Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8,** 47–56 doi: 10.1111/2041-210X.12648

A graphical framework for model selection criteria and significance tests: refutation, confirmation and ecology

Ken Aho[1]*, Dewayne Derryberry[2] and Teri Peterson[3]

[1]*Department of Biological Sciences, Idaho State University, Pocatello, ID 83209, USA;* [2]*Department of Mathematics and Statistics, Idaho State University, Pocatello, ID 83209, USA; and* [3]*Department of Management, Idaho State University, Pocatello, ID 83209, USA*

Summary

1. In this study, we use a novel graphical heuristic to compare the way four methods: significance testing, two popular information-theoretic approaches (AIC and BIC) and Good's Bayes/non-Bayes compromise (an underutilized hypothesis testing approach whose demarcation criterion adjusts for *n*), evaluate the merit of competing hypotheses, for example H_0 and H_A.

2. A primary goal of our work is to clarify the concept of strong consistency in model selection. Explicit considerations of this principle (including the strong consistency of BIC) are currently limited to technical derivations, inaccessible to most ecologists. We use our graphical framework to demonstrate, in simple terms, the strong consistency of both BIC and Good's compromise.

3. Our framework also locates the evaluated metrics (and ICs in general) along a conceptual continuum of hypothesis refutation/confirmation that considers *n*, parameter number and effect size. Along this continuum, significance testing and particularly AIC are refutative for H_0, whereas Good's compromise and particularly BIC are confirmatory for the true hypothesis.

4. Our work graphically demonstrates the well-known asymptotic bias of significance tests for H_A, and the incorrectness of using statistically non-consistent methods for point hypothesis testing. To address these issues, we recommend: (i) dedicated confirmatory methods with strong consistency like BIC for use in point hypothesis testing and confirmatory model selection; (ii) significance tests for use in exploratory/refutative hypothesis testing, particularly when conjoined with rational approaches (e.g. Good's compromise, power analyses) to account for the effect of *n* on *P*-values; and (iii) asymptotically efficient methods like AIC for exploratory model selection.

Key-words: Akaike Information Criterion, Bayes factor, Bayesian Information Criterion, confirmation test, graph, model selection, neutral model, null model, *P*-value, significance test

Introduction

When making formal statistical inferences, ecologists must choose from among a wide array of hypothesis testing methods and model selection approaches. Unfortunately, these choices are hampered by the absence of a general framework for defining and distinguishing the characteristics and *purpose* of particular methods. For instance, statisticians and statistical ecologists have thus far described the asymptotic efficiency of the Akaike Information Criterion (AIC, Akaike 1973) and strong consistency of the Bayesian Information Criterion (BIC, Schwarz 1978) in rather abstract terms (Hooten & Hobbs 2014, p. 15), and without a contextual framework for frequentist significance tests (FSTs). Further, existing comparisons of FSTs and information-theoretic criteria (ICs) generally disregard BIC and ignore the effects of sample size, effect size and parameter number. For example, Murtaugh (2014) described the relatedness of *P*-values to ΔAIC for nested alternative models differing by one parameter, but did not consider BIC, and more complex multiparameter settings. Conversely, Burnham & Anderson (1998, pp. 337–339) distinguished ΔAIC and significance testing for nested models with widely differing numbers of parameters, but ignored BIC, and the potential effect of sample size. To address this deficiency, we present a simple graphical heuristic – which simultaneously accounts for sample size, effect size and number of parameters – to compare four inferential approaches: frequentist significance testing, AIC, BIC and Good's Bayes/non-Bayes compromise (an underutilized hypothesis testing approach whose demarcation criterion adjusts for *n*). Our work is intended to: (i) provide a thorough but non-technical description of strong consistency in model selection and (ii) clarify the purpose and trade-offs of inferential methods and their correct (and incorrect) uses in ecology.

AIC AND BIC

AIC and BIC are often used by ecologists to identify models that balance uncertainty, caused by excessive complexity, and bias, resulting from model oversimplification. These metrics have the form:

*Correspondence author. E-mail: ahoken@isu.edu

$$\mathrm{AIC} = -2\hat{\ell} + 2p \qquad \text{eqn 1}$$

$$\mathrm{BIC} = -2\hat{\ell} + p\log(n), \qquad \text{eqn 2}$$

where $\hat{\ell}$ is the log-likelihood of the estimated model, and p = the total number of parameters estimated in the model, including σ in general linear models.

AIC is a maximum-likelihood estimator for Kullback–Leibler information (Kullback & Leibler 1951) that corrects for the bias associated with likelihood maximization. AIC is efficient (Box 1). That is, as sample size tends to infinity, a minimum AIC model minimizes the bias-corrected KL distance from the true data-generating process. Thus, as sample size increases, AIC will identify the true best predictive model from a group of candidate approximating models.

BIC is the asymptotic approximation, for the regular exponential family, of a Bayesian hypothesis testing procedure. Specifically, BIC is approximately two times the log of a Bayes factor (the ratio of the marginal densities for two models) resulting from the comparison of the model of interest and a saturated model. BIC has strong consistency for the true model (Box 1). Thus, as sample size approaches infinity, the true model, from a group of models, will have the smallest BIC value. BIC uses the highly diffuse unit-information prior distribution (Fox 2015). Notably, however, BIC is not strictly Bayesian because of its use of within-sample likelihood maximization (see Hooten & Hobbs 2014, p. 15).

Many alternatives to AIC and BIC exist, although they are seldom used by ecologists (Aho, Derryberry & Peterson 2014). Other approaches include ICs that are asymptotically efficient (Akaike 1969; Mallows 1973; Sugiura 1978), strongly consistent (Hannan & Quinn 1979; Chen & Chen 2008; Zhang & Shen 2010) and strictly Bayesian (Watanabe 2010). Further, several authors have attempted to define ICs that possess both the efficiency of AIC and strong consistency of BIC (Bozdogan 1987). However, these properties are inextricable. A method cannot simultaneously have strong consistency and the predictive optimality of efficiency (Theorem 1, Yang 2005).

The magnitude of IC outcomes will generally increase with n, and the empirical frame of reference for ICs will shift as other response variables are considered. Thus, (i) IC outcomes for models with different response variables are not comparable and (ii) it is the *difference* in IC outcomes, not IC values themselves, that provide insight into model optimality. To aid in interpreting ΔICs, a generalized likelihood ratio can be obtained with:

$$\exp(0{\cdot}5\Delta\mathrm{IC}). \qquad \text{eqn 3}$$

For BIC, this transformation will approximate a Bayes factor. For instance, if $\Delta\mathrm{BIC} = 8$, the generalized likelihood ratio/Bayes factor approximation is 54·6, indicating that the smaller BIC model has 54·6 times more empirical support than the larger BIC model.

Box 1. Consistency and efficiency

Consistency: Assume that the true model or hypothesis is present within a set of candidate models or hypotheses under consideration. As n tends to infinity, a method with *strong consistency* will identify the true model or hypothesis with probability one, whereas a method with *weak consistency* will identify the true model or hypothesis with probability tending to one.

Efficiency: As n goes to infinity, an *efficient* method will identify, with probability one, the model or hypothesis whose squared prediction error best approximates the optimal theoretical model or hypothesis.

GOOD'S BAYES/NON-BAYES COMPROMISE

Good (1992) proposed an intuitive compromise between FSTs and Bayesian hypothesis testing based on standardizing P-values to a sample size of 100 with the transformation: min $(0{\cdot}5, P\sqrt{n}/10)$. The approach acknowledges the proportionality of a Bayes factor to $1/\sqrt{n}$ under H_0 when the P-value is fixed. Good's compromise addresses a fundamental criticism of significance testing, namely that, because P-values are partially a function of sample size, significant results may become scientifically meaningless as n grows large (Oakes 1986; Royall 1997; Johnson 1999).

Based on previously established generalized bounds (Chiani, Dardari & Simon 2003; Chang, Cosman & Milstein 2011), the bounds of the probit function in the context of Good's compromise are (Appendix S1; §7, Supporting Information):

$$0{\cdot}5\log(n) - 0{\cdot}5\log(2\pi/e) - c_{\mathrm{U}} \leq \left\{\Phi^{-1}\left(1 - \frac{5\alpha}{\sqrt{n}}\right)\right\}^2 \leq \log(n) - c_{\mathrm{L}}, \qquad \text{eqn 4}$$

where $c_{\mathrm{L}} = \log(100\alpha^2)$ and $c_{\mathrm{U}} = 0{\cdot}5c_{\mathrm{L}}$ are constants that vary with α, and $\Phi^{-1}(p)$ is the probit function (standard normal inverse CDF) at probability p. Thus, for one-parameter models, Good's compromise has similar asymptotic properties to BIC.

HYPOTHESIS TESTING UNDER THE EVALUATED METRICS

Statistical methods are most often applied by biologists for the purpose of hypothesis testing (Quinn & Keough 2002, p. 32). As a result, we evaluate methods here (including ICs) from the perspective of parametric inference, as opposed to predictive efficacy (e.g. Brewer, Butler & Cooksley 2016).

Consider the hypotheses:

$$H_0 : \mu = c,$$

$$H_A : \mu = \mu_A \neq c.$$

In the preceding statement, H_0 defines an exact *fixed* value, c, whereas H_A defines an exact but *unknown* parameter value, μ_A, distinct from c. To standardize the way inferential methods assess the validity of H_0 and H_A, we use the likelihood ratio test

statistic, $X^2 = -2(\hat{\ell}_0 - \hat{\ell}_A)$, that is twice the difference in maximized log-likelihoods under models representing H_0 and H_A.

For simplicity, we impose classical constraints in which: (i) hypotheses represent nested models (H_0 in H_A) that differ with respect to the inclusion or exclusion of a single parameter, and (ii) data are a random sample from a normal distribution with a known variance, σ^2 (cf. Pötscher 1991; Stoica, Selén & Li 2004; Dziak *et al.* 2012; Murtaugh 2014).

Demarcation

We define the *line of demarcation* for choosing H_0 or H_A, when using ICs, to be $\Delta IC = 0$ under those hypotheses. That is, let IC_0 and IC_A be the IC values under H_0 and H_A, respectively. Then

$$IC_0 - IC_A > 0 \text{ favours } H_A$$
$$IC_0 - IC_A < 0 \text{ favours } H_0.$$

Under these conditions, the AIC and BIC lines of demarcation are $X^2 = 2$ and $X^2 = \log(n)$, respectively (Appendix S1 §1–4; cf., Söderström 1977; Teräsvirta & Mellin 1986; Foster & George 1994; van der Hoeven 2005; Claeskens & Hjort 2008). We note that alternative demarcation criteria have been proposed for ΔAIC (Royall 1997; Burnham & Anderson 1998, pp. 70–71), Bayes factors (Jeffreys 1961) and ΔBIC (Raftery 1995, table 6).

For FSTs using significance level $\alpha = 0{\cdot}05$, the line of demarcation is the critical value $X^2 = \chi^2_{(1,0{\cdot}95)} \approx 1{\cdot}96^2$, wherein $\chi^2_{(1,0{\cdot}95)}$ denotes the chi-squared inverse CDF at probability 0·95. Good's (1992) compromise suggests the adjusted significance level $\alpha_{adj} = 10\alpha/\sqrt{n}$, resulting in the line of demarcation:

$$X^2 = \left\{\Phi^{-1}\left(1 - \frac{\alpha_{adj}}{2}\right)\right\}^2 = \chi^2_{(1,1-\alpha_{adj})}. \qquad \text{eqn 5}$$

Distribution of X^2 under H_0

Under our assumptions, the likelihood ratio test statistic will asymptotically follow a χ^2_1 distribution under H_0 (Wilks 1938). Therefore,

$$E(X^2|H_0) = 1, \qquad \text{eqn 6}$$

$$\mathrm{Var}(X^2|H_0) = 2, \qquad \text{eqn 7}$$

where $X^2|H_0$ denotes the sampling distribution of X^2 under H_0.

Distribution of X^2 under H_A

Under a series of local alternative hypotheses, the sampling distribution of X^2 will be asymptotically non-central-chi-squared distributed with one degree of freedom, with non-centrality parameter, δ, reflecting the deviation of H_A from H_0 (Sugiura 1969; Shapiro 2009). For *our* purposes, $\delta = n(\mu_A - c)^2/\sigma^2$ (Appendix S1; §5).

The mean of the non-central-chi-squared distribution is $k + \delta$, where k = the number of (central) chi-squared degrees of freedom, and the variance is $2k + 4\delta$ (Chun & Shapiro 2009).

Let

$$\gamma = \sqrt{\frac{\delta}{n}} = \frac{\mu_A - c}{\sigma},$$

represent effect size, we have, for the current application:

$$E(X^2|H_A) = k + \delta = 1 + n\gamma^2, \qquad \text{eqn 8}$$

$$\mathrm{SD}(X^2|H_A) = \sqrt{2k + 4\delta} = \sqrt{2 + 4n\gamma^2} = 2\sqrt{0{\cdot}5 + n\gamma^2}, \qquad \text{eqn 9}$$

where $X^2|H_A$ denotes the sampling distribution of X^2 under H_A. Thus, when n is large and the alternative hypothesis is true, the mean of X^2 grows in proportion to n, and the standard deviation of X^2 grows in proportion to $2\sqrt{n}$.

A graphical heuristic

The mathematical definitions given in the previous section are graphically expressed in Fig. 1.

Consider the role that σ^2 and $(\mu_A - c)^2$ play in determining the slope of $E(X^2|H_A)$. As σ^2 increases and $(\mu_A - c)^2$ decreases, the slope for the mean function flattens, and larger samples will be required to establish the invalidity of false null hypotheses (Fig. 1). A counterbalancing factor is that the flatter the slope, the smaller the standard deviation because of these same quantities (eqns 8 and 9).

Figure 2 extends Fig. 1 by showing X^2 under H_0, and X^2 under H_A for the levels of γ considered in Fig. 1. Note that the distributions for H_A and H_0 are probablistically indistin-

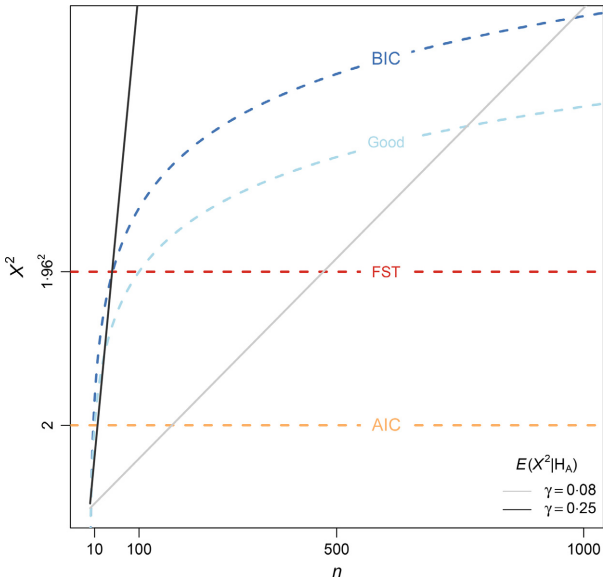

Fig. 1. Values of X^2 as a function of n. Dashed lines indicate lines of demarcation for Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Good's Bayes/non-Bayes compromise and frequentist significance tests (FSTs). If X^2 falls below the line, the respective framework supports H_0. Conversely, if X^2 falls above the line, the respective framework supports H_A. Solid lines show the means of X^2, under H_A, for differing effect sizes, γ. Any line with an intercept of 1 and linear in n with a positive slope is a candidate.

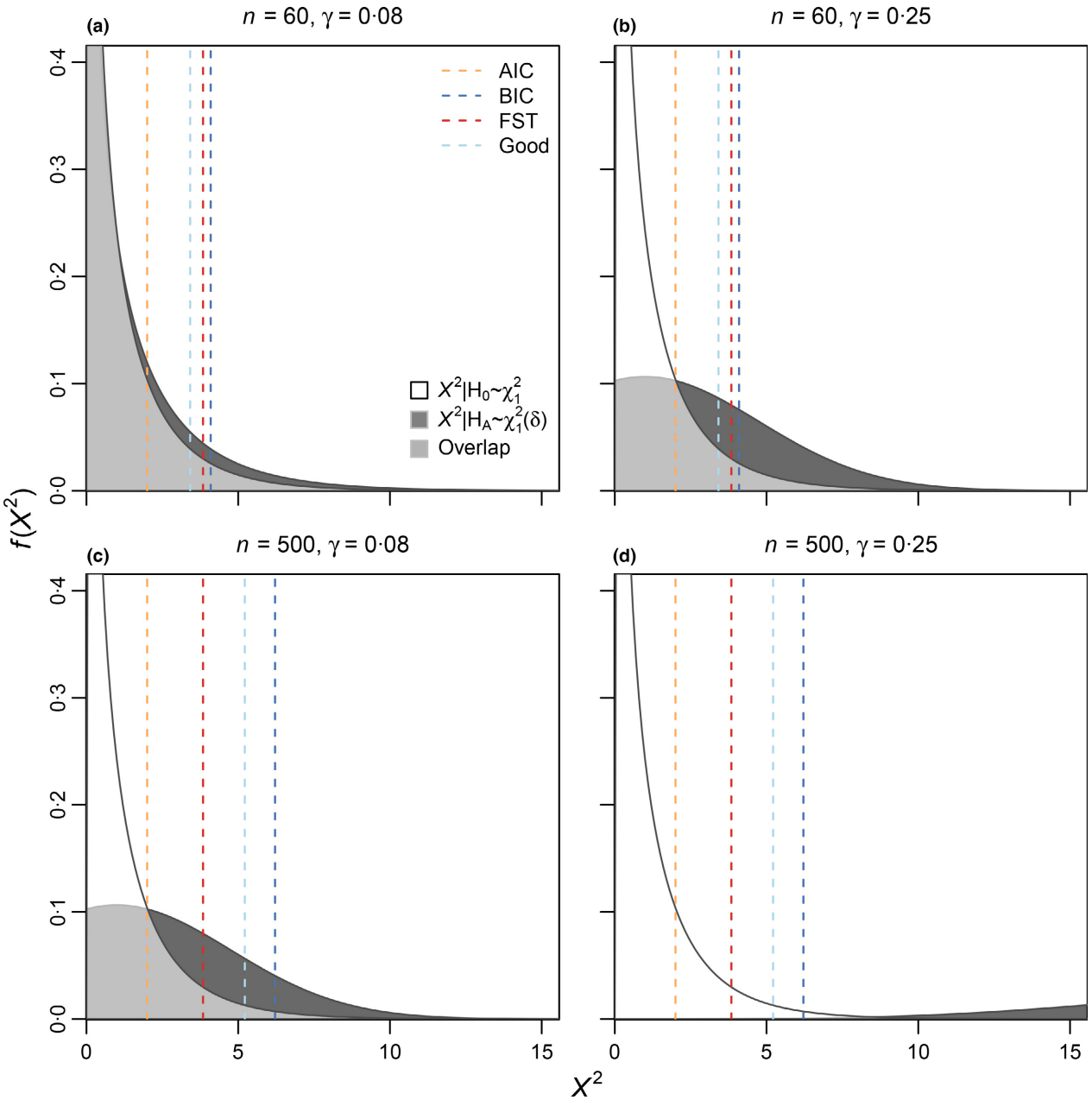

Fig. 2. Non-central-chi-squared with non-centrality parameter δ, that is $\chi^2_1(\delta)$, densities under H_A and central χ^2_1 densities under H_0, with respect to X^2 quantiles. Distributions of X^2 are shown for (a) $n = 60$, $\gamma = 0{\cdot}08$, (b) $n = 60$, $\gamma = 0{\cdot}25$, (c) $n = 500$, $\gamma = 0{\cdot}08$, (d) $n = 500$, $\gamma = 0{\cdot}25$. As in Fig. 1, dashed lines indicate lines of demarcation. See Appendix S3 for additional graphical perspectives on these relationships.

guishable for small sample sizes, but become increasingly distinct as n increases, particularly for larger effect sizes (Fig. 2d). The increasing variance of $X^2|H_A$ with n is also evident.

Figure 2 also depicts methodological power. Specifically, the power of a method is the area under the $X^2|H_A$ curve, above the demarcation bounds (dashed coloured lines). For instance, under BIC, and $|\gamma| = 0{\cdot}08$, a sample size of approximately 500 would be required to correctly select H_A over H_0 with probability (power) = 0·24 (Fig. 2c). For $|\gamma| = 0{\cdot}25$, power increases to approximately 0·999 for the same sample size (Fig. 2d). The ordering of demarcation lines is in agreement with Dziak *et al.* (2012) who previously defined AIC and BIC as methods that emphasize sensitivity (statistical power) and specificity (avoidance of type I error) in null hypothesis tests, respectively.

A CONTINUUM OF REFUTATION/CONFIRMATION

The likelihood ratio test statistic (the Y-axis in Fig. 1) measures the weight of evidence for H_A relative to H_0 (Johnson 2008). Indeed, the generalized likelihood ratio given in eqn 3 can be obtained directly from X^2. For example, under BIC we have $\exp(0{\cdot}5\Delta\text{BIC}) = \exp(0{\cdot}5(X^2 - \log(n)))$, which approximates a Bayes factor.

Along the X^2 continuum, lines of demarcation reveal the *views* of the metrics, regarding H_0 and H_A, given the same evidence. Note that, unlike AIC and FSTs, Good's compromise and BIC demand more evidence against H_0 for rejection as sample size increases. This causes the BIC line of demarcation to surpass those of AIC and FSTs at $n = 8$ and $n = 47$, respectively, and the line of demarcation for Good's compromise to exceed those of AIC and FSTs at $n = 11$ and $n = 101$, respectively (Fig. 1). Thus, as n increases, demarcation lines lower on the Y-axis represent methods whose focus is *refutation of the null hypothesis*, and lines higher on the Y-axis represent methods whose intent is *confirmation of the true hypothesis*. As a result, we define FSTs and particularly AIC as exploratory and refutative, relative to Good's compromise and BIC, which are progressively confirmatory.

BIC AND GOOD'S COMPROMISE ARE CONSISTENT

Confirmatory methods will often have the property of strong consistency. As noted earlier (Box 1), this requires:

$$\lim_{n\to\infty} Pr(\mathrm{H_0\ rejected}|\mathrm{H_A\ true}) = 1, \text{ and}$$
$$\lim_{n\to\infty} Pr(\mathrm{H_0\ rejected}|\mathrm{H_0\ true}) = 0. \qquad \text{eqn 10}$$

Thus, a method with strong consistency will always choose the correct hypothesis as sample size approaches infinity. Our heuristic demonstrates that BIC has this characteristic.

If H_A is true then the mean of X^2 grows with n at a much faster rate than the BIC line of demarcation, whereas the standard deviation grows at a slower rate than the mean (eqns 8 and 9). Thus, as sample sizes grow extremely large, the probability of an X^2 outcome being below the BIC line of demarcation goes to zero. An assumption of normality is not necessary here – only Chebyshev's inequality (Bienaymé 1853), which requires at least 75% of any distribution to be within two standard deviations of its mean. Conversely, if H_0 is true, the probability that X^2 falls above $\log(n)$ goes to zero as $n \to \infty$ (Figs 1 and 2). For a formal proof of strong consistency under our heuristic, see Appendix S1 §6.

As suggested by its name, Good's compromise is intermediate between BIC and frequentist significance testing with respect to hypothesis refutation/confirmation (Fig. 1). Nonetheless, the line of demarcation for Good's method increases with n, but more slowly than $\sqrt{n}$, insuring that it also has strong consistency (and asymptotic confirmation) for the true model or hypothesis (see Appendix S1 §7·3).

Whereas our paper is the first to demonstrate the strong consistency of Good's compromise, the strong consistency of BIC (and Bayes factors) is well known (e.g. Chen & Chen 2008; Casella *et al.* 2009). The proofs for this property, however, have relied on mathematics inaccessible to most ecologists. As a result, considerations of the strong consistency of BIC have remained (until now) terse in ecological publications (e.g. Hooten & Hobbs 2014, p. 15).

AIC AND SIGNIFICANCE TESTING ARE *NOT* CONSISTENT

The principle of strong (and weak) consistency does not hold for either AIC or frequentist significance testing. For FSTs:

$$Pr(\mathrm{H_0\ rejected}|\mathrm{H_0\ true}) = Pr\left(X^2 > b|\mathrm{H_0\ true}\right) = \alpha_b > 0,$$

where b is the demarcation value and α_b is the associated significance level. This condition is not dependent on sample size. Thus,

$$\lim_{n\to\infty} Pr(\mathrm{H_0\ rejected}|\mathrm{H_0\ true}) = \alpha_b > 0,$$

which violates consistency.

AIC does not consider n, and thus, its probability of type I error will also remain fixed at some nonzero value, which violates consistency (for our application, this is $Pr(\chi^2_1 \geq 2) \approx 0{\cdot}16$). Of course, as noted earlier, AIC was designed to be efficient, not consistent.

We emphasize that AIC and significance testing are both consistent when H_A is true. That is, both will correctly choose H_A over H_0, with probability 1 as n approaches infinity. Indeed, both approaches will have smaller type II error rates than BIC. Both methods, however, are biased against H_0 (Schervish 1996; Sellke, Bayarri & Berger 2001, p. 71) and by definition must incorrectly reject H_0 with some positive probability (e.g. α) when H_0 is true. Notably, this property is in conflict with a common pedagogic presentation of FSTs as a criminal trial in which the burden of proof is said to be on H_A (guilty verdict), not H_0 (innocent verdict) (e.g. Trosset 2009, p. 207).

COMPLEX MODELS

We can extend our heuristic to selection among models that differ widely with respect to numbers of parameters, reflecting a more conventional usage of AIC and BIC by ecologists. Under the assumptions given previously, consider a situation in which a null model, H_0, and a more complex alternative model, H_A, differ with respect to the presence or absence of d parameters. The demarcation surfaces of BIC and AIC are now $X^2 = \log(n)d$ and $X^2 = 2d$, respectively. The FST plane is $X^2 = \chi^2_{(d,1-\alpha)}$, where $\chi^2_{(d,p)}$ is the chi-squared inverse CDF with d degrees of freedom at probability P, and the demarcation surface for Goods compromise is $X^2 = \chi^2_{(d,1-\alpha_{\mathrm{adj}})}$. Figure 3 summarizes these ideas by depicting a diagonal slice through a three-dimensional space that defines X^2 as a function of all possible combinations of n and d (see Fig. S3-2 for a complete overview).

Earlier we emphasized the shared non-consistency of FSTs and AIC and revealed their resulting similarities for comparisons of nested models differing by one parameter (Figs 1 and 2). The monotonic relatedness of P-values and ΔAIC in this context has been noted previously by a large number of authors (e.g. Burnham & Anderson 1998; Murtaugh 2014; Brewer, Butler & Cooksley 2016). Figure 3, however, reveals that FST and AIC demarcation bounds may diverge substantially when comparing models with widely varying complexities.

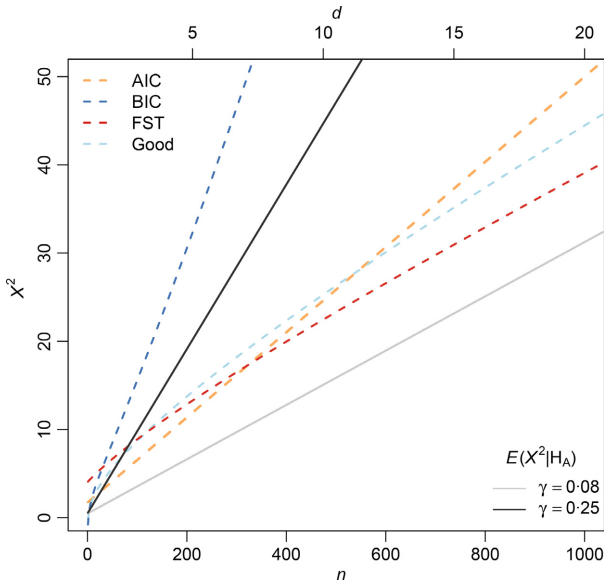

Fig. 3. Demarcation for Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), frequentist significance testing and Good's compromise as a function of particular combinations of increasing sample size, $n = 1, \ldots, 1000$, and differences in the number of parameters in null and alternative models, $d = 1, \ldots, 20$. A more complex depiction of X^2 as a function of *all possible* combinations of n and d is shown in Appendix S3.

A likelihood ratio FST considers the differing number of parameters in H_A compared with H_0 by altering the degrees of freedom in the χ^2_d null distribution. AIC and other ICs, however, address the magnitude of d explicitly through the use of penalty terms. Because of this difference, the demarcation plane of AIC increases rapidly with d, causing the AIC surface to exceed the FST surface at $d = 8$, and Good's surface when $d > 0{\cdot}5\chi^2_{(d,1-\alpha_{adj})}$ (Fig. 3). In application, this property will cause FST model selection (e.g. pick the smallest significant model at $\alpha = 0{\cdot}05$) to tend to choose more complex optimal approximating models compared with AIC.

A final difference between AIC and FST model selection is *not* evident in Fig. 3. This is that, unlike FSTs, ICs do not require that the H_0 model be parametrically nested in H_A, or even that these models have the same assumed error distributions. The only requirement is that the same response variable be used in all models under consideration. Akaike (1981), Burnham & Anderson (1998, p. 36, 133), Aho, Derryberry & Peterson (2014) and Fox (2015, p. 608) consider additional issues with model selection using FSTs, including logical problems, and simultaneous inference. While Good's approach remains ostensibly a compromise between BIC and conventional FSTs (Fig. 3), this approach is also ill-suited for generalized model selection for the same reasons that FSTs are ill-suited.

Because the BIC demarcation surface increases with both n and d, BIC has, except for extremely small sample sizes, lower type I error rates than AIC, FSTs, or Good's compromise. Thus, when a set of models under consideration includes the true model, and n is sufficiently large, BIC will select the true model and AIC and FST model selection (which emphasize sensitivity over specificity) will tend to choose models that are more complex than the true model. The *purpose* of AIC, however, is not asymptotic consistency, but identifying useful predictive models. This means that BIC may select underfit models, compared with AIC, because the general scenario $\log(n) > 2$ requires greater penalization from BIC. As a result, an investigator must choose a model selection 'worldview' (see Aho, Derryberry & Peterson 2014). If the goal is model confirmation, suggesting that the set of proposed models includes a specific well-justified model, then BIC is the appropriate criterion. On the other hand, if the goal is model exploration and practical generalization, reflecting a situation wherein 'all models are false or insufficient, but some are useful', then AIC is the appropriate choice.

Discussion

Our approach provides a graphical depiction of strong consistency and allows comparisons of FSTs, Good's compromise, AIC and BIC along a continuum of refutation/confirmation. The relative positioning of demarcation criteria in Figs 1–3 clarifies both the purpose and correct usage of these inferential methods.

Interestingly, our heuristic also provides insight into the general epistemology and ontogeny of model selection and hypothesis testing approaches. In conventional usage, Popper (1934) would not view H_0 as embodying a 'bold hypothesis'. Nonetheless, we feel that refutatory methods like FSTs acknowledge the tenets of severe falsificationism, and thus suggest the influence of the hypothetico-deductive archetype on R. A. Fisher (see Quinn & Keough 2002, p. 32). On the other hand, consistent methods are to varying degrees confirmatory and thus address the practical concerns of philosophers of science in the context of well-justified theories (Lakatos 1978), or for inferences that encompass a phenomenon that is observed (essentially) in its entirety ('empirical laws' *sensu* Carnap 1966, Ch. 23).

LIMITATIONS AND FUTURE WORK

We emphasize that our presentation is intentionally simplistic. For example, to clarify differences in hypothesis and model selection approaches, our graphs define strict standards distinguishing support for H_0 or H_A. In practice, however, we recommend *against* this potentially thoughtless approach. Following the advice of Fisher (1956, pp. 41–42, 80, 100) and (most?) modern statisticians, we recommend that X^2 (and, if applicable, P-values) be viewed as a continuous measure of evidence with respect to decision rule benchmarks, particularly when those conditions are allowed to vary rationally with power, sample size and/or parameter number.

Our approach also has conventional statistical assumptions, including the independence of outcomes. The performance of ICs in violation of these constraints can be considered with simulation (Brewer, Butler & Cooksley 2016). In future work, we intend to consider the generalities and limitations of our comparative heuristic by employing this approach.

Practical confirmational inferences

Our work also can be criticized on the grounds that consideration of strong consistency is necessarily asymptotic and that strong consistency may not insure useful confirmational inferences for small or realistic sample sizes. This limitation is illustrated by considering the refutation/confirmation perspectives of other ICs using their penalty terms (Fig. 4).

With respect to efficient and non-consistent methods, AICc (Sugiura 1978), which has the form: AIC + $(2p(p + 1)/(n - p - 1))$, converges to AIC as n becomes large. AICc, however, applies a larger penalty term than AIC for smaller sample sizes, making it *more confirmatory* than AIC, but (generally) *less confirmatory* than Good's compromise or BIC.

The 'corrected AIC' (CAIC) of Bozdogan (1987), a strongly consistent but non-efficient method, has the demarcation surface $d\log(n + 1)$, making it slightly *more confirmatory* than BIC (Fig. 4). Conversely, the Hannan–Quinn information criterion (HQC, Hannan & Quinn 1979), another strongly consistent but non-efficient method, uses the surface $2d\log(\log(n))$. Because of its double log transformation of n, HQC is *less confirmatory* than BIC and (given $d = 1$) Good's method, or even significance testing for $n < 922$ (Fig. 4). *Still less confirmatory* is the strongly consistent and non-efficient 'adjusted-BIC' (Rissanen 1978; Sclove 1987), with the surface $d\log((n + 2)/24)$.

Obviously, while non-consistent methods are inappropriate for confirmatory inferences, consistent methods (dashed lines in Fig. 4) may vary greatly with respect to their degree of refutation/confirmation for hypotheses, given $n \ll \infty$. Thus, consistency alone does not insure that a method will provide useful confirmational inferences (see Claeskens & Hjort 2008, p. 113).

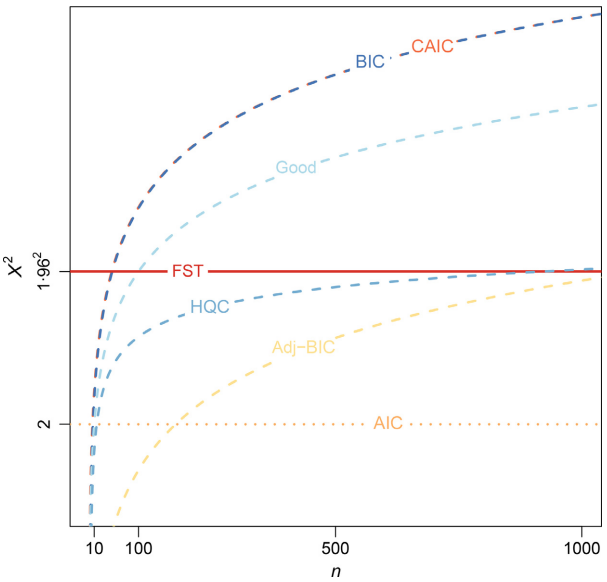

Fig. 4. Demarcation for H_0 and H_A models that differ with respect to the inclusion/exclusion of a single parameter. Dotted lines indicate asymptotically efficient methods, dashed lines indicate methods with strong consistency, and solid lines indicate methods which are neither efficient nor strongly consistent.

Confirmation of a hypothesis is a much more difficult matter than refutation (Popper 1934). This is because refutation can occur as the result of *one* observation, whereas confirmation requires *all possible* observations (*sensu* 'the problem of induction' §4, Hume 1748; 'black swan hypothesis', Taleb 2007). It follows that probabilistic confirmational *inferences* should require much larger sample sizes than those required for epistemically equivalent refutative inferences. Given a moderately large sample size, for example $n = 100$, both CAIC and BIC begin a rapid divergence from FSTs, and Good's compromise is equivalent to FSTs; however, HQC and adjusted-BIC will perform, for small d values, as strongly refutative methods, with high levels of type I error. At $n = 1000$, CAIC, BIC and Good's compromise are all distinctly confirmational compared with FSTs; however, HQC and adjusted-BIC remain ambiguous with respect to refutation/confirmation.

RECOMMENDATIONS

AIC, BIC, FSTs and Good's compromise address a fundamental scientific concern: quantifying the strength of empirical evidence to aid in choosing among models and hypotheses. As a result, all four approaches constitute potential catalysts for scientific progress. Like all tools, however, the methods may have limited usefulness when applied incorrectly or to inappropriate tasks.

Hypothesis testing

We have defined FSTs to be strongly refutative for H_0. As summarized by the inventor of both *P*-values and H_0:

> '. . .the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance to disprove the null hypothesis'.
>
> R. A. Fisher (1935, p. 18).

However, ecologists seldom believe that a null hypothesis representing 'zero effect' is literally true (Barber & Ogle 2014), and if this belief is correct, then Fig. 2 shows that a researcher need only sufficiently increase sample size to reject H_0 with probability (power) = 1. This outcome has been the basis for many criticisms of FSTs, on the grounds that (i) H_0 will always equate to a straw man hypothesis that must give way under increasing sample size, and (ii) a correct significant result can be the by-product of a trivial (biologically meaningless) effect size (Oakes 1986; Royall 1997; Johnson 1999; Trafimow & Marks 2015; Wasserstein & Lazar 2016). This issue become more striking when considering the negligible probability that true effect sizes *can ever be* 'exactly zero' for ecological systems because of the potential for confounded effects and coincidental associations (*sensu* 'ambient extraneous correlations', Lykken 1968; 'crud factor' Meehl 1990).

Related complications arise in the less frequent converse case that FSTs are applied to a well-justified point hypothesis given as H_0. Many proposed theoretical frameworks in ecology

(Appendix S2), including neutral models (Gotelli & Graves 1996), effectively describe natural systems, and investigators have generally quantified the validity of said theories by setting them as null hypotheses in FSTs (Appendix S2, Case study 1). This is done because an explicit fixed effect (generally 0) can be set for H_0, whereas H_A can only define 'some effect' distinct from H_0 (Quinn & Keough 2002; Aho 2013, Ch. 6). FSTs, however, do not allow empirical confirmation of null hypotheses. Instead, under Fisher's (1935) refutatory framework, we 'reject' or provisionally 'fail to reject' H_0.

The non-confirmatory character of FSTs becomes particularly relevant in the context of extremely large and high-dimensional modern data sets and simulation studies (White *et al*. 2014). Such formats have become increasingly prevalent, particularly in molecular and spatial ecology, prompting an explosion of new methods for data management (Altschul *et al*. 1990; Huson *et al*. 2007; Song *et al*. 2012; Gong, Geng & Chen 2015) and data analysis (Benjamini & Hochberg 1995; Benjamini & Yekutieli 2001; Yoo, Ramirez & Liuzzi 2014; Gandomi & Haider 2015). As models become complex or sample sizes become large, FSTs will reject a strongly truth-directed null hypothesis with high probability, whereas BIC, because of its emphasis on minimizing type I error, may suggest retention of H_0 (Fig. 3; Appendix S2, Case study 2).

On the other hand, *P*-values may provide a valuable exploratory tool for establishing 'some effect' and quantifying empirical departures from H_0 (e.g. Murtaugh 2014; Stanton-Geddes, Gomes De Freitas & De Sales Dambros 2014). This is particularly true when *P*-values and FSTs are conjoined with a method like Good's compromise that accounts for the effect of *n*. In a paper titled 'The Common Sense of *P*-values', de Valpine (2014) reaffirms our characterization of *P*-values as simultaneously falsificationist, exploratory, and widely useful. Specifically, the author argues that: '. . .the purpose of *P*-values is to convince a skeptic that a pattern in data is real', and 'When there is a scientific need for skeptical reasoning with noisy data, the logic of *P*-values is inevitable'. Nonetheless, FSTs are inappropriate for confirmatory testing, and because of a number of issues, including the impossibility of comparing non-nested models – or comparing models with differing error distributions – FSTs (and Good's compromise) are unsuitable for general model selection.

For confirmatory hypothesis testing of point hypotheses, we recommend using *non-ambiguously* confirmatory approaches with strong consistency like BIC. Other useful confirmational methods include Bayes Factors (Kass & Raftery 1995), which ignore priors, and Bayesian posterior probabilities (Jeffreys 1961), which consider priors, thus allowing 'coherent' assessments of complex hypotheses (Lavine & Schervish 1999). Non-Bayesian likelihood-based methods for fixed-response data (e.g. log-likelihood ratios) and non-consistent IC approaches (e.g. ΔAIC and Akaike weights) are ostensibly *n*-insensitive. These approaches, however, were not designed to be confirmatory, and furthermore are generally applied in the context of *n*-invariant decision standards (e.g. $\alpha = 0{\cdot}05$ and ΔAIC = 10) that ignore, among other things, the required increase in $E(X^2|H_A)$ with *n*.

Model selection

Our graphs demonstrate that model selection techniques represent trade-offs with respect to sensitivity (i.e. *Power* = 1 − *Pr*(type II error)) and specificity (1 − *Pr*(type I error)). Attention to type I error is purportedly emphasized over attention to type II error in classic hypothesis testing – although BIC generally has far lower rates of type I error than FSTs – because the former constitutes an incorrect statement, while the latter is merely a 'failure to reject' (Dziak *et al*. 2012). De-emphasis of type II error, however, may result in underfitting and loss of predictive power. Thus, one must choose a worldview. If, as in significance testing, the aim is non-confirmatory and we merely wish to identify – from a set of imperfect and potentially non-nested models – a useful predictive model, then we should use an efficient method like AIC. Conversely, when comparing and assessing models and hypotheses representing well-justified theories, we should rely on methods, such as BIC, that are both statistically consistent and confirmatory.

Acknowledgements

We thank Colden Baxter and Terry Bowyer (Idaho State University), Perry de Valpine (University of California, Berkeley), associate editor Bob O' Hara and three anonymous MEE reviewers for their useful comments on this manuscript.

Data accessibility

Data sets examined in the supplemental materials (Appendix S2) have been archived as the dataframes `savage` and `simberloff`, in the `R`-package *asbio* (Aho 2016, https://cran.r-project.org/web/packages/asbio/index.html).

References

Aho, K. (2013) *Foundational and Applied Statistics for Biologists using R*. CRC Press, Boca Raton, FL.

Aho, K. (2016) *asbio: A Collection of Statistical Tools for Biologists*. R package version 1.3-4. Available at: https://cran.r-project.org/web/packages/asbio/index.html

Aho, K., Derryberry, D. & Peterson, T. (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, **95**, 631–636.

Akaike, H. (1969) Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, **21**, 243–247.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory* (eds B. Petrov & F. Csáki), pp. 267–281. Tsahkadsor, Armenia, USSR, September 2–8, Akadémiai Kiadó, Budapest, Hungary.

Akaike, H. (1981) Modern development of statistical methods. *Trends and Progress in System Identification* (ed. P. Eykhoff), pp. 165–189. Pergamon, Paris, France.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Barber, J.J. & Ogle, K. (2014) To P or not to P? *Ecology*, **95**, 621–626.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 289–300.

Benjamini, Y. & Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165–1188.

Bienaymé, I.J. (1853) Considérations àl'appui de la découverte de laplace. *Comptes Rendus de L'Académie des Sciences*, **37**, 309–324.
Bozdogan, H. (1987) Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.
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**, 679–692.
Burnham, K.P. & Anderson, D.R. (1998) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, NY.
Carnap, R. (1966) *Philosophical Foundations of Physics*. Basic Books, New York, NY.
Casella, G., Girón, F.J., Martínez, M.L. & Moreno, E. (2009) Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, **37**, 1207–1228.
Chang, S.H., Cosman, P. & Milstein, L. (2011) Chernoff-type bounds for the Gaussian error function. *IEEE Transactions on Communications*, **59**, 2939–2944.
Chen, J. & Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759–771.
Chiani, M., Dardari, D. & Simon, M.K. (2003) New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Transactions on Wireless Communications*, **2**, 840–845.
Chun, S.Y. & Shapiro, A. (2009) Normal versus noncentral chi-square asymptotics of misspecified models. *Multivariate Behavioral Research*, **44**, 803–827.
Claeskens, G. & Hjort, N.L. (2008) *Model Selection and Model Averaging*. Cambridge University Press, Cambridge, UK.
Dziak, J.J., Coffman, D., Lanza, S. & Runze, L. (2012) Sensitivity and specificity of information criteria. *The Pennsylvania State University Technical Report Series*, **12–119**, 1–30.
Fisher, R.A. (1935) *The Design of Experiments*. Oliver and Boyd, Edinburgh, UK.
Fisher, R.A. (1956) *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, UK.
Foster, D.P. & George, E.I. (1994) The risk inflation criterion for multiple regression. *The Annals of Statistics*, **22**, 1947–1975.
Fox, J. (2015) *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, Thousand Oaks, CA, USA.
Gandomi, A. & Haider, M. (2015) Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management*, **35**, 137–144.
Gong, J., Geng, J. & Chen, Z. (2015) Real time GIS data model and sensor web service platform for environmental data management. *International Journal of Health Demographics*, **14**, 2.
Good, I. (1992) The bayes/non-bayes compromise: a brief review. *Journal of the American Statistical Association*, **87**, 597–606.
Gotelli, N.J. & Graves, G.R. (1996) *Null Models in Ecology*. Smithsonian Institution Press, Washington, DC.
Hannan, E. & Quinn, B. (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, **41**, 190–195.
van der Hoeven, N. (2005) The probability to select the correct model using likelihood-ratio based criteria in choosing between two nested models of which the more extended one is true. *Journal of Statistical Planning and Inference*, **135**, 477–486.
Hooten, M.B. & Hobbs, N.T. (2014) A guide to Bayesian model selection for ecologists. *Ecological Monographs*, **85**, 3–28.
Hume, D. (1748) *Philosophical Essays Concerning Human Understanding*. A. Millar, London, UK.
Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Research*, **17**, 377–386.
Jeffreys, H. (1961) *Theory of Probability*, 3rd edn. Oxford University Press, Oxford, UK.
Johnson, D.H. (1999) The insignificance of statistical significance testing. *The Journal of Wildlife Management*, **63**, 763–772.
Johnson, V.E. (2008) Properties of Bayes factors based on test statistics. *Scandinavian Journal of Statistics*, **35**, 354–368.
Kass, R.E. & Raftery, A.E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.
Kullback, S. & Leibler, R.A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.
Lakatos, I. (1978) *The Methodology of Scientific Research Programmes*. Cambridge University Press, Cambridge, UK.
Lavine, M. & Schervish, M.J. (1999) Bayes factors: what they are and what they are not. *The American Statistician*, **53**, 119–122.
Lykken, D. (1968) Statistical significance in psychological research. *Psychological Bulletin*, **70**, 151–159.
Mallows, C.L. (1973) Some comments on CP. *Technometrics*, **15**, 661–675.
Meehl, P.E. (1990) Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, **66**, 195–244.
Murtaugh, P.A. (2014) In defense of *P*-values. *Ecology*, **95**, 611–617.
Oakes, M. (1986) *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. Wiley, Chichester, UK.
Popper, K. (1934) *Logik der Forschung*. Mohr Siebeck, Tubingen, Germany.
Pötscher, B.M. (1991) Effects of model selection on inference. *Econometric Theory*, **7**, 163–185.
Quinn, G.P. & Keough, M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK.
Raftery, A.E. (1995) Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.
Rissanen, J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465–471.
Royall, R. (1997) *Statistical Evidence: A Likelihood Paradigm*. CRC Press, Boca Raton, FL, USA.
Schervish, M.J. (1996) P values: what they are and what they are not. *American Statistician*, **50**, 203–206.
Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.
Sclove, S.L. (1987) Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, **52**, 333–343.
Sellke, T., Bayarri, M.J. & Berger, J.O. (2001) Calibration of p values for testing precise null hypotheses. *American Statistician*, **55**, 62–71.
Shapiro, A. (2009) Asymptotic normality of test statistics under alternative hypotheses. *Journal of Multivariate Analysis*, **100**, 936–945.
Söderström, T. (1977) On model structure testing in system identification. *International Journal of Control*, **26**, 1–18.
Song, B., Su, X., Xu, J. & Ning, K. (2012) MetaSee: an interactive and extendable visualization toolbox for metagenomic sample analysis and comparison. *PLoS One*, **7**, e48998.
Stanton-Geddes, J., Gomes De Freitas, C. & De Sales Dambros, C. (2014) In defense of *p* values: comment on the statistical methods actually used by ecologists. *Ecology*, **95**, 637–642.
Stoica, P., Selén, Y. & Li, J. (2004) On information criteria and the generalized likelihood ratio test of model order selection. *IEEE Signal Processing Letters*, **11**, 794–797.
Sugiura, N. (1969) Asymptotic expansions of the distributions of the likelihood ratio criteria for covariance matrix. *The Annals of Mathematical Statistics*, **40**, 2051–2063.
Sugiura, N. (1978) Further analysis of the data by Akaike' s information criterion and the finite corrections. *Communications in Statistics – Theory and Methods*, **7**, 13–26.
Taleb, N.N. (2007) *The Black Swan: The Impact of the Highly Improbable*. Random House, New York, NY, USA.
Teräsvirta, T. & Mellin, I. (1986) Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics*, **13**, 159–171.
Trafimow, D. & Marks, M. (2015) Editorial. *Basic and Applied Social Psychology*, **37**, 1–2.
Trosset, M.W. (2009) *An Introduction to Statistical Inference and Its Applications with R*. CRC Press, Boca Raton, FL, USA.
de Valpine, P. (2014) The common sense of *P*-values. *Ecology*, **95**, 617–621.
Wasserstein, D.L. & Lazar, N.A. (2016) The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, **70**, 129–133.
Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.
White, J.W., Rassweiler, A., Samhouri, J.F., Stier, A.C. & White, C. (2014) Ecologists should not use statistical significance tests to interpret simulation model results. *Oikos*, **123**, 385–388.
Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, **9**, 60–62.
Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, **92**, 937–950.
Yoo, C., Ramirez, L. & Liuzzi, J. (2014) Big data analysis using modern statistical and machine learning methods in medicine. *International Neurourology Journal*, **18**, 50–57.
Zhang, Y. & Shen, X. (2010) Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining*, **3**, 350–358.

Received 9 May 2016; accepted 19 August 2016
Handling Editor: Robert B. O'Hara

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Mathematical derivations including, (1) definitions for ICs based on the normal likelihood function under H_0 and H_A, (2) demarcation lines justifications, (3) distributions of X^2 under H_0 and H_A, (4) a heuristic-based proof of the strong consistency of BIC, (5) bounds for the Good's compromise, (6) proof of the strong consistency of Good's Bayes/non-Bayes compromise.

Appendix S2. An abridged table of theories which effectively describe at least some ecological systems, and a small sample of case studies in which ecologists have (inappropriately) used significance testing for the purpose of confirmation.

Appendix S3. Additional figures.