




Error Statistics Using the Akaike and Bayesian Information Criteria

Henrique Cheng¹ · Beckett Sterner² 

Received: 13 March 2024 / Accepted: 26 October 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract

Many biologists, especially in ecology and evolution, analyze their data by estimating fits to a set of candidate models and selecting the best model according to the Akaike Information Criterion (AIC) or the Bayesian Information Criteria (BIC). When the candidate models represent alternative hypotheses, biologists may want to limit the chance of a false positive to a specified level. Existing model selection methodology, however, allows for only indirect control over error rates by setting a threshold for the difference in AIC scores. We present a novel theoretical framework for parametric Neyman-Pearson (NP) model selection using information criteria that does not require a pre-data null and applies to three or more non-nested models simultaneously. We apply the theoretical framework to the Error Control for Information Criteria (ECIC) procedure introduced by Cullan et al. (*J Appl Stat* 47: 2565–2581, 2019), and we show it shares many of the desirable properties of AIC-type methods, including false positive and negative rates that converge to zero asymptotically. We discuss implications for the compatibility of evidentialist and severity-based approach to evidence in philosophy of science.

1 Introduction

Statistical theory and practice are important sources of insights for philosophers investigating the concept of evidence. The development of the Akaike Information Criterion (AIC), named after Hirotugu Akaike, has transformed the theoretical foundations for statistical model selection in science (Burnham & Anderson, 2002; Wagenmakers et al., 2004), building deep connections to information theory and spawning a larger number of related information criteria with different statistical properties (Ding et al., 2018; Markatou et al., 2021). The AIC has been influential

✉ Beckett Sterner
bsterne1@asu.edu

¹ Exponent, Inc., 23445 North 19th Ave, Phoenix, AZ 85027, USA

² School of Life Sciences, Arizona State University, 427 East Tyler Mall, Tempe, AZ 85281, USA

for philosophers analyzing the justification for parsimony (simplicity) as a basis for choosing between alternative theories or models (Forster & Sober, 1994; Bandyopadhyay & Boik, 1999). While model selection using the AIC or other criteria is closely related to a likelihood ratio test, it also differs in important ways that have led philosophers to revise major leading theories of evidence, including evidentialism, severe testing, and Bayesianism (Lele, 2004; Bandyopadhyay & Brittan, 2006; Dennis et al., 2019).

Unlike classical Neyman-Pearson (NP) testing, model selection using the AIC or Bayesian Information Criterion (BIC) does not provide a general procedure for controlling how often one accepts or rejects a candidate model by chance. At the same time, the AIC and BIC apply to a broader array of statistical contexts that are common in scientific practice. In ecology and evolution, for example, scientists increasingly seek to compare many alternative models simultaneously without designating one model as an a priori null (Burnham & Anderson, 2002; Sullivan & Joyce, 2005; Hunt, 2006; Ripplinger & Sullivan, 2008; Aho et al., 2014). This appears to challenge the importance of methodological theories, such as error statistics (Mayo & Spanos, 2006), that view controlling error rates as necessary to acquiring strong statistical evidence. A key gap has been demonstrating how one can generalize the Neyman-Pearson approach to accommodate model selection methods using information criteria while benefiting from their statistical virtues (Dennis et al., 2019).

We address this gap by presenting a new theoretical framework for deriving the error rates of the Error Control for Information Criteria (ECIC) approach (Cullan et al., 2019). The ECIC approach subjects the model with the best (assumed here to be the lowest) observed information criterion score to a test that bounds the false positive rate at or below a pre-specified level. In this way, ECIC combines the flexibility of the AIC and BIC, where multiple unrelated models may be considered, with the methodology of NP testing that explicitly controls false positive rates. We suggest that ECIC can also be interpreted in similar ways as classical NP tests: one can apply it as a decision rule with desirable long-run error frequencies, or as a severe test for a single case (Mayo & Spanos, 2006).

We analyze the performance of an amended numerical algorithm for ECIC in the context of a correctly-specified set of candidate models, and we show the algorithm provides an expected false positive rate less than or equal to a user-specified level of α without the need to designate a null or conduct multiple testing. We also analyze the conditions under which ECIC will succeed at controlling error rates when the model set is misspecified. These are notable advances on the asymptotic rates of classical NP tests, which converge to α in a well-specified setting and may converge to 1 in a misspecified setting. In this manner, we establish a parametric approach to NP classification that parallels recent progress in non-parametric model selection (Tong et al., 2016, 2018).

The theoretical framework we develop also contributes to a deeper methodological understanding of the duality between information criteria and error probabilities. Many studies have now explored the AIC and BIC's relationships to rates of model selection errors in a wide range of model and data types (Dziak et al., 2020). In any application context, the procedure of choosing the model with the lowest AIC or BIC score corresponds to a locally specific probability of getting a false positive,

i.e., choosing a model that does not include the true (or closest-to-true) distribution. However, the chance of a false positive also varies substantially depending on the specific models being compared, the sample size, and the adequacy of the models to the data (Markon & Krueger, 2004; Kuha, 2004; Glattig et al., 2007; Sayyareh et al., 2010; Hegyi & Laczi, 2015; Brewer et al., 2016). While one can always reduce the error rate for a particular analysis by requiring the difference of model scores to exceed a higher threshold value, any fixed choice of threshold (e.g., $\Delta AIC > 2$) will also have varying error rates across contexts. To date, there has been no similarly general procedure for implementing a ΔAIC or ΔBIC threshold that corresponds to the user's desired false positive rate $\leq \alpha$.

Based on this technical advance, we suggest there is greater compatibility than generally appreciated between evidentialist and severity-based theories of evidence in philosophy (Spanos, 2010; Dennis et al., 2019; Bandyopadhyay & Brittan, 2006; Bandyopadhyay et al., 2016b). In particular, one valid interpretation of ECIC is that it implements a severe test using an evidence function in the sense of (Lele, 2004) without having to designate a pre-data null or assume the candidate models are well-specified (i.e., contain the true distribution). This suggests that some key points of contention on both sides of the evidentialist-severity debate are based on contingent features of particular model selection methods rather than essential points of difference between the two conceptions of evidence (Spanos, 2010; Dennis et al., 2019). To put it differently, ECIC undercuts the perception that AIC-type model selection renders severe testing obsolete.

In Sect. 1 we introduce the original ECIC algorithm from (Cullan et al., 2019) and present the amended ECIC algorithm we developed in light of our new theoretical analysis. The amended version avoids some limitations of the original, such as relying on potentially biased maximum likelihood estimates, and is more tractable for theoretical analysis. Section 2 introduces the novel theoretical framework and analyzes the performance of the amended algorithm in correctly versus incorrectly specified settings and where all parameters are known or must be estimated. In particular, we define a theoretical benchmark for performance in the absence of model uncertainty, and we compare the benchmark's error rates to the amended ECIC algorithm and the $\Delta IC > 2$ rule. Lastly, Sect. 3 discusses how our results advance the methodological debate about error statistical versus evidentialist approaches in ecology and evolution. We close by discussing the appropriate use of ECIC as a model selection tool, its known limitations, and potential strategies for addressing them.

2 Controlling False Positives Using Information Criteria

ECIC's primary contribution to model selection methodology is to provide a positive, error-statistical alternative to Dennis et al.'s argument that evidentialism "surpasses" severity as a theory of evidence (Dennis et al., 2019, p. 25). Scientists in fields such as ecology and evolutionary biology commonly use information criteria to select the best-supported model as a basis for drawing further conclusions about their data. Dennis et al. provide an excellent review of model selection approaches for an ecology audience with a special focus on the issue of how one can understand

statistical evidence under model misspecification, and they argue on this basis that an evidentialist approach is superior to classical statistical testing. In particular, they highlight how classical NP hypothesis testing (i.e., likelihood ratio tests between a null and alternate model) are fairly limited tools in practice and may perform very poorly using misspecified models. We follow their definition of misspecification here, which states that a parametric model is misspecified when its distributions are all non-identical to the true distribution under the Kullback–Leibler (KL) divergence, i.e., all have some divergence from the truth. Similarly, the concept of a “quasi-true” distribution in a model can be defined as the distribution with the smallest KL divergence from the true distribution. Dennis et al. refer to this as the distribution “closest” to the truth.

A particularly troublesome reality for NP tests (and Fisherian significance analysis) is that the total error rate – i.e., the rate of false positives (FP) plus false negatives (FN) – does not approach zero asymptotically. In correctly specified cases, the FN rate approaches zero, but since a fixed nominal FP rate α is calculated for all sample sizes, the total error rate approaches α . In misspecified modeling contexts, Dennis et al. demonstrate that the total error rate for classical NP testing can asymptotically approach 1. Conceptually, this alarming result occurs when the null and alternate models are, in terms of KL divergence, far from each other but both fairly close to the true model. The same root concerns also apply to generalized likelihood ratio tests (Vuong, 1989; Pesaran, 1990).

In light of these limitations, (Dennis et al., 2019) propose *evidence functions* as a superior interpretation of information-theoretic model selection methods (Lele, 2004). Conceptually, they describe an evidence function for a given *divergence measure* as “a data-based estimate of the difference of divergences of two approximating models from the underlying process that generated the data” (Dennis et al., 2019), p. 17). A divergence measure is a pseudometric such as the KL divergence that quantifies the discrepancy between two probability distributions. Evidence functions are a generalization of Royall’s likelihoodist concept of evidence, which compares the difference of log-likelihoods for two models to a fixed value k that is independent of sample size (Royall, 1997, 2000).

A list of intuitive conditions that an evidence function should satisfy is provided in (Lele, 2004). These are discussed more thoroughly in Sect. 4 where we demonstrate how ECIC meets these conditions in the applications relevant to this article. We emphasize for now that (Dennis et al., 2019) demonstrate that the difference between two BIC or AIC scores, denoted as ΔBIC_{12} and ΔAIC_{12} , respectively, both qualify as evidence functions in certain contexts. The AIC has several desirable asymptotic behaviors, such as predictive efficiency, and its popularity and ease of use have inspired researchers in many fields to develop a wide assortment of other information criteria that prioritize other statistically desirable properties (Ding et al., 2018). The BIC, for instance, uses a different model complexity penalty in order to ensure asymptotic statistical consistency in a setting that includes overlapping models.

From the evidentialist perspective, ΔBIC qualifies as an evidence function irrespective of the modeling context. On the other hand, ΔAIC does not qualify when the models being compared are nested or overlapping. The basic reasoning can be

stated using the general form of an information criterion $-2 \log(\hat{L}_i) - c_n r_i$, where \hat{L}_i is a model with parameters estimated using maximum likelihood, r_i is the dimension of the model's parameters, and c_n scales the penalty for model complexity. Differences in information criteria qualify as evidence functions irrespective of the modeling context when the rate of growth for c_n is $< n$ and $> \log(\log(n))$ (Nishii, 1988). Since we are concerned with whether one can control error rates using commonly applied criteria, the evidence function framework provides us a good setting to discuss the contributions of ECIC to model selection.

Our main technical result will be to show that ECIC allows one to construct NP tests using a transformation of the AIC or BIC that have false positive rates at or below a pre-specified level and preserve key virtues of evidence functions. Cullan et al. developed ECIC with the aim of maintaining the basic logical structure of hypothesis testing while avoiding some of the limitations of having to designate fixed null or alternative models in advance of the data or of having to conduct multiple tests when there are more than two candidate models (Cullan et al., 2019). Informally, ECIC can be understood as a form of composite testing. Post data, it treats the best scoring model as the alternative model and compares it against the other, worse-scoring candidate models as “null” models. Instead of conducting pair-wise tests, however, ECIC uses a minimax approach (Berger, 1985) to identify the most conservative decision threshold among the null models and uses it to determine whether the best scoring alternative model should be accepted.

More formally, Cullan et al. introduced ECIC using the AIC and BIC in a correctly specified setting, i.e., assuming a model set $\mathcal{M} = \{M_1, \dots, M_k\}$ where the model containing the true distribution $M_t(\theta_t)$ for $t \in \{1, \dots, k\}$ is contained in \mathcal{M} (Cullan et al., 2019). We understand a statistical model here to be a set of probability distributions indexed by one or more parameters. We use θ_{t^*} to index the unique quasi-true distribution as the distribution in \mathcal{M} closest to the true one based on KL divergence. Note that Cullan et al. did not formally consider the misspecified setting where $M_t \notin \mathcal{M}$.

Given data $D \in \mathcal{D}$, their general framework for model selection involves specifying criterion, preference, and decision functions. The *criterion function* $f : \mathcal{M} \times \mathcal{D} \rightarrow \mathbb{R}^k$ assigns a score to all $M_i \in \mathcal{M}$, $i = 1, \dots, k$. We slightly alter the range of their *preference function* $g : \mathbb{R}^k \rightarrow 1, \dots, k$ to identify the *index* b of the best scoring model instead of the model itself. The *decision function* $h : \mathcal{M} \times \mathbb{R}^k \rightarrow \{0, 1\}$ returns 1 if M_b is selected and 0 otherwise. Letting F denote the vector $(f(M_i, D) : i = 1, \dots, k)$, this yields four possible outcomes:

1. True Positive (TP): $b = t$ and $h(M_b, F) = 1$
2. False Positive (FP): $b \neq t$ and $h(M_b, F) = 1$
3. False Negative (FN): $b = t$ and $h(M_b, F) = 0$
4. True Negative (TN): $b \neq t$ and $h(M_b, F) = 0$

The crux of ECIC lies in defining the decision function h such that the FP rate is controlled at level α over many random draws from \mathcal{D} . In effect, one can think of this as calculating the locally correct ΔAIC or ΔBIC threshold to give the desired

false positive rate. However, this threshold will depend on which of the candidate models contain the true distribution. ECIC provides a way to do this without making pre-data assumptions that treat some models asymmetrically, i.e., without choosing a fixed null or comparing model pairs in a predetermined order.

The first step is to define a *difference-in-goodness-of-fit* (DGOF) statistic Δf_j that measures the difference between the best and second best score based on simulating data from model M_j (Wagenmakers et al., 2004).

$$\Delta f_j(D, \mathcal{M}) = f(M_b, D) - \min\{f(M_r, D) : r \neq b\}.$$

Following the classical NP approach, we are interested in controlling the potential for false positives, which we can represent jointly by the probability that a false model scores best, $P(b \neq t)$, and the probability the false model scores a certain amount better than the next best model, which is described by the distribution of negative values for Δf given $b \neq t$. ECIC estimates and considers both of these inputs in order to determine its decision function.

Letting Δf_{obs} denote the observed DGOF score and $q(\tau)$ the τ^{th} quantile of the distribution of Δf , the class of decision function h that ECIC uses is defined as

$$h_{q(\tau)} = \begin{cases} 1 & \text{if } \Delta f_{\text{obs}} < q(\tau) \\ 0 & \text{else.} \end{cases} \quad (1)$$

The goal is therefore to select τ so that the FP rate is equal to or less than α , recognizing that $q(\tau)$ must be estimated since the true distribution is unknown. To achieve this goal, we provide a numerical algorithm that uses parametric sampling from the “null” models M_j to estimate $P_j(b \neq j)$ and Δf_j . The algorithm uses the largest value of $P_j(b \neq j)$ to set a conservative value for $q(\tau)$ so that the FP rate will be less than or equal to α in the case that any $M_{j \neq b}$ ends up being the correct one.

To describe the numerical algorithm we present here, we make some changes in notation and steps in the original ECIC implementation summarized in Algorithm 1 to help facilitate comparison. Three points are worth highlighting about the original algorithm. First, steps 5 – 8 are achieved by simulating draws from $M_{j \neq b}$ and subsequently applying the criterion function to each draw. We use b here to denote the best observed model and b_{jl}^* to denote the best model on simulated data D_{jl} . Second, step 11 is the key to bounding the FP rate since it sets the lowest, i.e., strictest available, decision threshold for step 12. Third, the distribution estimated in line 8 of Algorithm 1 only includes negative values since it is conditioned on M_b being the observed best. If Δf_j is empty, then we can remove M_j from consideration as it has no probability of generating data that scores a different model best. As illustrated in Algorithm 1, if any $M_i \in \mathcal{M}$ has unknown parameters then maximum likelihood estimation is used and the models with estimated parameters are denoted as $M_i(\hat{\theta})$ or just \hat{M}_i for short.

Algorithm 1 Original ECIC Algorithm

-
- 1: Compute model scores $F_i = f(M_i, D_{obs})$
 - 2: Apply preference function g to compute best observed model index $b = g(F)$
 - 3: Compute observed DGOF $\Delta f_{obs} = F_b - \min\{F_i : i \neq b\}$
 - 4: **for all** $j \neq b$, assume M_j is true and **do**
 - 5: Simulate a set of $l = 1 \dots n$ datasets, D_{jl} , from \hat{M}_j
 - 6: Compute the best models $b_{jl}^* = g(f(D_{jl}))$ for $l = 1 \dots n$
 - 7: Estimate $\pi_j = P(b_{jl}^* = b | \hat{M}_j)$ using $\hat{\pi}_j = \frac{1}{n} \sum_l \mathbb{1}_{b_{jl}^* = b}$
 - 8: Estimate the distribution of $\Delta f_j = f(M_b) - \min\{f(M_r) : r \neq b\}$ using D_{jl}
 - 9: Compute the quantile $\hat{q}_j(\hat{\tau}_j)$ of Δf_j using $\hat{\tau}_j = \min\{\alpha / \hat{\pi}_j, 1\}$
 - 10: **end for**
 - 11: Set $\hat{q} = \min_{j \neq b} \{\hat{q}_j(\hat{\tau}_j)\}$ to estimate $q(\tau)$
 - 12: Apply $h = \begin{cases} 1 & \text{if } \Delta f_{obs} < \hat{q} \\ 0 & \text{else} \end{cases}$
-

As we will show in our results below, several changes are required in the original algorithm in order to achieve the desired α level in a general way. To begin, the formula for estimating $q(\tau)$ in line 9 of Algorithm 1 actually controls the FP rate at level $\alpha * (k - 1)$, where $k = |\mathcal{M}|$. An ostensible correction would be to replace the original $\hat{\tau}_j$ with $\min\{\frac{\alpha}{(k-1)\hat{\pi}_j}, 1\}$. However, even with this change, $\hat{\tau}_j$ does not necessarily suffice to control the FP rate at the nominal level α when parameter estimation is required. The root problem is that in lines 7 – 8, the original ECIC algorithm conditions its estimate of π_j and Δf_j on model b being observed best from the empirical data in line 2: the parameter estimates for \hat{M}_j conditioned on model b having scored best are not equal to the unconditioned expectations, and hence will typically be biased compared to the standard MLEs (Cullan et al., 2019).

In response, we establish and analyze Algorithm 2 in the remainder of the paper. Critically, note that the amended algorithm does not reference the best observed model, M_b , in its core steps for estimating the decision quantile $q(\tau)$. Instead, it estimates the chance that a model other than M_j scores best when the distribution \hat{M}_j is assumed true. We then calculate Δf_j based on the score difference between the best false model (i.e., M_r where $r = \min\{f(M_s) : s \neq j\}$) and any other model on the simulated data from \hat{M}_j . Figure 1 illustrates how Algorithm 2 works with three candidate models in the correctly specified case where one model includes the true distribution.

Algorithm 2 Amended ECIC Algorithm

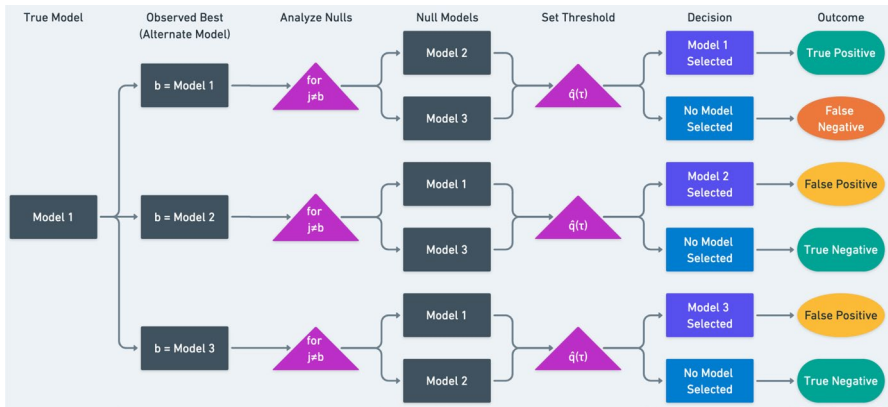


Fig. 1 Flowchart illustrating the basic logic of Algorithm 2 in the correctly-specified case with three candidate models. From left to right, the first box indicates that Model 1 is true and generates the observed data, although this is not known in the remainder of the procedure. The next stage shows that any one of the three candidate models may score best (Algorithm 2 steps 1-3). The first column of pink triangles indicates the for loop in the algorithm (steps 4-9) where the remaining candidate models are analyzed to determine π_j and Δf_j , resulting in the decision threshold $\hat{q}(\tau)$ shown in the second column of pink triangles (steps 10-13). The blue boxes mark the decision made to accept or not accept the best observed model (step 14), and the corresponding outcome in the colored ovals

-
- 1: Compute vector of model scores $F_i = f(M_i, D_{obs})$
 - 2: Apply preference function g to compute best observed model index $b = g(F)$
 - 3: Compute observed DGOF $\Delta f_{obs} = F_b - \min\{F_j : j \neq b\}$
 - 4: **for all** $j \neq b$, assume M_j is true and **do**
 - 5: Simulate a set of $l = 1 \dots n$ datasets, D_{jl} , from \hat{M}_j
 - 6: Compute the best models $b_{jl}^* = g(f(D_{jl}))$ for $l = 1 \dots n$
 - 7: Estimate $\pi_j = P(b_{jl}^* \neq j | \hat{M}_j)$ using $\hat{\pi}_j = \frac{1}{n} \sum_l \mathbb{1}_{b_{jl}^* \neq j}$
 - 8: Estimate the distribution of $\Delta f_j = f(M_v) - \min\{f(M_r) : r \neq v\}$ given $v \neq j$
 - 9: **end for**
 - 10: Compute $\hat{\pi}_{max} = \max_{j \neq b}(\hat{\pi}_j)$
 - 11: Compute $\tau_{min} = \min\{\frac{\alpha}{\hat{\pi}_{max}}, 1\}$
 - 12: Compute the quantiles $\hat{q}_j(\tau_{min})$ of Δf_j for $j \neq b$
 - 13: Estimate $q(\tau)$ using $\hat{q}(\tau) = \min_j\{\hat{q}_j(\tau_{min})\}$
 - 14: Apply $h = \begin{cases} 1 & \text{if } \Delta f_{obs} < \hat{q} \\ 0 & \text{else} \end{cases}$
-

3 Error Control Properties of ECIC

We now turn to address both the finite sample and asymptotic behaviors of the FP and FN rates using ECIC. We will highlight how ECIC can work under well-specified and misspecified model sets, i.e., when the true distribution is and is not included within the candidate models, respectively. In particular, we consider four model selection contexts: the correctly specified setting where all model parameters are known (CK), the correctly-specified setting where all model parameters are unknown (CU), the misspecified setting where all model parameters are known (MK), and the misspecified setting where all parameters are unknown (MU). We begin by defining a theoretical performance benchmark that assumes knowledge of the true distribution, which we use to compare ECIC against in practice where model uncertainty may be substantial. From there, we conduct simulations to highlight interesting elements of the benchmark versus in-practice performance across varying sample sizes. We ultimately show that although finite sample behavior may differ depending on the specific context for \mathcal{M} , the FP and FN rates both asymptotically approach zero in all cases. This contrasts with classical NP tests where the total error rate converges to α at best, and may be higher if the model set is misspecified.

3.1 Defining a Theoretical Benchmark

As mentioned, the formula in (1) must be estimated since the true distribution $M_t(\theta_t)$ is unknown. In order to provide a comparative baseline for the quality of our estimation in practice, we introduce a decision function that assumes knowledge of $M_t(\theta_t)$ to serve as a theoretical benchmark. The perfect decision function

$$h^{\text{perfect}} = \begin{cases} 1 & \text{if } b = t \\ 0 & \text{else} \end{cases}$$

is not very informative for our purposes since it does not follow the basic logic of ECIC where models from \mathcal{M} are used to determine a quantile for the decision threshold. We instead use

$$h^B = \begin{cases} 1 & \text{if } \Delta f_{\text{obs}} < \hat{q}(\tau^B(t)) \\ 0 & \text{else} \end{cases} \tag{2}$$

where

$$\tau^B(t) = \begin{cases} \min \left\{ \frac{\alpha}{P_r(b \neq t | \hat{M}_t)}, 1 \right\} & \text{if } b \neq t \\ \min \left\{ \frac{\alpha}{\min_{r \neq t} (P_r(b = t | \hat{M}_r))}, 1 \right\} & b = t \end{cases} \tag{3}$$

The role of $\min_{r \neq t} (P_r(b = t | \hat{M}_r))$ in the $b = t$ condition of (3) is to ensure that when the best observed model is the true one, we choose the largest DGOF quantile, and hence the most permissive threshold, for deciding on M_b . This is desirable if $b = t$ since selecting M_b would evade a false negative. When the distributions are known,

we drop the need for MLEs to find \hat{M}_t and \hat{M}_i and use just M_t and M_i instead to designate the specific distributions. We use the notation $P_j(X)$ in this case to mean the probability of X given M_j is the assumed true distribution. We emphasize that (3) does not necessarily globally minimize error rates but rather serves to illustrate the performance possible for the ECIC approach when we remove uncertainty about the model, M_t , that contains the true distribution.

We now explore the error control properties of the benchmark decision function (2) in each modeling context. Our approach is to write down equations for the FP and FN rates by adding up the chances of h^B selecting the observed best model M_b across the possible observed best models and alternate models used to set the decision threshold (see Fig. 1). Beginning with the Benchmark CK (BCK) setting, the FP rate for a fixed sample size n is:

$$\begin{aligned}
 P^{\text{BCK}}(\text{FP}) &= P_t(\cup_j b = j \neq t \& \Delta f_{\text{obs}} < q(\tau^B(t))) \\
 &= \sum_{j \neq t} P_t(b = j \& \Delta f_{\text{obs}} < q(\tau^B(t))) \\
 &= \sum_{j \neq t} P_t(b = j) P_t(\Delta f_{\text{obs}} < q(\tau^B(t)) | b = j) \\
 &= \sum_{j \neq t} P_t(b = j) \min \left\{ \frac{\alpha}{P_t(b \neq t)}, 1 \right\} \\
 &\leq \sum_{j \neq t} P_t(b = j) \frac{\alpha}{P_t(b \neq t)} \\
 &= \frac{\alpha P_t(b \neq t)}{P_t(b \neq t)} \\
 &= \alpha
 \end{aligned}
 \tag{4}$$

Note that the *min* term enters in 4 by substituting for the definition of $\tau^B(t)$ when $b \neq t$. The correction factor $P_t(b \neq t)$ serves to rescale α to account for the total probability of a getting a false positive by picking any model other than M_t . The intuition is that as the chance of false positives goes to zero, we can set a proportionately more permissive quantile than α and still get a FP rate less than or equal to α . When $P_t(b \neq t) < \alpha$, we no longer need a decision threshold at all. This is key to ensuring that both FP and FN rates go to zero asymptotically, in contrast to classical NP testing. Asymptotically, using a statistically consistent information criterion ensures that $\lim_{n \rightarrow \infty} P_t(b \neq t) \rightarrow 0$, so that the minimum in (4) eventually becomes 1 and $\lim_{n \rightarrow \infty} P^{\text{BCK}}(\text{FP}) \rightarrow 0$.

Next, let us look at the asymptotic behavior of the FN rate:

$$\begin{aligned}
 P^{\text{BCK}}(\text{FN}) &= P_t(b = t \& \Delta f_{\text{obs}} \geq q(\tau^B(t))) \\
 &= P_t(b = t) P_t(\Delta f_{\text{obs}} \geq q(\tau^B(t)) | b = t) \\
 &= P_t(b = t) \left(1 - \min \left\{ \frac{\alpha}{\min_{r \neq t} (P_r(b = t))}, 1 \right\} \right)
 \end{aligned}
 \tag{5}$$

Here, the other case of $\tau^B(t)$ enters in (5) because $b = t$. It is difficult to find a neat analytical solution because the FN rate may depend on which r sets the *min* for a given n . Nonetheless, the asymptotic behavior is straightforward. In this case, $\lim_{n \rightarrow \infty} \min_{r \neq t}(P_r(b = t)) \rightarrow 0$, so the value of the minimum in (5) always eventually reaches 1, which implies $\lim_{n \rightarrow \infty} P^{\text{BCK}}(\text{FN}) = 0$.

In the BCU setting, the main difference is that the models are no longer single distributions and we use the maximum likelihood method to estimate parameters, including for M_t . If we update (3) to reflect this, we get an expression with \hat{M}_t and \hat{M}_r that now behaves stochastically. We therefore aim to demonstrate $E[\text{FP}] \lesssim \alpha$, i.e., that the expectation is less than or approximately equal to the nominal level. We start with the FP rate as before and arrive at α times a ratio of probabilities with the MLE in the denominator:

$$\begin{aligned}
 P^{\text{BCU}}(\text{FP}) &= \sum_{j \neq t} P_t(b = j) P_t(\Delta f_{\text{obs}} < q(\tau^B(t)) | b = j) \\
 &= \sum_{j \neq t} P_t(b = j) \min \left\{ \frac{\alpha}{P_t(b \neq t | \hat{M}_t)}, 1 \right\} \\
 &\leq \alpha \sum_{j \neq t} \frac{P_t(b = j)}{P_t(b \neq t | \hat{M}_t)} \\
 &= \alpha \frac{P_t(b \neq t)}{P_t(b \neq t | \hat{M}_t)}
 \end{aligned} \tag{6}$$

We can estimate the denominator, $P_t(b \neq t | \hat{M}_t)$, by sampling parametrically from \hat{M}_t , counting the times the condition is met, and dividing by the number of samples. This estimator is a function of a Binomial distribution: $\hat{\pi}_t \sim 1/n * B(n, p_{\hat{M}_t})$. If $E[B(n, p_{\hat{M}_t})] = np_{\hat{M}_t}$, so that the MLEs provide an unbiased estimate of the true conditional probability, then $E[\hat{\pi}_t] = P_t(b \neq t)$.

To calculate the expected FP rate expressed in (6), we then need to handle the fact that $\hat{\pi}_t$ appears in the denominator. We can do that by looking at the properties of $f(X) = n/X$ where $X \sim B(n, p_{\hat{M}_t})$. In particular, we can use a Taylor series expansion based on the facts that

$$f(X) = n/X, f'(X) = -n/X^2, \text{ and } f''(X) = 2n/X^3,$$

together with the first two non-zero terms of the Taylor series:

$$\begin{aligned}
 E[f(X)] &\approx f(E[X]) + f''(E[X])/2 * Var(X) \\
 &= f(np_{M_t}) + f''(np_{M_t})/2 * Var(\hat{\pi}_t) \\
 &= \frac{n}{np_{M_t}} + \frac{2n}{2 * (np_{M_t})^3} * (np_{M_t}(1 - p_{M_t})) \\
 &= \frac{1}{p_{M_t}} + \frac{(1 - p_{M_t})}{np_{M_t}^2} \\
 &\approx \frac{1}{p_{M_t}}
 \end{aligned}$$

The last approximation follows because we can make n arbitrarily large by drawing more samples from \hat{M}_j when estimating π_j . Now we can finish by showing

$$\begin{aligned}
 E[P^{BCU}(\text{FP})] &\leq E \left[\alpha \frac{P_t(b \neq t)}{P_t(b \neq t | \hat{M}_t)} \right] \\
 &= \alpha P_t(b \neq t) E \left[\frac{1}{\hat{\pi}_{\hat{M}_t}} \right] \\
 &= \alpha P_t(b \neq t) E \left[\frac{n}{X} \right] \\
 &\approx \alpha P_t(b \neq t) \left(\frac{1}{p_{M_t}} \right) \\
 &= \alpha
 \end{aligned}$$

More weakly, for the asymptotic behavior we can assume just that the MLEs are asymptotically consistent. Then we have that $\lim_{n \rightarrow \infty} \hat{M}_t \rightarrow M_t$ for any model in \mathcal{M} . Thus, by the same reasoning used in the BCK setting, $\lim_{n \rightarrow \infty} P^{BCU}(\text{FP}) \rightarrow 0$. Similarly, the FN rate:

$$P^{BCU}(\text{FN}) = P_t(b = t) \left(1 - \min \left\{ \frac{\alpha}{\min_{r \neq t} (P_r(b = t | \hat{M}_r))}, 1 \right\} \right)$$

likewise approaches zero asymptotically.

We now turn to the two misspecified settings where the true distribution is not contained in \mathcal{M} . We denote M_{t^*} as the model closest to the true distribution in terms of KL divergence and use it in lieu of M_t in (3). A true positive in this context would therefore be selecting M_b when $b = t^*$. As before, we consider the cases where the parameters are known versus unknown. In the Benchmark MK setting we thus have:

$$\begin{aligned}
 P^{\text{BMK}}(\text{FP}) &= \sum_{j \neq t^*} P_t(b = j) P_t(\Delta f_{\text{obs}} < q(\tau^{\text{B}}(t^*)) | b = j) \\
 &= \sum_{j \neq t^*} P_t(b = j) \min \left\{ \frac{\alpha}{P_{t^*}(b \neq t^*)}, 1 \right\} \\
 &\leq \sum_{j \neq t^*} P_t(b = j) \frac{\alpha}{P_{t^*}(b \neq t^*)} \\
 &= \alpha \frac{P_t(b \neq t^*)}{P_{t^*}(b \neq t^*)}
 \end{aligned}$$

In general, it is possible that $P_t(b \neq t^*) > P_{t^*}(b \neq t^*)$, since the closest distribution M_{t^*} in the candidate models may be a poor proxy for the true distribution M_t . The benchmark decision function therefore does not guarantee a ceiling on FP at a finite sample size.

For the false negative rate, we have:

$$P^{\text{BMK}}(\text{FN}) = P_t(b = t^*) \left(1 - \min \left\{ \frac{\alpha}{\min_{r \neq t^*} (P_r(b = t^*))}, 1 \right\} \right)$$

The results in the BMU setting are similar after accounting for conditioning on \hat{M}_{t^*} or \hat{M}_r . Both the finite-sample and asymptotic properties for the error rates in the Benchmark MK and MU settings are similar to those in the CU setting — that is, the finite-sample FP error rates are determined by the discrepancy between probabilities under the true and closest models, but all asymptotic error rates still approach zero as long as $\lim_{n \rightarrow \infty} P^{\text{BCU}}(\text{FP}) \rightarrow 0$.

3.2 Practical Performance for Correctly Specified Case

Having set out a reasonable theoretical baseline, we can move to assess ECIC’s practical performance in the correctly specified setting with both known (CK) and unknown (CU) parameters. We present results from a simulation study involving normal distributions for the former and spline regression fits for the latter.

3.2.1 Known Parameters

The FP rate for a fixed sample size n in the CK setting is:

$$\begin{aligned}
 P^{CK}(\text{FP}) &= P_t(\cup_j b = j \neq t \& \Delta f_{\text{obs}} < \hat{q}) \\
 &= \sum_{j \neq t} P_t(b = j) \min \left\{ \frac{\alpha}{\max_{r \neq j} (P_r(b \neq r))}, 1 \right\} \\
 &\leq \sum_{j \neq t} P_t(b = j) \frac{\alpha}{\max_{r \neq j} (P_r(b \neq r))} \\
 &\leq \sum_{j \neq t} P_t(b = j) \frac{\alpha}{P_t(b \neq t)} \\
 &= \alpha \frac{P_t(b \neq t)}{P_t(b \neq t)} \\
 &= \alpha
 \end{aligned}$$

Since for all r , $\lim_{n \rightarrow \infty} P_r(b \neq r) \rightarrow 0$, it is easy to see that $\lim_{n \rightarrow \infty} P^{CK}(\text{FP}) \rightarrow 0$ by using the same reasoning in the Benchmark CK setting. The FN rate can be expressed as:

$$\begin{aligned}
 P^{CK}(\text{FN}) &= P_t(b = t \& \Delta f_{\text{obs}} \geq \hat{q}) \\
 &= P_t(b = t) P_t(\Delta f_{\text{obs}} \geq \hat{q} | b = t) \\
 &= P_t(b = t) \left(1 - \min \left\{ \frac{\alpha}{\max_{r \neq t} (P_r(b \neq r))}, 1 \right\} \right).
 \end{aligned}$$

As above, this also approaches zero asymptotically.

Figure 2 presents a simulation study for Gaussian distributions with known parameters across sample sizes that are multiples of 50 between 50 and 3, 500 inclusive. Here, $\mathcal{M} = \{N(3.9, 1), N(4, 1), N(4.2, 1)\}$, $M_t = N(4, 1)$, and the information criterion used is the negative log likelihood since all models represent single distributions. We set $\alpha = 0.05$, sample 7, 000 draws from the true model for each sample size to generate observed data, and simulate 7, 000 data sets for performing ECIC. The top three graphs in Fig. 2 are the FP rates using the Benchmark CK threshold, ECIC, and Burnham and Anderson’s ΔIC rule of thumb from (Burnham & Anderson, 2002), respectively. This latter method selects the model with the lowest observed IC score if Δf_{obs} is less than a fixed value set by the analyst, which we set to the commonly used value of -2 . The bottom three graphs correspond to the FN rates under the same model selection approaches used in the top graphs.

Overall, we see that both the benchmark and practical implementations of ECIC achieve control over the FP rate at the desired level, with small exceptions due to our numerical procedure that we discuss in more detail below. For the $\Delta IC < -2$ approach, we see a value of ≈ 0.02 for the FP rate at a sample size of 50, followed by a subsequent spike and drop back down. This is due to the observed DGOFs being relatively close to zero at a sample size of 50. In fact, 6,880 of the 7,000 samples here have an observed DGOF greater than -2 , which leads to a high rejection rate in general. Since each model is observed as best at about similar frequencies, this leads to a lower FP rate and higher FN rate. We therefore see that in this model selection context, the $\Delta IC < -2$ rule corresponds to a slightly more conservative α level than 0.05.

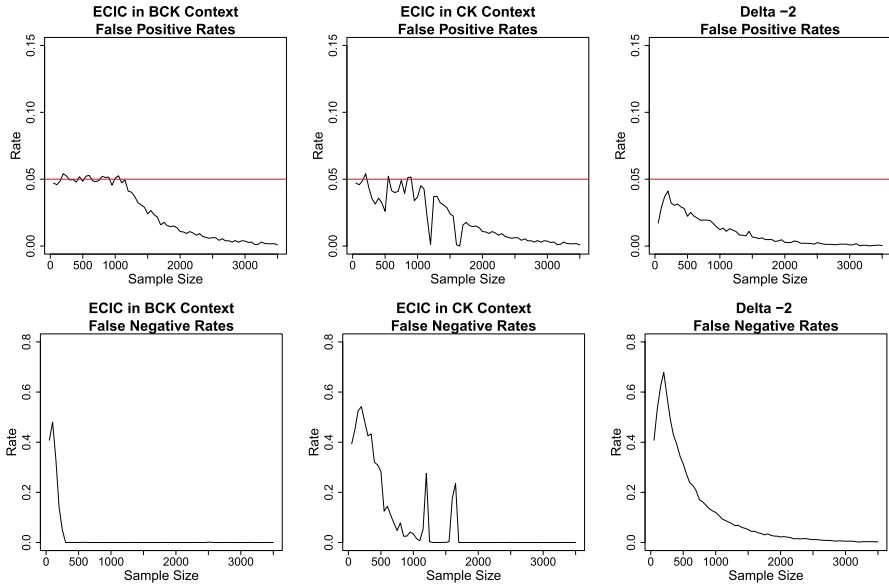


Fig. 2 Simulation study where the top three graphs correspond to linearly interpolated FP rates using thresholds for ECIC in the BCK context, ECIC in the CK context, and Burnham and Anderson’s rule of thumb with a threshold of -2 where the red lines mark the nominal α level. The bottom graphs correspond to linearly interpolated FN rates under the same methods

As we would expect, then, the practical version of ECIC set to $\alpha \leq 0.05$ shows a slightly improved FN rate compared to the Δ/C rule, modulo two peaks due to a discretization effect. The ECIC benchmark illustrates the ideal behavior we expect: the FP rate is held at α until the sample size is large enough that the total chance of a FP is below α , at which point the decision threshold becomes progressively weaker until the best-scoring model is always accepted. For the FN rates, we can see, as expected, a sharper spike and heavier right-skewed tail in the CK versus the BCK setting. Additionally, both the practical and benchmark procedures show the FN rate declining to zero. A key takeaway here is that although each approach successfully controls the FP rate in this instance, there is no theoretical guarantee of this happening in the general case with the rule of thumb approach.

We identify two artificial effects in ECIC’s performance due to using a finite-sample DGOF distribution. First, the small upticks above 0.05 for smaller sample sizes arise from numerical error due to discretization. That is, these upticks flatten out as a greater number of draws from each model are made.

Second, the sudden upticks of FN rates in the CK graph arises from an artifact of discretization for the simulated data sets. For example, the first uptick in the CK graph that occurs a little bit after the sample size of 1,000 is due to there being only one negative DGOF value, ≈ -4.38 , in the estimated distribution used to set the decision quantile. Whenever the true model, $N(4, 1)$, is observed best, there are 1,926 observed DGOF values that are greater than -4.38 , which leads to the estimated FN rate of $1,926/7,000 \approx 0.28$. If the number of simulated data sets were

increased, then values between -4.38 and zero would be drawn and a less stringent decision threshold might be set.

3.2.2 Unknown Parameters

As with the BCU setting, the dependence on ML estimation for the model parameters in the CU setting frustrates the establishment of a deterministic analytical bound for $P^{CU}(FP)$. However, we are still able to demonstrate that $E[FP] \lesssim \alpha$ since the FP rate is still bounded by $\alpha \frac{P(b \neq t | \hat{M}_t)}{P(b \neq t | \hat{M}_t)}$:

$$\begin{aligned}
 P^{CU}(FP) &= \sum_{j \neq t} P_t(b = j) P_t(\Delta f_{\text{obs}} < \hat{q} | M_t, b = j) \\
 &= \sum_{j \neq t} P_t(b = j) \min \left\{ \frac{\alpha}{\max_{r \neq j} (P_r(b \neq r | \hat{M}_r))}, 1 \right\} \\
 &\leq \sum_{j \neq t} P_t(b = j) \frac{\alpha}{\max_{r \neq j} (P_r(b \neq r | \hat{M}_r))} \\
 &\leq \alpha \sum_{j \neq t} \frac{P_t(b = j)}{P_t(b \neq t | \hat{M}_t)} \\
 &= \alpha * \frac{P_t(b \neq t)}{P_t(b \neq t | \hat{M}_t)}
 \end{aligned} \tag{7}$$

As for the FN rate,

$$\begin{aligned}
 P^{CU}(FN) &= P_t(b = t) P_t(\Delta f_{\text{obs}} \geq \hat{q} | b = t) \\
 &= P_t(b = t) \left(1 - \min \left\{ \frac{\alpha}{\max_{r \neq t} (P_r(b \neq r | \hat{M}_r))}, 1 \right\} \right)
 \end{aligned}$$

Both of these errors rates approach zero since $P(b \neq r | \hat{M}_r)$ will decay to zero for all r .

For our CU simulations, we let \mathcal{M} consist of 3 cubic spline regression models using B-spline bases with knots placed at the quintiles, sextiles, and septiles of the interval $[-10, 10]$. The true model places knots at the quintiles of $[-10, 10]$ and has a coefficient vector of $\beta = (0.3, -0.6, 0.4, -0.6, 0.5, -0.5, 0.1)$. Thus, in this simulation the knot locations are fixed for each model but the parameters are unknown. We included sample sizes that are multiples of 10 between 20 and 260 inclusive and distributed observed points uniformly beginning at -10 and ending at 10 . We set $\alpha = 0.05$, sampled 2,000 draws from the true model with added independent and identically distributed $N(0, 0.06^2)$ noise for each sample size to generate observed data, and simulated 3,000 data sets for performing ECIC. The IC used is the BIC since it will consistently select the true model as sample size increases under conditions which are met in the current setup (Vrieze, 2012). A lower number of sample draws and simulations are used here compared to the CK exercise to make the simulations faster.

To give a sense of what these spline fits might look like, in Fig. 3 we fit all three models to the first of the 2,000 simulated draws of 20 points. It should be clear in this figure that visual inspection alone may be insufficient to select an appropriate model, showing the utility of a model selection procedure. Figure 4 presents the error rates resulting from our simulations. We see that there is slight uncontrol of the FP rate for some smaller sample sizes under the benchmark threshold, whereas the in-practice threshold maintains control for all sample sizes. This former observation is perhaps unsurprising given that the bound derived in the BCU calculation is approximate, and its error only diminishes asymptotically as the number of samples used in simulating the DGOF distributions increases.

The rule of thumb approach can result in an FP rate that significantly exceeds the nominal rate, as it does for $n = 20$ in which it is a little over double that of 0.05. Although the rule of thumb approach catches up to the performance of ECIC for moderately larger sample sizes, the high degree of instability for smaller sample sizes highlights the benefit of the theoretical framework of ECIC. For the FN rates, the benchmark threshold and rule of thumb approach perform similarly with rates beginning around 0.20 that taper off to rates closer to zero around sample sizes of 100. On the other hand, it is only until sample sizes of around 250 that the FN rates using the in-practice threshold begin to taper off to around zero. This exemplifies the conservative nature of ECIC, in that strong evidence must be present in order for it to select a model, as well as the potential cost of demanding control over only the FP rates.

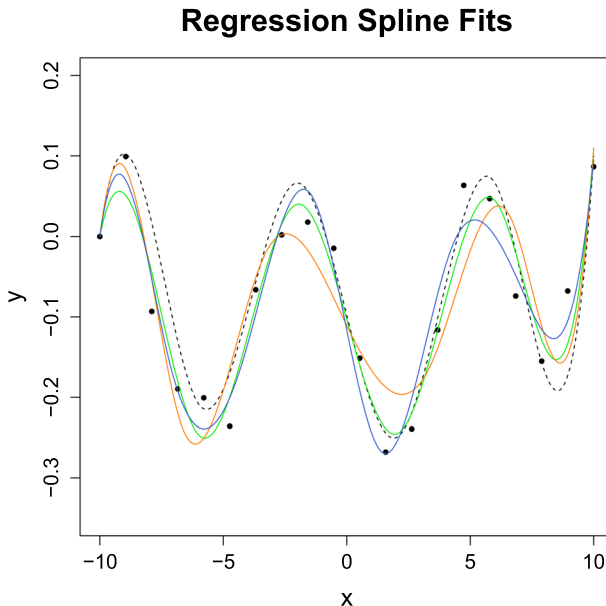


Fig. 3 A fit of all three spline regression models in the BCU/CU context to a simulated draw of 20 points. The dashed black line represents the (true) model with knots at the quintiles of $[-10, 10]$. The green, orange, and blue lines represents a cubic spline regression fit to the data using knots at the quintiles, sextiles, and septiles of $[-10, 10]$, respectively

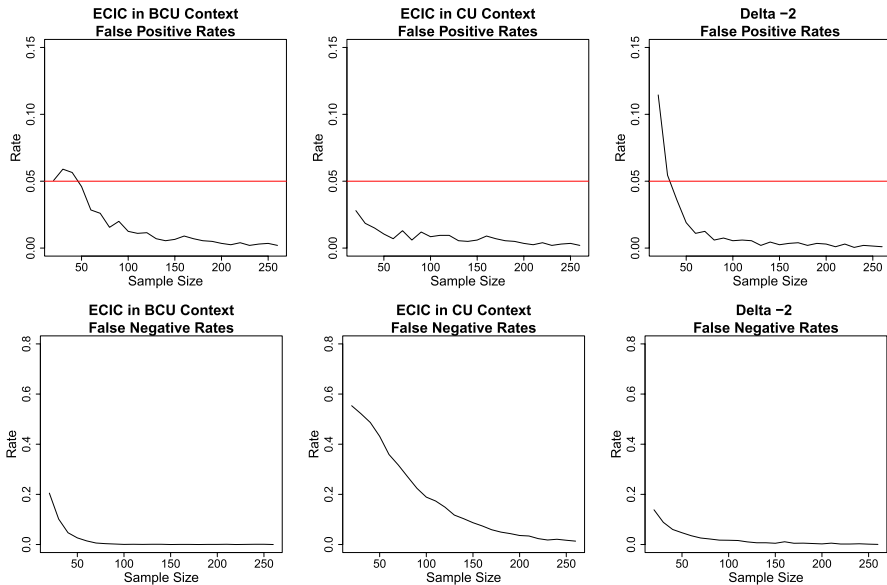


Fig. 4 Simulation study where the top three graphs correspond to linearly interpolated FP rates using thresholds for ECIC in the BCU context, ECIC in the CU context, and Burnham and Anderson's rule of thumb with a threshold of -2 where the red lines mark the nominal α level. The bottom graphs correspond to linearly interpolated FN rates under the same methods

3.3 Misspecified Model Sets In Practice

Following (Dennis et al., 2019), we extend our analysis of ECIC to consider model selection contexts where the candidate model set \mathcal{M} does not include the true distribution. In both the known and unknown parameter settings, we find that ECIC controls the FP rate for finite samples under restricted circumstances and hence does not guarantee α universally. However, what can still be guaranteed is the asymptotic decay of both the FP and FN rates to zero.

3.3.1 Known Parameters

The error rate calculations for the MK setting are the same as the CK setting, except that we must now consider the model M_{f^*} that is closest to M_t in terms of KL divergence to be the most desirable model. The error rates for MK can be expressed as:

$$\begin{aligned}
 P^{MK}(\text{FP}) &= \sum_{j \neq t^*} P_t(b = j) \min \left\{ \frac{\alpha}{\max_{r \neq j} (P_r(b \neq r))}, 1 \right\} \\
 &\leq \alpha \sum_{j \neq t^*} \frac{P_t(b = j)}{\max_{r \neq j} (P_r(b \neq r))} \\
 P^{MK}(\text{FN}) &= P_t(b = t^*) \left(1 - \min \left\{ \frac{\alpha}{\max_{r \neq t^*} (P_r(b \neq r))}, 1 \right\} \right)
 \end{aligned}$$

DELETE will be below α when

$$\sum_{j \neq t^*} \frac{P_t(b = j)}{\max_{r \neq j} (P_r(b \neq r))} < 1$$

Unfortunately, whether this inequality holds in practice is difficult to assess because we have assumed that the true distribution is outside the candidate model set. Nonetheless, both error rates will still converge to zero because $\max_{r \neq j} (P_r(b \neq r))$ and $P_t(b \neq t^*)$ will converge to zero asymptotically.

Figure 5 presents a simulation study across sample sizes that are multiples of 50 between 50 and 2,000 inclusive. Here, $\mathcal{M} = \{N(3.8, 1), N(4.1, 1), N(4.15, 1)\}$, $M_t = N(4, 1)$, $M_{t^*} = N(4.1, 1)$ and the IC used is the negative log likelihood. We set $\alpha = 0.05$, sample 7,000 draws from the true model for each sample size to generate observed data, and simulate 7,000 data sets for performing ECIC. As

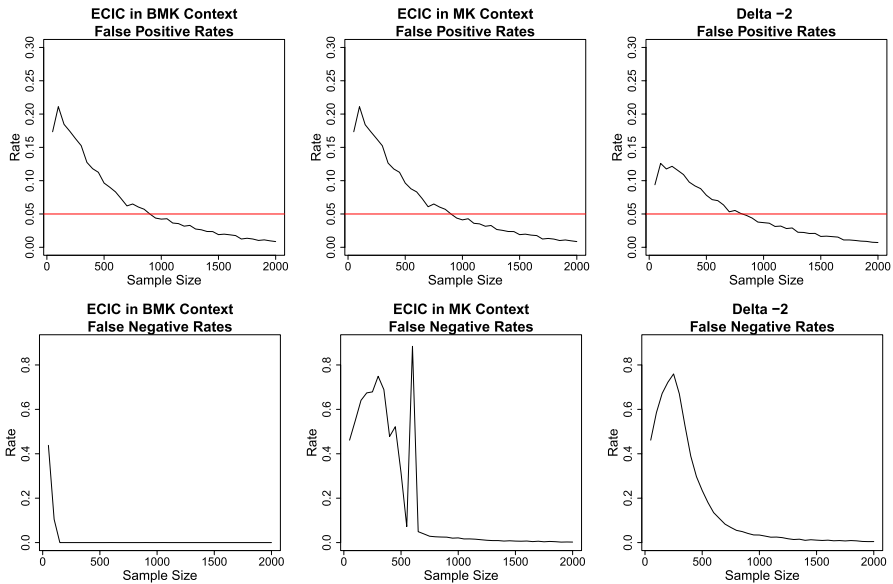


Fig. 5 Simulation study where the top three graphs correspond to linearly interpolated FP rates using thresholds for ECIC in the BMK context, ECIC in the MK context, and Burnham and Anderson’s rule of thumb with a threshold of -2 . The bottom graphs correspond to linearly interpolated FN rates under these methods

previously warned, the FP rates for ECIC under the benchmark and in-practice thresholds are inflated above 0.05 until samples sizes are larger than about 750. The rule of thumb approach also experiences inflation above the nominal rate, albeit less dramatically. A similar sudden spike in FN rates as in the CK case appears in the MK case as well. Lastly, all error rates for ECIC eventually decay to zero with increasing sample size as expected.

3.3.2 Unknown Parameters

Lastly, the MU setting closely resembles the CU setting. The error rates can be expressed as:

$$P^{MU}(\text{FP}) \leq \alpha \sum_{j \neq t^*} \frac{P_t(b=j)}{\max_{r \neq j} (P_r(b \neq r | \hat{M}_r))}$$

$$P^{MU}(\text{FN}) \leq P_t(b=t^*) \left(1 - \frac{\alpha}{\max_{r \neq t^*} (P_r(b \neq r | \hat{M}_r))} \right).$$

Using results from the previous cases, we can see that both approach zero asymptotically.

The simulation setup for the MU case will exactly resemble the CU setup except that the first model in \mathcal{M} places knots at the noniles of $[-10, 10]$ and the IC used is the AIC (more on this shortly). We generate the exact same data from the true model, which places knots at the quintiles of $[-10, 10]$, as we did in the CU example. In Fig. 6, we fit all three models to the first of the 2,000 simulated draws of 20 points and again see that it is difficult to select a model by visual inspection alone.

A slight complication with the MU simulation is exactly how to define the model that is closest to the truth and thus should be selected. In the MK simulation we could consider the Gaussian distribution whose mean had the closest absolute distance to the true mean as closest. Since parameters are not fixed in the current context, such a straightforward scheme is not possible. We use the AIC to determine the closest model since it will select the model that minimizes KL divergence with increasing sample size (Vrieze, 2012). Thus, after examining the best observed AIC scores for the data we simulated, we found that the regression model with knots at the noniles was overwhelmingly selected as sample size increased. Therefore, we consider this model to be the one closest to the true model.

Figure 7 displays the results of our simulations. We actually see in this case that FP rates using both the benchmark and in-practice threshold for ECIC remain bounded below 0.05, whereas the rule of thumb has inflated values until the sample size reaches about 180. Compared to the CK simulation, the FN rates drop significantly slower with increasing sample size, but an asymptotic decrease to zero is apparent for each method, nonetheless. Together with the MK case, this illustrates how ECIC is sometimes able to control FP rates for misspecified model sets, but this depends on the particulars of the model selection context.

Regression Spline Fits

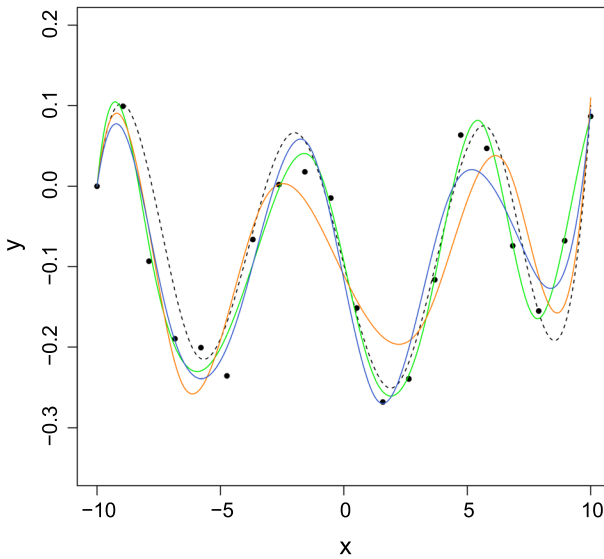


Fig. 6 A fit of all three spline regression models in the BMU/MU context to a simulated draw of 20 points. The dashed black line represents the (true) model with knots at the quintiles of $[-10, 10]$. The green, orange, and blue lines represents a cubic spline regression fit to the data using knots at the noniles, sextiles, and septiles of $[-10, 10]$, respectively

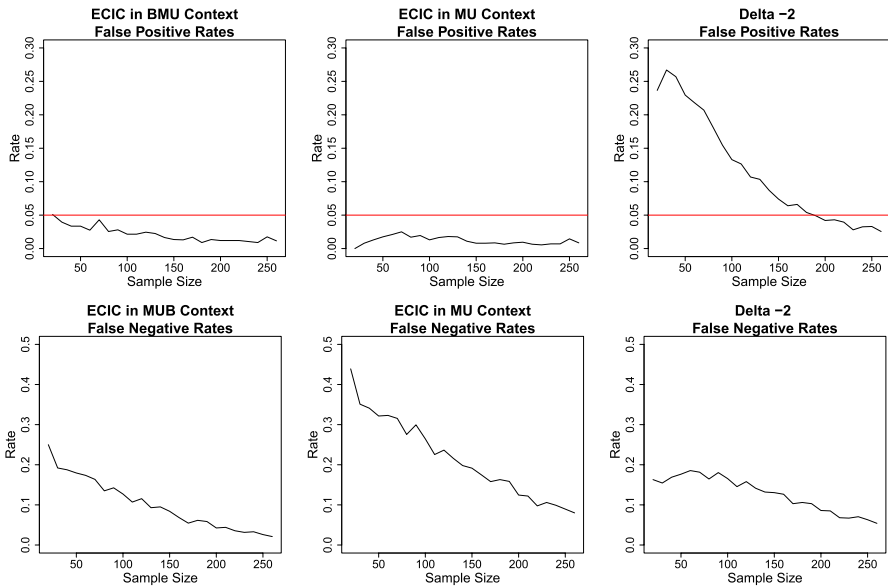


Fig. 7 Simulation study where the top three graphs correspond to linearly interpolated FP rates using thresholds for ECIC in the BMU context, ECIC in the MU context, and Burnham and Anderson's rule of thumb with a threshold of -2 where the red lines mark the nominal α level. The bottom graphs correspond to linearly interpolated FN rates under the same methods

4 Discussion

Our results provide a proof-of-concept result that NP testing can be generalized to operate in a fully information-theoretic model selection paradigm. Improving on prior work by (Cullan et al., 2019), we presented a revised algorithm for error control using the AIC and BIC, and we established some key theoretical properties of the algorithm's finite sample and asymptotic behaviors. In this section, we discuss some of the broader implications of our results for ongoing debates about statistical methodology in biology. Our focus will be on addressing the criticisms in (Dennis et al., 2019) of the practical limitations of error statistics based on classical NP tests. Overcoming these practical limitations may open the door to new epistemic interpretations and norms for applying severity as a form of evidence, e.g., in response to (Bandyopadhyay et al., 2016b; Taper & Ponciano, 2016).

The ultimate project is to understand the epistemic implications of using the AIC and BIC, which have a deep theoretical foundation in information theory, for model selection practices in the sciences. Classical Fisherian significance and NP tests both use a test statistic that compares the data to a null model selected before observing the data, and which may or may not be true. Information-theoretic model selection, in contrast, uses statistics such as the AIC or BIC to estimate a candidate model's relative divergence from the true distribution (Ponciano & Taper, 2019). How does this underlying shift in reference point and theoretical background affect our understanding of statistical evidence?

Evidentialist statistics represents one answer, which developed as an outgrowth of Royall's likelihoodist theory of statistical evidence (Royall, 1997, 2000) in response to modeling practices in ecology (Lele, 2004; Taper & Ponciano, 2016; Dennis et al., 2019; Taper & Lele, 2021). A key advance was defining evidence functions as a generalization of likelihoods that keep some of their most important features while accommodating a broader array of measures, such as the AIC and BIC, for the divergence between distributions (Lele, 2004; Dennis et al., 2019). On the evidentialist approach, evidence is defined by comparing the scores of two or more models using an evidence function. The larger the difference, the stronger the evidence. An observed score difference can also be contextualized by calculating probabilities of misleading or weak evidence.

Error statistics, based on the concept of severity, has often been viewed as a mutually exclusive alternative to evidentialism. Historically, error statistics developed from Karl Popper's falsificationist theory of the scientific method in response to a growing appreciation of the specific reasoning practices scientists use in designing and interpreting experiments (Mayo, 1996; Matthewson & Weisberg, 2009; Spanos & Mayo, 2015). The key advance here was the concept of a severe test: "Hypothesis H passes a severe test with [evidence] e if (i) e fits [alternative] H^* and (ii) the test procedure T has a very low probability of producing a result that fits H as well as (or better than) e does, if H were false or incorrect" (Matthewson & Weisberg, 2000, p. S198).

Previously, advocates for error statistics have been critical of information-theoretic methods precisely because they do not guarantee any particular level of error

probabilities (Spanos, 2010). At the same time, many of the favorite examples used by advocates of error statistics rely on problematic model selection methods from an evidentialist perspective. Classical NP tests, for example, do not guarantee that the probability of strong evidence for the true model converges to one as the sample size increases (Dennis et al., 2019). The main frequentist and Bayesian justifications of severity as a theory of evidence also rely on the true model assumption (Mayo & Spanos, 2006; Bandyopadhyay & Brittan, 2006). However, many scientists view this assumption as false, or at least deeply flawed, in fields such as ecology (Burnham & Anderson, 2002; Anderson, 2008; Aho et al., 2014; Sterner & Lidgard, 2021).

Our results using ECIC suggests these criticisms may be overcome in part through technical advances in model selection methods. However, we also need to consider how ECIC relates to existing theories of evidence. We approach this by considering ECIC as a statistical procedure that may be interpreted as compatible with one or more philosophical interpretations. We suggest it is compatible with error statistics, evidentialism, and the behavioristic view that Jerzy Neyman sometimes adopted for classical NP tests.

ECIC can be interpreted as implementing severe testing in an analogous way that Mayo and Spanos proposed for classical NP tests (Mayo & Spanos, 2006). Let model \hat{M}_{j^*} be the most conservative model that determines the value of the quantile threshold, $q(\tau)$. Then one can re-express the decision function h as: if $P(\Delta f < \Delta f_{obs} | \hat{M}_{j^*}) < P(\Delta f < \hat{q}(\tau) | \hat{M}_{j^*})$ then $h = 1$, else $h = 0$. This shows how $P(\Delta f < \Delta f_{obs} | \hat{M}_{j^*})$ is analogous to a p-value, because it expresses a conservative estimate of the probability that one would get an observed score difference at least as large if the observed best model was false. One can therefore define a severity function using h in the same way as for classical NP tests.

Another compatible interpretation is the “behavioristic” view of the decision function h as a rule that can be applied to many datasets, e.g., as part of quality testing in a factory or classifying patterns of variation in a large population of genetic sequences. On this view, ECIC is useful as a way to constrain the long-run frequentist properties of the model selection decision.

From an evidentialist perspective, what matters is the Δf_{obs} function at the core of ECIC, which qualifies as an evidence function according to the five main properties proposed in (Lele, 2004). The first two properties are translation and scale invariance. Since $P(\Delta f_{obs} < \hat{q}) = P(\Delta f_{obs} + c < \hat{q} + c)$ and $P(\Delta f_{obs} < \hat{q}) = P(c\Delta f_{obs} < c\hat{q})$ for any real constant $c \neq 0$, the observed DGOF and elements of each of the $j \neq b$ bootstrapped distributions can be translated and scaled by the same value without changing the decision of ECIC. The third property is invariance to one-to-one reparameterization of the parameter space. This follows directly from the invariance property for general transformations of MLEs (Casella & Berger, 2002), which implies that the ordering of the IC scores computed in both the initial data scoring and bootstrapping steps of the ECIC algorithm would be preserved. The fourth property is invariance to one-to-one transformations of the data. This follows directly from the fact that likelihood ratios, and thus differences in the IC scores, are invariant to one-to-one transformations of random variables due to a cancellation

of the Jacobian terms. This again preserves the ordering of IC scores throughout the algorithm. The final property is the probability of strong evidence for the true hypothesis converging to 1 as the sample size increases. For ECIC at a specified level α , we can interpret this as the FP and FN rates both converging to 0 with increasing sample size, which we demonstrated in Section 3. ECIC can therefore be understood as using a transformation of an evidence function, i.e. the decision procedure h as a function of Δf_{obs} , to control error rates.

Table 1 summarizes how ECIC leverages the benefits of the AIC and BIC to achieve many of the same benefits for severe testing as are available for the evidentialist approach. We discuss each of the items in more detail here:

- *Equal status for null and alternatives:* As with the evidentialist approach, ECIC treats all the candidate models a priori symmetrically. Post-data, it does distinguish between the the observed best-scoring model M_b and the remaining models in the candidate model set. One can therefore think of ECIC as comparing a post-data alternative model (the observed best model) to a set of null models (the remaining candidate models), but this is not identical to the procedure used in classical NP testing.
- *Allows evidence for Null:* Evidence for M_b , which in principle can be any $M_i \in \mathcal{M}$, is encapsulated in the observed DGOF Δf_{obs} , which is also the basis for carrying out a severe test. If passed, Δf_{obs} is deemed to be sufficient evidence in support of M_b and is insufficient otherwise. It is therefore possible for ECIC to find evidence for any of the candidate models, which is not the case for the null model in classical NP tests.
- *Accommodates multiple models:* ECIC can, in principle, support any finite set of non-overlapping candidate models. It further inherits the property of IC-based model selection where all pairs of models can be compared.

Table 1 A comparison of inferential characteristics between classical frequentist tests, ECIC, and evidential statistics

Inferential characteristic	P-value	Classical NP	ECIC	Evidence
Equal status for null and alternatives	NA	No	Yes	Yes
Allows evidence for Null	No	No	Yes	Yes
Accommodates multiple models	No	Awkward	Yes	Yes
All error rates go to zero as sample size increases	No	No	Yes	Yes
Total error rate always decreases with increasing sample size	No	No	Unknown	Yes
Can be used with non-nested models	NA	Not standard	Yes	Yes
Evidence and error rates distinguished	No	No	Yes	Yes
Robust to model misspecification	Yes	No	Partial	Yes
Promotes exploration of new models	Yes	No	Yes	Yes

The table is updated from (Dennis et al., 2019) to highlight how ECIC addresses key practical criticisms of the error statistical approach by showing it is not limited to classical testing methods. See the main text for more explanation

- *All error rates go to zero as sample size increases*: This was demonstrated in Section 2 for the CK, CU, MK, and MU contexts. ECIC therefore benefits from the statuses of the ΔBIC and ΔAIC (under certain circumstances) as evidence functions, but it provides an alternative way of understanding evidence than the simple magnitude of the ΔIC score.
- *Total error rate always decreases with increasing sample size*: The total error rate for ECIC would be defined as the sum of false positives and false negatives. Since our implementation of ECIC relies on a numerical algorithm involving stochastic elements, it is unlikely that the total error rate will always decrease in a strictly monotonic way with increasing sample size. One could potentially investigate the expected behavior of $FP + FN$ using the theoretical framework we introduced, but we have not done so here, and therefore we list it as unknown at this point.
- *Evidence and error rates distinguished*: ECIC is compatible with both severity and evidentialist interpretations. ECIC can be understood as a decision procedure based on a transformation of an evidence function. It can also be understood as a severe test because the decision is based on the probability that one would get at least as good an observed score difference if the observed best model is false. As a result, it permits a clear distinction between evidence and error rates, although not necessarily in the same way that evidentialists do.
- *Robustness to model misspecification*: The standard set by (Dennis et al., 2019) is the asymptotic decay of error rates as sample size increases, which ECIC has been shown to meet. However, ECIC explicitly sets a higher standard for robustness that is not necessarily met for misspecified model sets, i.e., control of FP rates at a nominal level α . We have seen that ECIC sometimes succeeds at this in misspecified contexts, but this depends on the particularities of how the candidate models relate to the external true distribution and is not universally guaranteed. We therefore list it as partially satisfied by ECIC at this time.
- *Promotes exploration of new models*: This criterion is somewhat vague, but we understand the main thrust of Dennis et al.'s critique to be the assumption of classical NP tests that the null and alternative hypothesis form a closed, complementary set of possibilities. Since ECIC accommodates an indefinite number of candidate models, it is consistent with commonly recommended practices of exploring model adequacy as a means to identify new models, especially from an error statistical perspective. For this reason we list it as "Yes."

Because ECIC changes so much about the practical features of severe testing, it may also suggest new ways to draw epistemic conclusions from the results of a severe test. We highlight a few points of interest in lieu of a more comprehensive discussion that remains for future work. One central question for model selection, for example, is whether we have good reason to believe a hypothesis is true when it passes one or more severe tests. This has been a point of contention with evidentialists, who call it the "true-model" assumption of error statistics (Bandyopadhyay et al., 2016b; Bandyopadhyay & Brittan, 2006), p. 75). Since ECIC can (in some cases) operate successfully in a misspecified model context, this assumption is therefore not required for severe testing. Instead, the assumption might be better labeled as the

“quasi-true” model assumption, since ECIC can provide us with a severe test that the best observed model contains the distribution closest to the truth.

ECIC is designed to be stringent in ruling out relevant alternatives to M_b , where relevance is determined by inclusion in the candidate set. The key benefit of this is that some degree of control is provided for FP rates in finite samples. However, as we saw in simulations for simple modeling contexts, this can also result in nontrivial FN rates for finite samples. Thus, ECIC is most suited for situations in which it is acceptable or perhaps even informative to conclude insufficient evidence for M_b .

The preceding notion can be naturally tied to an emphasis on *representational fidelity*, or how well a model describes the causal structure of a data generating process (Matthewson & Weisberg, 2008). That is, in conceiving which candidate models may be included in \mathcal{M} , it may be most informative to include ones with distinct explanatory features. The idea here would be to have the selection of M_b implicating reliable information about the underlying mechanics of a data generating process. On the other hand, not selecting M_b would implicate some degree of ambiguity for such mechanics.

A final point to be made here concerns the cardinality of \mathcal{M} . As mentioned before, there is no restriction for this beyond finiteness in principle. However, given the conservative nature of ECIC, including superfluous models in \mathcal{M} may unnecessarily frustrate the chances of a model being selected. For example, in keeping with the spirit of our spline examples in Section 3, ECIC would perform poorly if \mathcal{M} were to include many spline regression models that are all fairly competitive in terms of a chosen information criterion. This, however, would also apply to ΔIC procedures, and so it is not a unique problem to ECIC. Putting these suggestions together, the ideal setup for ECIC would involve a manageable-sized set of models that each have distinct explanatory features and would yield useful information both when a decision is and is not made. While it is difficult to designate a precise number for what qualifies as a manageable model set, simulation studies under plausible conditions can provide some guidance.

5 Conclusion

We identify several areas for future work emerging from ECIC’s finite sample and asymptotic properties. On the philosophical side, an important next step is to develop a more comprehensive and systematic epistemology for severe testing using ECIC. This would further clarify its general implications and novelty in the philosophy and methodology of statistics.

It may also be possible to extend our theoretical framework to provide stronger guarantees for the misspecified model case and to cover the important class of model selection problems where one or more candidate models are nested. For the former issue, it would be important to improve the reliability of ECIC under model misspecification. A plausible future strategy to achieve a more general guarantee for FP would be to use non-parametric bootstrapping to estimate the DGOF distribution and estimator $\hat{\pi}_{NP}$ (Taper & Lele, 2021). A smoothing approach may be required for small sample sizes, however, to avoid artifacts in

the distribution of score differences. This would also restrict the scope of the approach to settings where an unbiased non-parametric bootstrap estimator exists.

In addition, it may be possible to generalize the ECIC algorithm to apply to any combination of nested and non-nested models. Shao and Rao provide an important result here using the Generalized Information Criterion for linear regression (Shao & Rao, 2000). More generally, when two models are nested in \mathcal{M} , e.g., $M_i \in M_j$, we must provide a more nuanced interpretation of when a model counts as a false positive in order to apply ECIC. In particular, if $M_i(\theta_i)$ is the true distribution, then there is no unique value of t , since $M_i(\theta_i) \in M_j$. Nonetheless, the convention is to designate selecting the simpler model as correct, so that a statistically consistent criterion should eventually (almost) always score $M_i(\theta_i)$ as best.

This poses a challenge for NP testing, since when the BIC converges on selecting $M_i(\theta_i)$, it will also appear that \hat{M}_j has a very high false positive rate. In this context, the algorithm we defined for non-nested models will control the FP rate as desired, but the FN rate for $M_i(\theta_i)$ will converge to 1 instead of 0. It is odd to say that choosing a subset of M_j should count as a false positive, though, suggesting that this context will require us to reconsider exactly how we define our candidate models to reflect their nested logical relationships.

From an error statistical perspective, one strategy may be to recognize that hypothesis tests are not as perfectly precise in practice as the mathematical models we use to represent them. Instead, we can interpret the BIC as actually the sum of two values: (1) a model fitting penalty that corrects for expected bias in estimating KL divergence from the true distribution, and (2) a tacit effect size penalty that privileges simpler models. The first part corresponds to the AIC in the non-nested case, while the second part reflects the additional penalty imposed by the BIC, i.e., $k \log(n) - 2k$ for $\log(n) > 2$. In effect, we can understand this additional penalty as asserting that any distribution in M_j within a KL divergence of $k \log(n) - 2k$ of M_i should be treated as part of M_i . If M_j is related to M_i by setting a parameter to zero, this assertion corresponds to restricting M_j to sufficiently large values of the parameter, and hence we can interpret it as an implicit effect size requirement that modulates the boundaries of the smaller model in parameter space. In other words, it may be possible to transform the nested models case into a non-nested case by excluding distributions in the larger model from consideration when they are too close to the smaller model.

On the applied side, it may be possible to improve ECIC's performance by integrating methods and theoretical results from adjacent topics, e.g., research on error exponents for Markov model classification (Chambaz, 2006; Eguchi & Copas, 2006; Leong et al., 2007) and non-parametric NP classification (Tong et al., 2016, 2018). These may suggest strategies to improve ECIC's finite sample error rates and provide analytical bounds for more complex asymptotic cases.

Funding Funding was provided by John Templeton Foundation (Grant No. 62220).

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

- Aho, K., Derryberry, D. W., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, 95(3), 631–636. <https://doi.org/10.1890/13-1452.1>
- Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. London: Springer.
- Bandyopadhyay, P. S., & Boik, R. J. (1999). The curve fitting problem: A Bayesian rejoinder. *Philosophy of Science*, 66(S3), S390–S402.
- Bandyopadhyay, P. S., & Brittan, G. G. (2006). Acceptability, evidence, and severity. *Synthese*, 148(2), 259–293. <https://doi.org/10.1007/s11229-004-6222-6>
- Bandyopadhyay, P. S., Brittan, G. G., & Taper, M. L. (2016). Error-statistics, evidence, and severity. In P. S. Bandyopadhyay, G. G. Brittan, & M. L. Taper (Eds.), *Belief, evidence, and uncertainty: Problems of epistemic inference* (pp. 73–91). Cham: Springer International Publishing.
- Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.
- Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7(6), 679–692.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and inference: A practical-theoretic approach*. New York: Springer.
- Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Wadsworth Group. Duxbury.
- Chambaz, A. (2006). Testing the order of a model. *The Annals of Statistics*, 34(3), 1166–1203.
- Cullan, M., Lidgard, S., & Sterner, B. (2020). Controlling the error probabilities of model selection information criteria using bootstrapping. *Journal of Applied Statistics*, 47(13–15), 2565–2581.
- Dennis, B., et al. (2019). Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution*, 7, 372.
- Ding, J., Tarokh, V., & Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6), 16–34. <https://doi.org/10.1109/MSP.2018.2867638>
- Dziak, J. J., et al. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 21(2), 553–565. <https://doi.org/10.1093/bib/bbz016>
- Eguchi, S., & Copas, J. (2006). Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma. *Journal of Multivariate Analysis*, 97(9), 2034–2040. <https://doi.org/10.1016/j.jmva.2006.03.007>
- Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1), 1–35.
- Glattig, G., et al. (2007). Choosing the optimal fit function: Comparison of the Akaike information criterion and the F-test. *Medical physics*, 34(11), 4285–4292.
- Hegyí, G., & Laczi, M. (2015). Using full models, stepwise regression and model selection in ecological data sets: Monte Carlo simulations. *Annales Zoologici Fennici*, 52(5), 257–279. <https://doi.org/10.5735/086.052.0502>
- Hunt, G. (2006). Fitting and comparing models of phyletic evolution: Random walks and beyond. *Paleobiology*, 32(4), 578–601.
- Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research*, 33(2), 188–229. <https://doi.org/10.1177/0049124103262065>
- Lele, S. R. (2004). *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 191–216). Chicago: The University of Chicago Press.
- Leong, A. S., Dey, S., & Evans, J. S. (2007). Error exponents for Neyman–Pearson detection of Markov chains in noise. *IEEE Transactions on Signal Processing*, 55(10), 5097–5103. <https://doi.org/10.1109/TSP.2007.897863>
- Markatou, M., Karlis, D., & Ding, Y. (2021). Distance-based statistical inference. *Annual Review of Statistics and Its Application*, 8(1), 301–327. <https://doi.org/10.1146/annurev-statistics-031219-041228>
- Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics*, 34, 593–610. <https://doi.org/10.1007/s10519-004-5587-0>

- Matthewson, J., & Weisberg, M. (2008). The structure of tradeoffs in model building. *Synthese*, 170, 169–190.
- Matthewson, J., & Weisberg, M. (2009). Learning from error, severe testing, and the growth of theoretical knowledge. In D. G. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 28–57). Cambridge: Cambridge University Press.
- Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.
- Mayo, D. G. (2000). Experimental practice and an error statistical account of evidence. *Philosophy of Science*, 67(S3), S193–S207.
- Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, 57(20), 323–357.
- Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27(2), 392–403.
- Pesaran, M. H. (1990). Non-nested hypotheses. *Econometrics* (pp. 167–173). London: Palgrave Macmillan.
- Ponciano, J. M., & Taper, M. L. (2019). Model projections in model space: A geometric interpretation of the AIC allows estimating the distance between truth and approximating models. *Frontiers in Ecology and Evolution*, 7, 413.
- Ripplinger, J., & Sullivan, J. (2008). Does choice in model selection affect maximum likelihood analysis? *Systematic Biology*, 57(1), 76–85. <https://doi.org/10.1080/10635150801898920>
- Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. London, UK: Chapman & Hall.
- Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 95(451), 760–768.
- Sayyareh, A., Obeidi, R., & Bar-Hen, A. (2010). Empirical comparison between some model selection criteria. *Communications in Statistics-Simulation and Computation*, 40(1), 72–86. <https://doi.org/10.1080/03610918.2010.530367>
- Shao, J., & Rao, J. S. (2000). The GIC for model selection: A hypothesis testing approach. *Journal of Statistical Planning and Inference*, 88(2), 215–231. [https://doi.org/10.1016/S0378-3758\(00\)00080-X](https://doi.org/10.1016/S0378-3758(00)00080-X)
- Spanos, A. (2010). Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics*, 158(2), 204–220. <https://doi.org/10.1016/j.jeconom.2010.01.011>
- Spanos, A., & Mayo, D. G. (2015). Error statistical modeling and inference: Where methodology meets ontology. *Synthese*, 192, 3533–3555. <https://doi.org/10.1007/s11229-015-0744-y>
- Sterner, B., & Lidgard, S. (2024). Objectivity and underdetermination in statistical model selection. *The British Journal for the Philosophy of Science*, 75(3), 717–739. <https://doi.org/10.1086/716243>.
- Sullivan, J., & Joyce, P. (2005). Model selection in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 36(1), 445–466. <https://doi.org/10.1146/annurev.ecolsys.36.102003.152633>
- Taper, M. L., Lele, S. R., Ponciano, J. M., Dennis, B., Jerde, C. L. (2021) Assessing the global and local uncertainty of scientific evidence in the presence of model misspecification. *Frontiers in Ecology and Evolution*, 9, 679155. <https://doi.org/10.3389/fevo.2021.679155>
- Taper, M. L., & Ponciano, J. M. (2016). Evidential statistics as a statistical modern synthesis to support 21st century science. *Population Ecology*, 58, 9–29. <https://doi.org/10.1007/s10144-015-0533-y>
- Tong, X., Feng, Y., & Li, J. J. (2018). Neyman–Pearson classification algorithms and NP receiver operating characteristics. *Science Advances*. VOLME? PAGE NUMBERS?<https://doi.org/10.1126/sciadv.aao1659>
- Tong, X., Feng, Y., & Zhao, A. (2016). A survey on Neyman–Pearson classification and suggestions for future research. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(2), 64–81. <https://doi.org/10.1002/wics.1376>
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2), 228.
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 57(2), 307. <https://doi.org/10.2307/1912557>
- Wagenmakers, E. J., et al. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48(1), 28–50.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.