while there are possibly $2^q - 2$ other candidate (sub)models in between these. As all parametric models are nested we denote the different parametric submodels by subscript S (rather than the more generic 'pm' notation used elsewhere) to emphasise the nesting. We shall also on occasion use 'narr' and 'wide' for the smallest and biggest model. Before continuing, note that when δ is zero (in all q dimensions), the framework is no longer locally misspecified, but yielding the important special case where the narrow model is correct for all n. The results derived below will hold also for this special case.

The motivation for working with the local misspecification framework when deriving the original FIC stems from the fact that, in some sense, the limit of $n\,\mathrm{mse}(\widehat{\mu}_S)$ stabilises within this framework. This is naturally also the case for reasonable estimators of this quantity. It is therefore most convenient to scale the FIC formulae developed in the main paper by a factor n when studying their limit properties.

Before presenting the FIC formulae for the two variants, let us introduce some notation. Let I_k denote the k-dimensional identity matrix. Let then π_S be the appropriate $|S| \times q$ dimensional matrices of zeros and ones extracting the subvectors (and submatrices) corresponding to submodel S by multiplying the full q dimensional vectors (and $q \times q$ dimensional matrices) by π_S from the left (and π_S^{t} from the right). Let us also write $c_0 = \partial\mu(F_{\theta,\gamma})/\partial\theta$

and $c_1 = \partial\mu(F_{\theta,\gamma})/\partial\gamma$ for the partial derivatives of μ taken at (θ_0, γ_0). Next,

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix} \quad \text{and} \quad J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}$$

is respectively the Fisher information matrix of the wide model evaluated at the narrow model, and its inverse. It will also turn out convenient to define $Q = J^{11}$, $\omega = J_{10}J_{00}^{-1}c_0 - c_1$ and $G_S = \pi_S^{\mathrm{t}}(\pi_S^{\mathrm{t}}Q^{-1}\pi_S)^{-1}\pi_S Q^{-1}$. Let finally $(U(y)^{\mathrm{t}}, V(y)^{\mathrm{t}}\pi_S^{\mathrm{t}})^{\mathrm{t}}$ denote the score function of model S evaluated at the narrow model. All these quantities also have empirical analogues based on inserting wide model estimates $\widehat{\theta}$ and $\widehat{\gamma}$ for θ_0 and γ_0.

The truncated formulae for the original FIC's estimator of $n\,\mathrm{mse}(\widehat{\mu}_S)$ may then be written as

$$\begin{aligned} \mathrm{FIC}_{S,\mathrm{orig}} &= \max\{0, (\widehat{\omega}^{\mathrm{t}}(I_q - \widehat{G}_S)D_n)^2 - (\widehat{v}_{\mathrm{wide}} - \widehat{v}_S)\} + \widehat{v}_S, \\ \mathrm{FIC}_{\mathrm{wide,orig}} &= \widehat{v}_S \end{aligned} \tag{S5.2}$$

with $D_n = \sqrt{n}(\widehat{\gamma} - \gamma_0)$, and $\widehat{\omega}$ and $\widehat{G}_S$ empirical analogues of ω and G_S. Further, $\widehat{v}_S$ and $\widehat{v}_{\mathrm{wide}}$ are empirical analogues of respectively $v_S = \tau_0^2 + \omega^{\mathrm{t}}G_S Q G_S^{\mathrm{t}}\omega$ and $v_{\mathrm{wide}} = \tau_0^2 + \omega^{\mathrm{t}}Q\omega$, where $\tau_0^2 = c_0^{\mathrm{t}}J_{00}^{-1}c_0$ is the variance associated with the narrow model. The truncated formulae for the new n-scaled FIC scores (estimating $n\,\mathrm{mse}(\widehat{\mu}_S)$) are

$$\begin{aligned} \mathrm{FIC}_{S,\mathrm{new}} &= \max\{0, (\sqrt{n}\widehat{b}_S)^2 - (\widehat{v}_{\mathrm{np}} + \widehat{v}_{S,\mathrm{new}} - 2\widehat{v}_{c,S})\} + \widehat{v}_{S,\mathrm{new}}, \\ \mathrm{FIC}_{\mathrm{np,new}} &= \widehat{v}_{\mathrm{np}} \end{aligned} \tag{S5.3}$$

where $\widehat{v}_{\mathrm{np}}$ is the estimated analogue of v_{np}, which under the limit of (S5.1) converges to the expectation $\mathrm{E}_{F_{\theta_0}}\{\mathrm{IF}(Y_i; F_{\theta_0})^2\}$. Following the notation of this section, a subscript S is appended to the parametric quantities. An additional subscript 'new' is also appended to the variance for the parametric submodels here, to distinguish it clearly from the corresponding quantity for the original FIC. Despite this slight change of notation, the estimators are exactly as specified in the main paper.

To study and compare the limiting behaviour of these FIC scores under the framework of (S5.1), we need to derive the joint limit of $\Lambda_{n,\mathrm{np}} = \sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \mu_n)$ and $\Lambda_{n,S} = \sqrt{n}(\widehat{\mu}_S - \mu_n)$, with $\mu_n = \mu(G_n)$. To this end, let (C0*–C4*) denote the analogues of conditions (C0–C4) from the main paper, with θ replaced by $(\theta^{\mathrm{t}}, \gamma^{\mathrm{t}})^{\mathrm{t}}$ and G replaced by F_{θ_0}. Define in addition the following new conditions:

(C5*) The variables $|U_j(Y_i)^2 V_k(Y_i)|$, $|V_j(Y_i)^2 V_k(Y_i)|$ and $|\mathrm{IF}(Y_i; F_{\theta_0})^2 V_k(Y_i)|$ have finite means under F_{θ_0} for each dimension j, k;

(C6*) The conclusion of Proposition 2 in the main paper holds with G replaced by F_{θ_0}, i.e. $\widehat{v}_{\mathrm{pm}}, \widehat{v}_{\mathrm{np}}, \widehat{v}_c, \widehat{v}_\kappa, \widehat{b}$, and $\widehat{b^2}$ are all consistent under F_{θ_0}.

To state the joint limit of $\Lambda_{n,\mathrm{np}}$ and $\Lambda_{n,S}$, we shall need the following stochastically independent variables: $\Lambda_0 \sim \mathrm{N}(0,\tau_0^2)$, $D \sim \mathrm{N}_q(\delta,Q)$ and $X_0 \sim \mathrm{N}(0, v_{\mathrm{np}} - v_{\mathrm{wide}})$. That $v_{\mathrm{np}} \geq v_{\mathrm{wide}}$ and indeed $v_{\mathrm{np}} \geq v_S$ for all S follow from details of the proof given below, as in the remark following Proposition 1 in the main paper. The X_0 is in fact the limiting distribution of $X_{n,0} = \sqrt{n}(\widehat{\mu}_{\mathrm{np}} - \widehat{\mu}_{\mathrm{wide}})$, as borne out from the proof.

Proposition S1. *Assume that (C0*–C5*) hold. Under the framework of* (S5.1)*, we have that as* $n \to \infty$

$$\begin{aligned}\Lambda_{n,S} &\xrightarrow{d} \Lambda_S = \Lambda_0 + \omega^{\mathrm{t}}(\delta - G_S D) \sim \mathrm{N}(\omega^{\mathrm{t}}(I_q - G_S)\delta, v_S),\\ \Lambda_{n,\mathrm{np}} &\xrightarrow{d} \Lambda_{\mathrm{np}} = \Lambda_{\mathrm{wide}} + X_0 = \Lambda_0 + \omega^{\mathrm{t}}(\delta - D) + X_0 \sim \mathrm{N}(0, v_{\mathrm{np}}),\end{aligned}$$

converging jointly and with limiting covariance $\mathrm{Cov}(\Lambda_S, \Lambda_{\mathrm{np}}) = v_S$.

Proof. Note first that the assumed conditions (C0*–C2*) and (C5*) imply the conditions specified in Hjort and Claeskens (2003, Section 11) being used to prove the lemmas there and also the core results given in Claeskens and Hjort (2008, Chapters 6–7). Hence, $\Lambda_{n,S} \to_d \Lambda_S$ with the proposed form follows by Hjort and Claeskens (2003, Lemma 3.3). We shall now show that $\Lambda_{n,\mathrm{np}} \to_d \Lambda_{\mathrm{np}}$, with Λ_{np} on the given form. Observe that under the assumed conditions, a Taylor expansion gives $\mu_n = \mu_0 + c_1^{\mathrm{t}}\delta/\sqrt{n} + o(1/\sqrt{n})$, where $\mu_0 = \mu(F_{\theta_0,\gamma_0})$. Another Taylor argument allows us to write

$$\Lambda_{n,\mathrm{np}} = \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \mathrm{IF}(Y_i; F_{\theta_0}) - c_1^{\mathrm{t}}\delta + o_{\mathrm{pr}}(1). \tag{S5.4}$$

To arrive at the precise form of Λ_{np}, we need to check the limiting behaviour of the sum $\sum_{i=1}^n n^{-1/2}\mathrm{IF}(Y_i; F_{\theta_0})$ jointly with $\sum_{i=1}^n n^{-1/2}\{U(Y_i)^{\mathrm{t}}, V(Y_i)^{\mathrm{t}}\}^{\mathrm{t}}$. By the assumed conditions, we may write $g_n(y) = f(y;\theta_0)\{1 + V(y)\delta/\sqrt{n} + R(y;\delta/\sqrt{n})\}$ for some R with the property that $f(y;\theta_0)R(y;t) = o(\|t\|^2)$ uniformly in y. Extending the arguments in Hjort and Claeskens (2003, Section 11) yields

$$\begin{aligned}
&\mathrm{E}_{G_n}\{\mathrm{IF}(Y_i; F_{\theta_0})\} = c_1^{\mathrm{t}}\delta/\sqrt{n} + o(1/\sqrt{n}), && \mathrm{E}_{G_n}\{U(Y_i)\} = J_{01}\delta/\sqrt{n} + o(1/\sqrt{n}),\\
&\mathrm{E}_{G_n}\{V(Y_i)\} = J_{11}\delta/\sqrt{n} + o(1/\sqrt{n}), && \mathrm{E}_{G_n}\{\mathrm{IF}(Y_i; F_{\theta_0})^2\} = v_{\mathrm{np}} + o(1),\\
&\mathrm{E}_{G_n}\{\mathrm{IF}(Y_i; F_{\theta_0})U(Y_i)\} = c_0 + o(1), && \mathrm{E}_{G_n}\{\mathrm{IF}(Y_i; F_{\theta_0})V(Y_i)\} = c_1 + o(1),\\
&\mathrm{E}_{G_n}\{U(Y_i)^{\mathrm{t}}U(Y_i)\} = J_{00} + o(1), && \mathrm{E}_{G_n}\{U(Y_i)^{\mathrm{t}}V(Y_i)\} = J_{01} + o(1),\\
&\mathrm{E}_{G_n}\{V(Y_i)^{\mathrm{t}}V(Y_i)\} = J_{11} + o(1).
\end{aligned}$$

We have here in particular used that $\mathrm{IF}(Y_i; F_{\theta_0})U(Y_i)$ and $\mathrm{IF}(Y_i; F_{\theta_0})V(Y_i)$ have means c_0 and c_1, which follows by arguments similar to those used to prove $c = d$ in the proof

of Corollary 1 in the main paper. The triangular Lindeberg conditions are fulfilled for $n^{-1/2}\{\mathrm{IF}(Y_i;F_{\theta_0}),U(Y_i)^{\mathrm{t}},V(Y_i)^{\mathrm{t}}\}^{\mathrm{t}}$ (cf. Hjort and Claeskens (2003, Section 11)), which ensures that

$$\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\begin{pmatrix}\mathrm{IF}(Y_i;F_{\theta_0})\\U(Y_i)\\V(Y_i)\end{pmatrix}\xrightarrow{d}\begin{pmatrix}X'\\U'\\V'\end{pmatrix}\sim \mathrm{N}_{1+p+q}\left(\begin{pmatrix}c_1^{\mathrm{t}}\delta\\J_{01}\delta\\J_{11}\delta\end{pmatrix},\begin{pmatrix}v_{\mathrm{np}}&c_0^{\mathrm{t}}&c_1^{\mathrm{t}}\\c_0&J_{00}&J_{10}\\c_1&J_{01}&J_{11}\end{pmatrix}\right). \tag{S5.5}$$

The limit of the first element above inserted into (S5.4) gives $\Lambda_{n,\mathrm{np}}\to_d\Lambda_{\mathrm{np}}=X'-c_1^{\mathrm{t}}\delta\sim\mathrm{N}(0,v_{\mathrm{np}})$ by Slutsky's theorem. As in Claeskens and Hjort (2008, Chapter 6.3), let $\Lambda_0=c_0^{\mathrm{t}}J_{00}^{-1}U'$ and $D=\delta+Q(V'-J_{10}J_{00}^{-1}U')$. By arguments used in Claeskens and Hjort (2008, Chapter 6.3) and Hjort and Claeskens (2003, Section 11), these are independent and have the correct limit. Define in addition

$$X_0=X'-c_1^{\mathrm{t}}\delta-\Lambda_0-\omega^{\mathrm{t}}(\delta-D).$$

We need to check that this definition of X_0 is stochastically independent of both Λ_0 and D, and have the variance claimed. The following two covariances will be helpful:

$$\begin{aligned}\mathrm{Cov}(X',\Lambda_0)&=\mathrm{Cov}(X',c_0^{\mathrm{t}}J_{00}^{-1}U')=c_0^{\mathrm{t}}J_{00}^{-1}\mathrm{Cov}(X',U')=c_0^{\mathrm{t}}J_{00}^{-1}c_0=\tau_0^2,\\\mathrm{Cov}(X',D)&=\mathrm{Cov}\{X',Q(V'-J_{10}J_{00}^{-1}U')\}\\&=[Q\{\mathrm{Cov}(X',V')-J_{10}J_{00}^{-1}\mathrm{Cov}(X',U')\}]^{\mathrm{t}}\\&=\{Q(c_1-J_{10}J_{00}^{-1}c_0)\}^{\mathrm{t}}=-\omega^{\mathrm{t}}Q.\end{aligned}$$

Thus, we also get

$$\begin{aligned}\mathrm{Cov}(X_0,\Lambda_0)&=\mathrm{Cov}\{X'-c_1^{\mathrm{t}}\delta-\Lambda_0-\omega^{\mathrm{t}}(\delta-D),\Lambda_0\}\\&=\mathrm{Cov}(X',\Lambda_0)-\mathrm{Cov}(\Lambda_0,\Lambda_0)=\tau_0^2-\tau_0^2=0\\\mathrm{Cov}(X_0,D)&=\mathrm{Cov}\{X'-c_1^{\mathrm{t}}\delta-\Lambda_0-\omega^{\mathrm{t}}(\delta-D),D\}\\&=\mathrm{Cov}(X',D)+\omega^{\mathrm{t}}\mathrm{Cov}(D,D)=-\omega^{\mathrm{t}}Q+\omega^{\mathrm{t}}Q=0,\end{aligned}$$

proving stochastic independence between X' and both Λ_0 and D. The variance of X_0 is found by

$$\begin{aligned}\mathrm{Var}(X_0)&=\mathrm{Var}\{X'-c_1^{\mathrm{t}}\delta-\Lambda_0-\omega^{\mathrm{t}}(\delta-D)\}\\&=\mathrm{Var}(X')+\mathrm{Var}(\Lambda_0)+\omega^{\mathrm{t}}\mathrm{Var}(D)\omega-2\mathrm{Cov}(X',\Lambda_0)+2\mathrm{Cov}(X',D)\omega\\&=v_{\mathrm{np}}+\tau_0^2+\omega^{\mathrm{t}}Q\omega-2\tau_0^2-2\omega^{\mathrm{t}}Q\omega=v_{\mathrm{np}}-v_{\mathrm{wide}}.\end{aligned}$$

We now have to check that Λ_{np} is indeed equal to $\Lambda_{\mathrm{wide}}+X_0$ for $\Lambda_{\mathrm{wide}}=\Lambda_0+\omega^{\mathrm{t}}(\delta-D)$. This follows by re-arranging terms: $\Lambda_{\mathrm{np}}=X'-c_1^{\mathrm{t}}\delta=X_0+c_1^{\mathrm{t}}\delta+\Lambda_0+\omega^{\mathrm{t}}(\delta-D)-c_1^{\mathrm{t}}\delta=X_0+\Lambda_{\mathrm{wide}}$.

The joint convergence follows since all quantities are functions of X', U' and V'. Finally, the covariance $\mathrm{Cov}(\Lambda_S, \Lambda_{\mathrm{np}})$ is found by

$$\begin{aligned}\mathrm{Cov}(\Lambda_S, \Lambda_{\mathrm{np}}) &= \mathrm{Cov}\{\Lambda_0 + \omega^{\mathrm{t}}(\delta - G_S D), \Lambda_0 + \omega^{\mathrm{t}}(\delta - D) + X_0\} \\ &= \mathrm{Cov}(\Lambda_0, \Lambda_0) + \omega^{\mathrm{t}} G_S \mathrm{Cov}(D, D)\omega = \tau_0^2 + \omega^{\mathrm{t}} G_S Q \omega = v_S,\end{aligned}$$

and the proof is complete. ∎

The following lemma presents the limiting behaviour of the FIC scores specified in (S5.2) and (S5.3).

Lemma S1. *Assume that (C0*–C6*) hold. Under the framework of* (S5.1)*, we then have that as* $n \to \infty$

$$\begin{aligned}\mathrm{FIC}_{S,\mathrm{orig}} &\xrightarrow{d} \mathrm{FIC}^{\lim}_{S,\mathrm{orig}} = \max\{0, (\omega^{\mathrm{t}}(I_q - G_S)D)^2 - (v_{\mathrm{wide}} - v_S)\} + v_S \\ \mathrm{FIC}_{\mathrm{wide,orig}} &\xrightarrow{d} \mathrm{FIC}^{\lim}_{\mathrm{wide,orig}} = v_{\mathrm{wide}},\end{aligned}$$

and

$$\begin{aligned}\mathrm{FIC}_{S,\mathrm{new}} &\xrightarrow{d} \mathrm{FIC}^{\lim}_{S,\mathrm{new}} = \max\{0, (\omega^{\mathrm{t}}(I_q - G_S)D - X_0)^2 - (v_{\mathrm{np}} - v_S)\} + v_S \\ \mathrm{FIC}_{\mathrm{wide,new}} &\xrightarrow{d} \mathrm{FIC}^{\lim}_{\mathrm{wide,new}} = \max\{0, X_0^2 - (v_{\mathrm{np}} - v_{\mathrm{wide}})\} + v_{\mathrm{wide}}, \\ \mathrm{FIC}_{\mathrm{np,new}} &\xrightarrow{d} \mathrm{FIC}^{\lim}_{\mathrm{np,new}} = v_{\mathrm{np}},\end{aligned} \tag{S5.6}$$

Proof. Let us first check the limiting behaviour of D_n and $\sqrt{n}\widehat{b}_S$. The former here has by Claeskens and Hjort (2008, Theorem 6.1) precisely the limit D. The limit of the latter may be found by rewriting terms:

$$\begin{aligned}\sqrt{n}\widehat{b}_S &= \sqrt{n}(\widehat{\mu}_S - \widehat{\mu}_{\mathrm{np}}) = \Lambda_{n,S} - \Lambda_{n,\mathrm{np}} \xrightarrow{d} \Lambda_S - \Lambda_{\mathrm{np}} \\ &= \Lambda_0 + \omega^{\mathrm{t}}(\delta - G_S D) - \Lambda_0 - \omega^{\mathrm{t}}(\delta - D) - X_0 = \omega^{\mathrm{t}}(I_q - G_S)D - X_0.\end{aligned}$$

By (C6*) all variance and covariance estimators are consistent since consistency under G_n is equivalent to consistency under the narrow model F_{θ_0}. Note in particular that since all models are correct in the limit, we have $v_{c,S} = v_S$, such that all three estimators $\widehat{v}_S, \widehat{v}_{S,\mathrm{new}}$ and $\widehat{v}_{c,S}$ are consistent for v_S. Inserting the limit variables of the quantities in (S5.2) and (S5.3), noting that $G_{\mathrm{wide}} = I_q$, and gathering common factors, completes the proof through Slutsky's theorem. ∎

Note that since $\delta = 0$ is a valid special case of (S5.1), limit distribution results for the case where the narrow model is indeed fully correct for all n, may be read off directly from the above lemma. As noted before Proposition S1, we have $v_{\mathrm{np}} \geq v_{\mathrm{wide}}$. In cases where the

largest parametric variance is indeed equal to $v_{\rm np}$, it is seen that X_0 becomes degenerate at zero, and the original and new FIC scores for the parametric models coincide exactly in the limit experiment. Hence, for $v_{\rm wide}$ close to $v_{\rm np}$ one would expect the ranking of the parametric models to be similar for the original and new FIC schemes. When the nonparametric variance is much larger than that of the widest parametric model, there is typically no guarantee that the rankings stemming from the two criteria are similar. In particular, the FIC score of the wide model is constant in the limit for the original FIC scheme, while it is random for the new FIC scheme.

Finally, recall that Corollary 1 in the main paper stated that when a parametric model is indeed fully correct, the new FIC would prefer that model over the nonparametric with a probability converging to $\Pr(\chi^2_1 \le 2) \doteq 0.843$. This gave the FIC a new interpretation as an implied focused test of a parametric model, having asymptotic significance level 0.157. These results may be properly generalised under the local misspecification framework considered here. In particular, if the true distribution corresponds to (S5.1), the limiting probability that FIC would select a parametric model S over the nonparametric is

$$\Pr_{G_n}(\text{FIC}_{S,\text{new}} \le \text{FIC}_{\text{np,new}}) \to \Pr\left(\chi^2_1\left(\frac{\{\omega^{\rm t}(I_q - G_S)\delta\}^2}{v_{\rm np} - v_S}\right) \le 2\right),$$

where $\chi^2_1(x)$ is a non-central chi-squared distributed variable with 1 degree of freedom and non-centrality parameter x. Observe that if $\delta = 0$, we are back at the 0.843 probability as before, and when δ departs from zero in the right dimensions, the probability tends to 0. Thus, by daring to trust such a local misspecification framework, one learns not only how the FIC selects model when a parametric model is exactly correct or incorrect, but also how it behaves in between these two 'extremes'.

Remark 1. *Limiting selection with several unbiased parametric models.* Lemma 1 and Corollary 1 of the main paper concern asymptotic selection probabilitites when a) none of the parametric models are correct or at least asymptotically unbiased ($b = 0$), and b) exactly one of them have this property. Although that cover most of the natural cases, it does not yield asymptotic selection probabilities when there are several nested parametric models with a simpler model being (asymptotically) correct. Lemma S1 provides also a framework for studying that situation. The asymptotic selection probabilities for this case depend on the focus, the nesting of the parametric models and whether the truncated or untruncated version of the (new) FIC is used, and must therefore be derived on a case by case basis. For a particular setting with a specified value of δ, the asymptotic selection probabilities may be found by comparing simulated realisations of the $\text{FIC}^{\rm lim}_{\cdot,\text{new}}$ on the right hand side of (S5.6). Since $\delta = 0$ corresponds to the narrow model being true, the procedure may be utilized also

when the true model is fixed and equal to the narrow model for all n.

S5.2. Model averaging and post selection estimators

Finally, we shall use this local misspecification framework to derive the limiting distribution of model averaging estimators based on the new FIC as discussed in Section 6 of the main paper. This is merely an extension of Hjort and Claeskens (2003, Theorem 4.1) which gave the limiting distribution for model averaging estimators related to the original FIC, the AIC and so on. As we shall see, the limit distributions for the new FIC will turn out differently than for the original FIC.

Lemma S2. *Let Ω_M denote all candidate models and*

$$\widehat{\mu}_{\text{final}} = \sum_{j \in \Omega_M} a_j(D_n, X_{n,0}) \widehat{\mu}_j,$$

with $\sum_{j \in \Omega_M} a_j(d,x) = 1$ for all (d,x). Assume that (C0–C6*) hold, and that for all $j \in \Omega_M$, a_j has at most a countable number of discontinuities and that it depends stochastically only on $D_n, X_{n,0}$, and possibly terms converging to constants in distribution under* (S5.1)*. Then*

$$\sqrt{n}(\widehat{\mu}_{\text{final}} - \mu_n) \xrightarrow{d} \sum_{j \in \Omega_M} a_j(D, X_0) \Lambda_j.$$

Proof. There is joint convergence for all $\Lambda_{n,j}, j \in \Omega_M$ and stochastic weights $a_j(D_n, X_{n,0})$. This follows by (S5.5) and the arguments leading to the joint limit in Proposition S1. The continuous mapping theorem then completes the proof. ∎

Note in particular that any function (with at most a countable number of discontinuities) of $\text{FIC}_{j,\text{new}}$, $j \in \Omega_M$ may be applied. This follows since the only stochastic part here is $\sqrt{n}\widehat{b}_S$ which may be re-written as

$$\begin{aligned}\sqrt{n}\widehat{b}_S &= \sqrt{n}(\widehat{\mu}_S - \widehat{\mu}_{\text{np}}) = \sqrt{n}(\widehat{\mu}_S - \widehat{\mu}_{\text{wide}}) - \sqrt{n}(\widehat{\mu}_{\text{wide}} - \widehat{\mu}_{\text{np}}) \\ &= \widehat{\omega}^{\text{t}}(I_q - \widehat{G}_S) D_n + o_{\text{pr}}(1) - X_{n,0}.\end{aligned}$$

Hence, the weight functions suggested in (6.2) in the main paper are treatable by the above model averaging scheme. That is also the case for the post-FIC-selection estimator, putting all weight on the estimator with the smallest (new) FIC score. We note that the limit distributions above are nonlinear mixtures of normals and as such often highly non-normal, sometimes exhibiting multiple modes, etc. For further consequences and generalisations, also for bootstrapping and bagging, see Hjort (2014).

References

Agresti, A. (2002). *Categorical Data Analysis [2nd ed.].* Wiley, New York.

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

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.

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

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

Hjort, N. L. (2014). Discussion of Efron's 'Estimation and accuracy after model selection'. *Journal of the American Statistical Association* **110** 1017–1020.

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

Jullum, M. and Hjort, N. L. (2016). Parametric or nonparametric: The FIC approach. *Submitted: Statistica Sinica*.

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

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