Ecology, 102(10), 2021, e03475

Parsimonious model selection using information theory: a modified selection rule

Luke A. Yates,[1] Shane A. Richards, and Barry W. Brook

School of Natural Sciences, University of Tasmania, Hobart, Tasmania 7005 Australia

Citation: Yates, L. A., S. A. Richards, and B. W. Brook. 2021. Parsimonious model selection using information theory: a modified selection rule. Ecology 102(10):e03475. 10.1002/ecy.3475

Abstract. Information-theoretic approaches to model selection, such as Akaike's information criterion (AIC) and cross validation, provide a rigorous framework to select among candidate hypotheses in ecology, yet the persistent concern of overfitting undermines the interpretation of inferred processes. A common misconception is that overfitting is due to the choice of criterion or model score, despite research demonstrating that selection uncertainty associated with score estimation is the predominant influence. Here we introduce a novel selection rule that identifies a parsimonious model by directly accounting for estimation uncertainty, while still retaining an information-theoretic interpretation. The new rule, which is a modification of the existing one-standard-error rule, mitigates overfitting and reduces the likelihood that spurious effects will be included in the selected model, thereby improving its inferential properties. We present the rule and illustrative examples in the context of maximum-likelihood estimation and Kullback-Leibler discrepancy, although the rule is applicable in a more general setting, including Bayesian model selection and other types of discrepancy.

Key words: cross validation; information theory; model selection; overfitting; parsimony; post-selection inference.

Introduction

A typical goal in eco-evolutionary research is to determine the type and functional form of the mechanisms by which a biological response (e.g., fitness, growth, density, competition, etc.) is invoked, as well as to quantify the relative importance of "predictors" (Southwood 1978, Bolker 2008). This can be done, for instance, by experimentation (e.g., before-after-control-impact designs), or via observation of pattern (using correlative models). In terms of analysis, these problems have historically been tackled using null-hypothesis significance testing, but there is a growing body of literature on the theory and implementation of alternative approaches that avoid an a priori bias toward the null being true, including model selection and model averaging, regularization methods such as ridge regression or the lasso, and machine learning (Burnham and Anderson 2002, Hastie et al. 2009). If the goal of the analysis is prediction, then model averaging, regularization, or machine learning will generally perform best (Hooten et al. 2015), but see Richards (2005) and Richards et al. (2011). In ecology, however, the goal is often explanation, which is why model selection, with the underlying aim of selecting a single parsimonious model from an initial set of candidates, is a popular approach. However, there are challenges with using model selection to choose a single "best" model for inference; these include deciding which score or criterion to use, concern for overfitting or underfitting, accounting for model selection uncertainty, and applying the principle of parsimony.

In the usual approach to model selection, a score is estimated for each model belonging to an initial set of candidates. Candidate models are generated a priori, ideally on the basis of carefully considered hypotheses; the selected best model, subject to validation checks, is the model with the lowest estimated score. The most frequently used estimation indices in ecology are information criteria, such as the Akaike information criterion (AIC) and its common variants. These criteria are estimates of the information-theoretic quantity (relative, expected) Kullback-Leibler discrepancy, providing a theoretically rigorous foundation for the selection framework (Akaike 1973). Cross validation can also be used to estimate Kullback-Leibler discrepancy, but has been historically less often used in ecology, perhaps due to concerns about computational cost (Hooten et al. 2015). However, its application is rapidly increasing, especially in movement ecology, habitat-selection studies, and species distribution modeling (Roberts et al. 2017, Valavi et al. 2019). The Bayesian information criterion (BIC) is an often-used alternative, which, despite its name, is not information-theoretic. It is sometimes favored because it tends to select a simpler model than information-theoretic approaches.

Consistent selectors, such as BIC, are defined by their asymptotic property of selecting the data-generating

Manuscript received 13 October 2020; revised 16 February 2021; accepted 13 May 2021. Corresponding Editor: Caz M. Taylor.
[1] E-mail: Luke.Yates@utas.edu.au

model, given that this "true" model is in the model set. Relative to estimates of Kullback-Leibler discrepancy, consistent selectors penalize increasing complexity, thus offering the allure of a more parsimonious selection. However, in a setting where the true model is not one of the candidates, called an *M*-open setting, the use of information criteria or cross-validated scores is more appropriate (Vehtari and Ojanen 2012). Consistent selectors are suitable for *M*-closed settings, where the true model is a candidate; however, the vast majority of ecological problems are almost certainly *M*-open. Although it is tempting to use BIC to mitigate potential overfitting, consistent selectors are liable to underfit in an *M*-open setting, which is arguably a more serious issue than overfitting, due to increased bias, poor predictive performance and the potential failure to interpret important effects (Burnham and Anderson 2002).

It is not predominantly the choice of score that causes overfitting, providing the score is an unbiased estimate of Kullback-Leibler discrepancy, but instead a failure to account for the uncertainty associated with score estimation. We never have the true model scores, only estimates (Richards 2005), yet the usual practice of ignoring estimation uncertainty, by selecting the lowest-scoring model, leads to overfitting due to random chance alone (Piironen and Vehtari 2017); see Fig. 1. The probability of overfitting is increased when the number of models is high, and the (relative) uncertainty of score estimates is large with respect to score differences, especially when several models have similar performance to the lowest-scoring model.

In this paper, we introduce a novel selection rule that identifies a parsimonious model by directly accounting for estimation uncertainty in the chosen score. It is useful in ecological applications because it selects a model that has predictive power close to the Kullback-Leibler best model, yet favors simpler, thus more easily interpreted, models. Our rule, which is a modification of the ordinary one-standard-error rule (Breiman et al. 1984), uses correlation-adjusted standard errors to calibrate selection, in this way mitigating overfitting by accounting for uncertainty, while still retaining an information-theoretic interpretation using model scores such as AIC or cross validation. We use two case studies to illustrate the general applicability and functioning of the approach. We discuss how our method compares to other techniques for parsimonious model selection, including Bayesian reference methods and other calibration approaches. We also discuss the important (but usually ignored) issue of valid inference after model selection. Valid inference is a distinct problem from that of mitigating overfitting, or deciding whether to use a single model or model averaging as the basis for inference.

Methods

Estimating model scores

In the information-theoretic approach to model selection, models are scored using expected, relative, Kullback-Leibler discrepancy, a measure of the information lost when an alternative model is used in place of the (usually unknown) true model. Viewed as a predictive score, the discrepancy is the expected out-of-sample log-likelihood, or log predictive density, which can be estimated by adding a bias correction to the within-sample log-likelihood. Estimates of this type, usually multiplied by −2 to place them on the deviance scale, are called information criteria, the most well-known of which is AIC, given by

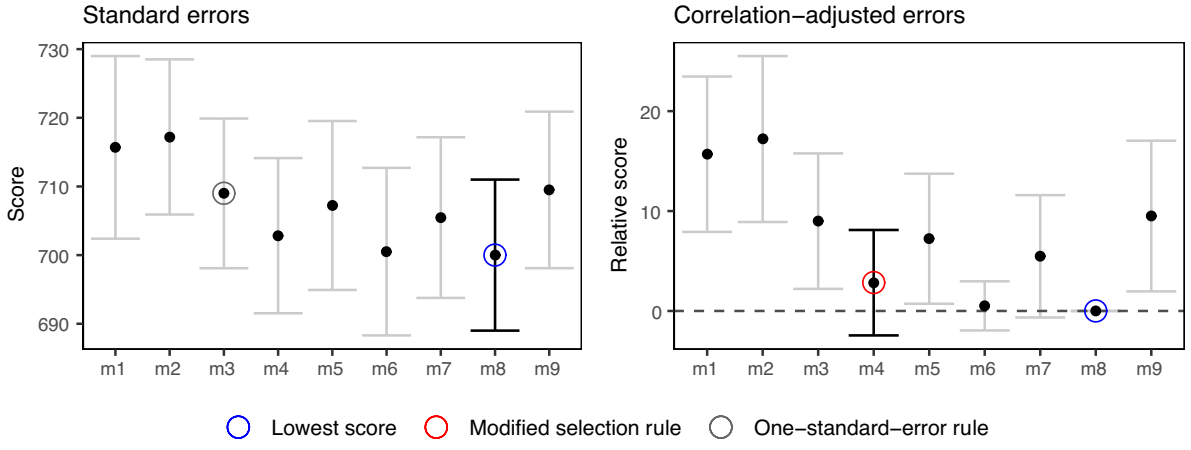

Fig. 1. Hypothetical model scores. The labels on the *x*-axis index candidate models by increasing complexity and the lowest score estimate is attributed to model m8. The left plot depicts score estimates (circles) and one standard error of the estimate (error bars). In this example, high uncertainty of the estimates makes it likely that model m8 is overfit. The ordinary one-standard-error rule selects model m3, the simplest model whose score lies within one standard error of the lowest-scoring model. However, this is likely underfit, since the score estimates are usually highly correlated. The modified rule uses correlation-adjusted standard errors (right plot), selecting model m4, the simplest model whose score estimate lies within one correlation-adjusted standard error of the lowest estimated score.

$$\mathrm{AIC} = -2\ell + 2p \tag{1}$$

where ℓ is the maximum log-likelihood of all data used for fitting and p, the number of estimated model parameters, is the bias-correction term (Akaike 1973). Other commonly used information criteria include $\mathrm{AIC_c}$, for small-sample-bias correction, and QAIC for over-dispersed data.

Cross validation provides a more direct means to estimate Kullback-Leibler discrepancy. In K-fold cross validation, out-of-sample prediction is simulated by splitting the data into K equal-sized folds (parts), generating K data subsets by omitting one fold at a time from the full data set. The model is fit to each subset in turn, providing an out-of-sample prediction for each data point. The corresponding cross-validation estimate of the score is

$$\mathrm{CV} = \sum_{i=1}^{n} -2\ell_i^* \tag{2}$$

where ℓ_i^* is the out-of-sample (predictive) log-likelihood of the ith data point, and n is the total number of data points.

To preserve the information-theoretic interpretation of the score, it is necessary to use unbiased (or constant-biased) estimators of Kullback-Leibler discrepancy. Information criteria are fast to compute, typically requiring a single model fit, however, the unbiasedness of the estimate imposes restrictions on the models and the data; for example, AIC requires large n and $\mathrm{AIC_c}$ requires models to be linear with homoscedastic errors (Hurvich and Tsai 1989). Cross validation is far more flexible, requiring only that the data points be conditionally independent (Roberts et al. 2017). For $K < n$, K-fold cross validation does introduce bias, however it is easily corrected using the method of Burman (1989). The special case $K = n$, called leave-one-out cross validation, has minimal bias, obviating the need for correction in most instances; however, it can be computationally expensive.

Estimating standard error

When a score estimate S is an information criterion (e.g., AIC), the standard error of the estimate depends only on the log-likelihood, since the bias correction (e.g., $-2p$) is fixed for each model. There are two main ways to estimate the standard error of the score:

1) the pointwise sample estimate

$$\hat{\sigma}_S = \sqrt{\frac{n}{n-1}\sum_{i=1}^{n}\left(2\ell_i - 2\overline{\ell}\right)^2} \tag{3}$$

where $\overline{\ell} = \frac{1}{n}\sum \ell_i$;

2) alternatively, bootstrapped samples can be used to calculate the non-parametric bootstrap sample estimate

$$\hat{\sigma}_S = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(2\ell^b - 2\overline{\ell}^b\right)^2} \tag{4}$$

where $b = 1, \ldots, B$ indexes bootstrap samples of size n, each resampled with replacement from the full data set, and $\overline{\ell}^b = \frac{1}{B}\sum \ell^b$ with $\ell^b = \sum \ell_i^b$ (Efron and Tibshirani 1994).

For large n, the pointwise estimate may suffice under the assumption of asymptotic normality (it is certainty faster to compute), however the bootstrap offers a more versatile solution, where B can be made arbitrarily large to improve statistical properties, limited only by computation time but without making distributional assumptions (Efron and Tibshirani 1994). We use the bootstrap estimate (Eq. 4) for all analyses in this paper.

The bootstrap estimate (Eq. 4) may also be applied when S is a cross-validated estimate. The standard error for K-fold cross validation is usually approximated by replacing ℓ with ℓ^* in Eq. 3 (Breiman et al. 1984), although it is well-known to be biased as the ℓ_i^* are not independent; in fact, there exists no unbiased estimator for the standard error (Bengio and Grandvalet 2004). The cross-validation estimate (Eq. 2) can be written as $\mathrm{CV} = -2\ell + 2\kappa$, where $\kappa = \mathrm{CV}/2 + \ell$ is an adaptive bias-correction term, often called the effective number of parameters (Gelman et al. 2014). The use of Eq. 4 to estimate σ_{CV} is justified because the contribution of κ is generally an order of magnitude smaller than that of ℓ (Efron 2004: Remark B). This approximation enables flexible nonparametric estimation of the standard error by estimating κ just once, using the full data set, thus avoiding the computational cost of applying cross validation to each bootstrap sample.

A modified selection rule

The original one-standard-error rule, introduced in the context of classification and regression trees (CART; Breiman et al. 1984), uses the standard error of the lowest-scoring model (i.e., the estimated best-performing model) for the possible selection of a simpler model; the rule is illustrated in Fig. 1. This choice of threshold has a "natural" interpretation, since one standard error is the average (root-mean-square) distance of score estimates from the mean (for the best-scoring model), but it ignores the often substantial correlation of the errors between models, thus over-estimating the relative variation. However approximate, the rule performs well in the original CART setting, but does not generalize to model- and variable-selection problems, where it can substantially underfit, as we show in the examples of the following section. Despite a proclivity to underfit in selection problems, use of the original one-standard-error rule is often recommended (Hastie et al. 2009, James et al. 2013) and it is implemented in commonly used R packages such as bestglm and caret (Kuhn 2008, McLeod et al. 2020, R Core Team 2020).

To account for the correlation of estimates between models, the performance threshold must be evaluated

pairwise, where each candidate, together with the lowest-scoring model, characterizes a unique pair. Given that model selection is based on the set of differences of the estimated model scores, a natural starting point for a correlation-adjusted performance threshold is the variance of the relative estimate, $\Delta S_k = S_k - S_{\min}$, where k indexes the candidate models and $S_{\min} = \min\{S_k\}$ is the lowest-scoring model. Since ΔS_i is an estimate of the difference of two random variables, with the subscript '$k = \min$' determined by the initial score estimates and fixed thereafter, the variance is given by the identity

$$\sigma^2_{\Delta S_k} = \sigma^2_{\min} + \sigma^2_k - 2\rho_{\min,k}\sigma_{\min}\sigma_k \tag{5}$$

where ρ is the correlation coefficient.

The standard error $\sigma_{\Delta S_k}$ appears in the recent Bayesian literature, where it is estimated using Eq. 3, up to an overall factor of $\sqrt{n}$, with $\ell_{k,i}$ is replaced by $\ell_{k,i} - \ell_{\min,i}$, the pointwise difference of (posterior) expected log predictive densities (Vehtari et al. 2017, Piironen et al. 2020). The values $\ell_{k,i} - \ell_{\min,i}$, or indeed the bootstrap values $\ell^b_k - \ell^b_{\min}$, sample the null-hypothesis distribution that the kth model does not improve over the lowest-scoring model. To investigate the suitability of $\sigma_{\Delta S_k}$ as performance threshold for a modified one-standard-error rule, we introduce the following generalized definition of a correlation-adjusted error:

$$\sigma^2_{\text{adj},k} = \alpha\sigma^2_{\min} + \beta\sigma^2_k + \gamma\rho_{\min,k}\sigma_{\min}\sigma_k \tag{6}$$

where α, β, and γ are coefficients to be determined. The case $\alpha = \beta = 1$, $\gamma = -2$, corresponds to $\sigma^2_{\Delta S_k}$ (Eq. 5).

To constrain the coefficients in Eq. 6, we demand that the correlation-adjusted definition, when used as the performance threshold for a modified one-standard-error rule, selects, in the limit of zero correlation, the same model that would be selected if using the usual standard error, and that it selects the lowest-scoring model in the limit of unit correlation, that is

$$\begin{aligned} \sigma_{\text{adj},k} &\to \sigma_{\min} \quad \text{as} \quad \rho_{\min,k} \to 0 \\ \sigma_{\text{adj},k} &\to 0 \qquad\quad \text{as} \quad \rho_{\min,k} \to 1 \end{aligned} \tag{7}$$

The constraints imply $\gamma = -\sigma_{\min}/\sigma_k$ and $\alpha = 1 - \beta\sigma^2_k/\sigma^2_{\min}$, which after substitution into Eq. 6, determine the following definition of a correlation-adjusted standard error, conditional on the lowest-scoring model

$$\sigma_{\text{adj},k} \equiv \sigma_{\min}\sqrt{1 - \rho_{\min,k}} \tag{8}$$

The standard error $\sigma_{\min}$ can be estimated using either Eq. 3 or Eq. 4, and an estimate of the matrix element $\rho_{\min,k}$ is easily computed in R by applying the base function cor to a data frame containing either bootstrapped or pointwise values of -2ℓ, where each row corresponds to a sample value and each column to a model; see the supplementary materials for code examples.

Using the definition (Eq. 8), with the complexity of a model taken to be the (effective) number of estimated parameters, the modified selection rule is stated as follows:

> Select the least complex model whose score estimate lies within one correlation-adjusted standard error of the lowest score estimate. If two or more models satisfy this condition, that is they have the same model complexity, then either (1) select the lowest scoring of these models or (2) retain of all these models as a selected best set.

A standard-error plot, such as Fig. 1, shows the score estimate and correlation-adjusted standard error for each model, providing an easily interpreted summary of model performance and a clear illustration of the modified selection mechanism.

In many instances, we expect the modified rule to select a single model, either because only one model fulfils the initial selection condition, or because option (1) has been chosen explicitly. However, when two or more models satisfy the initial selection condition, the selection of a single model could be deemed inappropriate on the basis that the data are simply ambiguous to alternative mechanisms. If the correlation of the score estimates for these models is high, it is likely that the models have similar structure but differ by one or more correlated predictors. On the other hand, if the correlation is low, yet score estimates are comparable, then each model is capturing independent aspects of the true data-generating mechanism. This situation suggests the potential merits of either (1) model averaging, (2) retaining a best set of models, or (3) considering alternative model structures (Garthwaite and Mubwandarikwa 2010). In this scenario, data simulation based on predictive distributions provides a rigorous means to validate candidate models, assessing their adequacy to explain the data (Gabry et al. 2019).

Finally, we comment on the use of $\sigma_{\Delta S_k}$ (see Eq. 5), in place of $\sigma_{\text{adj},k}$ (Eq. 8), as a preferred performance threshold for the modified selection rule. In practice, it may not make much difference which error definition is used, since both forms account explicitly for the correlation of score estimates, leading to substantially reduced lengths of the error bars of the best-performing models relative to the non-adjusted error $\sigma_{\min}$. In our experience, the difference in size of the two error estimates is often small relative to the differences in the model scores, thus commonly leading to selection of the same model.

At times, however, the two error definitions will select different models, in which case the interpretation of the selected models is different. A key difference is that $\sigma_{\Delta S_k}$ does not satisfy the constraints (Expression 7). These constraints, defined at the limiting values of the possible correlation values, impose logical and intuitive conditions on the modified selection mechanism, appealing to the underlying rationale of the original one-standard-error

rule, that the standard error of the lowest-scoring model is a "natural" threshold for the selection of a simpler model (here adjusted to account for sampling variability). On the other hand, $\sigma_{\Delta S_k}$ permits a probabilistic interpretation, such that an alternative model, whose score estimate falls within $\sigma_{\Delta S_k}$ of the lowest-scoring model, performs no worse than the lowest-scoring model with probability approximately 0.16 (Piironen et al. 2020). Thus, we suggest the use of $\sigma_{\Delta S_k}$ to calculate probabilities related to model performance, and $\sigma_{\text{adj},k}$ for use in the modified one-standard error rule. The latter permits a straight-forward interpretation in the context of parsimonious model selection and it is the definition we apply in the following examples.

Examples

To illustrate the use of the modified selection rule in practice, we apply it to two data sets, both of which have been studied previously in the model-selection literature. All analyses are done in R (R Core Team 2020); the code is available in the supplementary materials.

Goby survival

The Goby survival data set contains measurements from experimental manipulations on *Elacatinus evelynae* and *E. prochilos* in the U.S. Virgin Islands, 2000–2002 (Wilson 2004). The data are compiled from five experiments, $x = 1, 2, ..., 5$, collected over three years, and include information on animal density d and habitat quality q. The fraction of surviving gobies $T(t)$ is modeled using the Weibull distribution; the aim of the analysis is to investigate the effect of density and habitat quality on mortality rate. Following the analysis of Bolker (2008), we consider a set of 10 candidate models characterized by the dependence of the scale parameter s on the variables x, d, and q. The most complex model, denoted xqdi, is

$$T \sim \text{Weibull}(a, s_x(d, q))$$

$$s_x(d, q) = \exp(\alpha_x + \beta q + (\gamma + \delta q)d)$$

with s depending on $x, q, d,$ and an interaction i between q and d. Models of lower complexity, with various dependencies omitted, are labelled in the same manner, together with zero, which denotes the simplest model (a single shape and scale parameter). The data set contains 369 rows.

The selection results are shown in Fig. 2. We use (Eq. 2) to estimate model scores using leave-one-out cross validation, and the standard errors are calculated using 1,000 bootstrap samples. The lowest scoring model (qd) has four parameters and includes both quality and density as predictors. The modified rule selects the simpler model (d), comprising three parameters with no dependence on quality. The same results are obtained when using AIC instead of cross validation. In this example, the lowest-scoring model when using BIC coincides with the one chosen by the modified selection rule. The ordinary one-standard-error rule selects the null model due to the large (non-adjusted) error bars.

Body fat data

Here we consider a classical variable-selection problem using multivariate linear regression. The body-fat data set comprises body-fat densities for 252 men, ages 21–81, together with each subject's age, mass, height, and 10 body circumference measurements (neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist; Johnson 1996). We take $y = 1/\text{density}$ as the response variable and exhaust all possible linear combinations of the 13 predictors to generate a total 8,192 candidate models. This selection problem has been studied by Burnham and Anderson (2002), using AIC and BIC, and by Hoeting et al. (1999), in the context of Bayesian model averaging. As the number of models is much greater than the number of data points, we anticipate large selection uncertainty and a high probability of overfitting when selecting the model with the lowest estimate of Kullback-Leibler discrepancy.

The selection results are shown in Fig. 3. To illustrate the versatility of the rule, we use the small-sample-corrected information criterion AIC_c since the ratio of predictors to data points is low and the models are linear (Hurvich and Tsai 1989). (Alternatively, leave-one-out cross validation can be used, which selects a larger model than AIC_c using score minimization, but selects the same model when applying the modified rule.) Standard errors for the scores are calculated using 1,000 bootstrap samples. The modified rule selects a simpler model than the lowest-AIC_c selection, the former nested within the latter, containing four predictors instead of six. The same four predictors comprise the lowest scoring model using BIC. The modified rule identifies a second model with the same complexity and a higher score; the models differ by one predictor, the correlation of the differing predictors is $r = 0.68$. The ordinary one-standard-error rule appears to seriously underfit, selecting only two predictors, both of which are contained in the models selected by the modified rule. The equivalence of the modified selection with the BIC selection should be taken as a coincidence; see also (Hoeting et al. 1999: Table 9) for comparison with posterior model probabilities.

Finally, to illustrate the utility of the rule for a smaller set of candidate models, we applied the modified rule to the reduced set of body-fat models, based on a priori considerations and transformed predictors, introduced by Burnham and Anderson (2002; see reference and supplementary materials for details). In this instance, the modified rule selects the lowest-scoring model, therefore making the same selection as the usual AIC_c approach, but with increased confidence that overfitting has not occurred.

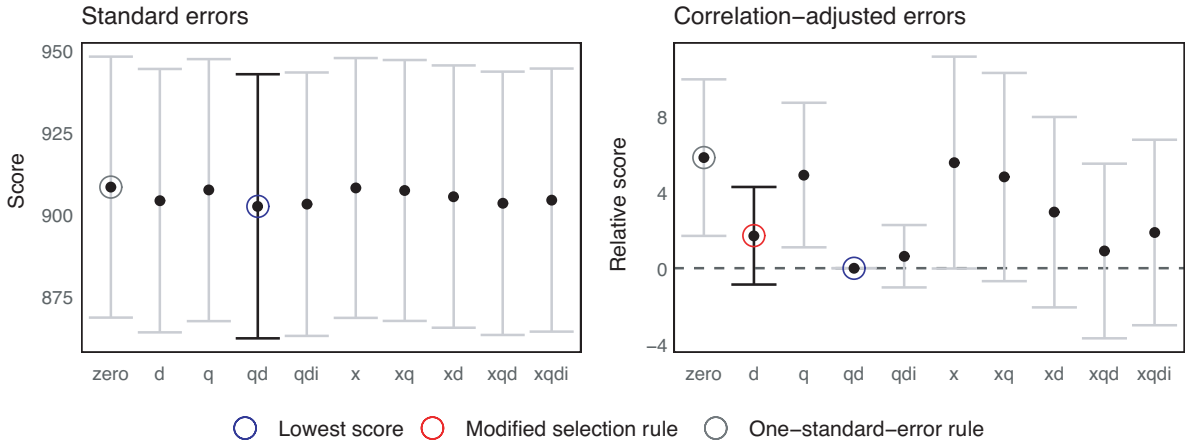

FIG. 2. Score estimates for Goby-survival models. The labels on the *x*-axis denote candidate models ordered by increasing complexity. The lowest-scoring model (qd) includes both quality and density as covariates, however, after accounting for estimation uncertainty, the modified rule selects the simpler model (d), which excludes habitat quality. The ordinary one-standard-error rule selects the null model due to the large (non-adjusted) error bars shown in the left figure.

DISCUSSION

We have introduced an easily implemented method to select a parsimonious model. This is useful in ecology where explanation, rather than the most accurate prediction, is often the goal, and for which the simpler structure of a parsimonious model facilitates interpretation and the inference of influential processes. The method estimates both score uncertainty and the correlation of model scores, using this information to address the shortcomings of naïvely selecting the lowest-scoring model. The standard-error plots give a useful visual summary of the model selection problem, displaying scores, estimation uncertainty and the modified selection mechanism. For parsimonious model selection, our rule provides an information-theoretic alternative to the common, but usually inappropriate, use of consistent selectors, such as BIC.

Our modified selection rule is a simple calibration of the commonly adopted rule to select the model with the lowest score. The need to calibrate the significance of the differences between model scores is identified in the

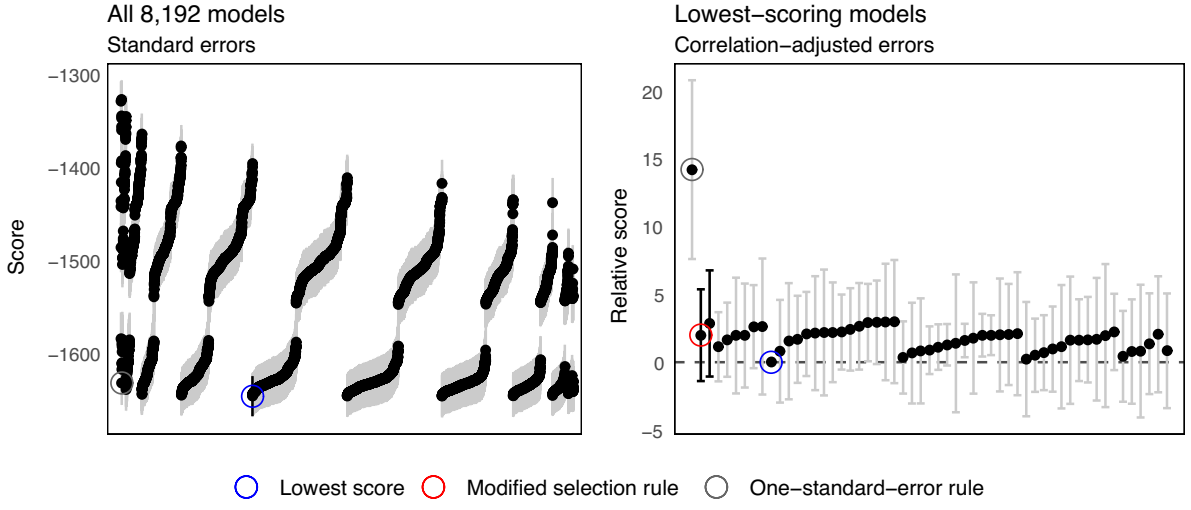

FIG. 3. Score estimates for body-fat models. The left figure shows all 8,192 models, first grouped by complexity, then ordered by score estimate. The models at the lowest point in these groups have comparable estimates, thus the lowest-scoring model will have a high chance of overfitting due to random variation alone. The correlation-adjusted standard errors in the right figure show this effect more clearly. The modified rule selects a four-predictor model, which is two predictors less than the lowest-scoring model. The two black error bars in the right plot indicate that two models of equal complexity fulfil the selection condition of the modified rule; the lowest-scoring of the two is circled.

literature (Vehtari and Ojanen 2012). The main reasons for this are that estimation variance makes selection of lowest-scoring model suboptimal, with a proclivity to overfit, and that a preference for parsimony is not built into the selection mechanism. Importantly, calibration, when used in conjunction with unbiased (or constant-biased) score estimation, respects the intended measure of model performance, in contrast to non-calibrated approaches that mitigate overfitting using estimators with complexity-dependent bias (Cawley and Talbot 2010, Arlot and Lerasle 2016).

Existing approaches to calibrate selection are mostly limited to Bayesian analyses, and usually require specification of an arbitrary threshold. Bayesian reference methods are known to perform well, but they are complex, requiring the construction of a large reference model that is used to calibrate the explanatory power of simpler models (Vehtari and Lampinen 2004). These methods, as well as parameterized techniques such as binomial calibration (McCulloch 1989) or the Gaussian approach of Bernardo (1999), require that an arbitrary threshold parameter be specified. In the context of AIC scores, Richards (2008) introduced a simple, yet effective, selection rule for eliminating complex nested models that fail to improve on the score of simpler alternatives. The original one-standard-error rule uses the standard error of the lowest score as a "natural" calibration threshold, but failure to account for correlation means that the original rule does not generalize well to model-selection problems, with a tendency to underfit.

Here we have made explicit use of the correlation of score estimates between models, a quantity that is rarely computed, despite its usefulness to both model selection and model averaging. The correlation of predictive performance provides a measure of model similarity, for which the corresponding (adjusted) standard errors provide an adaptive alternative to the static thresholds for ΔAIC that are often used to qualify the support for a given model, such as the rule-of-thumb that $\Delta\text{AIC} < 2$ indicates "substantial" support (Burnham and Anderson 2002). In the context of model averaging, the benefits of using averaged predictions are known to be greatest when the correlation between models is low (Garthwaite and Mubwandarikwa 2010, Dormann et al. 2018); thus, the plots of correlation-adjusted standard errors, or indeed the full covariance matrix, provide a useful diagnostic to assess the potential merits of retaining a best set of models in lieu of selecting a single best model.

While we have focused herein on information-theoretic applications, our modified rule is broadly applicable, working with almost any type of model score. Parameter estimation and model scoring are not always based on likelihood. Other fitness functions, such as misclassification rate or the continuous-ranked-probability score, are often more suitable (Gneiting and Raftery 2007), yet the problem of estimation uncertainty and subsequent overfitting remains. The use of nonparametric techniques such as the bootstrap and cross validation, adjusted appropriately for hierarchical or autocorrelative structures (Roberts et al. 2017), enable the modified rule to be applied in a wide variety of settings, although parametric estimation of the variance terms may be more accurate when a specific distribution can be assumed (Efron 2004).

We have focused on selection, but also wish to draw attention to the important yet neglected topic of inference after selection. Breiman and Spector (1992) called it the "quiet scandal in the statistical community," whereby the data are initially used to select a model, and then, acting as though the model was decided upon a priori, the same data are used to infer parameter values and associated confidence intervals. This approach ignores selection uncertainty, leading to exaggerated effect sizes and optimistic confidence intervals (Hjort and Claeskens 2003). In the present context of selecting a single best model, however calibrated to mitigate overfitting, the need to account for post-selection bias remains an important subsequent stage in the analysis, as it is in all model-selection settings.

What can be done to improve post-selection inference? This is an active area of research (Claeskens and Hjort 2008, Berk et al. 2013, Charkhi and Claeskens 2018, Kabaila and Wijethunga 2019). For predictive inference, the post-selection confidence interval of the mean response can be estimated using techniques based on model averaging (Hjort and Claeskens 2003, Efron 2014). Valid confidence intervals for model parameters are more difficult to estimate, yet are crucial for the explanatory goals of inferential analyses; ordinary model averaging does not address this problem (Banner and Higgs 2017). General purpose approaches such as PoSI (post-selection inference; Berk et al. 2013, Bachoc et al. 2019) are suitable for use with our modified rule, though possibly conservative, while more specialized estimators exist for the lasso (Lee et al. 2016) and more recently AIC (Charkhi and Claeskens 2018). Didactic publications and robust software implementation are much needed to bring these recent (and highly technical) methods into the toolkit of ecological analysts. Post-selection inference remains an important direction for future work in both statistics and ecological inference. Low-tech solutions include the perennial advice to do as much work as possible a priori, thus reducing the number of candidate models.

Finally, given the many difficulties of model selection in a frequentist setting, one might wonder whether Bayesian methods could be used to avoid model selection altogether. Bayesian data analysis provides a broadly inclusive framework to incorporate and marginalize over uncertainties. Large, globally encompassing models, that are regularized via priors and hierarchical structures, are an enticing one-stop-shop, but they require an intimidating (at times prohibitive) amount of prior knowledge (Hjort and Claeskens 2003). Indeed, the need for discrete model selection in a Bayesian setting is well-recognized, and has been reiterated in the recent literature (Gelman

et al. 2014, Hooten et al. 2015). Projection methods, related to reference methods discussed above, and information-theoretic scores, such as the widely applicable information criterion (WAIC) and Bayesian cross validation, are topics of active research (Piironen and Vehtari 2017, Vehtari et al. 2017). Score-based model selection has been incorporated into the popular R package rstanarm (Goodrich et al. 2020), or loo (Vehtari et al. 2017) for more general use, facilitating the routine use of model selection in Bayesian analyses, wherein our modified selection rule can also be applied.

Conclusion

The notion of selecting the "best," the "true," or the "most parsimonious" model are unattainable ideals in the real world of noisy data and approximate models of complex natural phenomena. Here we have introduced a practical rule to select a useful model. The model is useful because it is selected on the basis of its expected utility (score), while at the same time giving preference to a parsimonious selection when it is justified by the extent of estimation uncertainty, thereby mitigating overfitting. Defined in this way, the rule provides an objective (and intuitive) selection mechanism, where expert knowledge is incorporated a priori via the generation of candidate models and the choice of estimation method, model score, and selection rule. This makes the preference for parsimony explicit. Further, when applied using unbiased estimates of Kullback-Leibler discrepancy, the selected model retains an information-theoretic interpretation. The rule is versatile, easy to apply, and permits a straightforward interpretation, aided by visual plots.

Acknowledgments

This work was funded by the Australian Research Council grant FL160100101. The authors would like to thank Jackie Wilson for her permission to use the marine gobies data, and William Link and Jean-FranÓois Le Galliard for constructive critical comments on an earlier version of the manuscript.

Literature Cited

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov, and F. Csaki, editors. Second International Symposium on Information Theory (Tsahkadsor, 1971). Academiai Kiado, Budapest, Hungary.

Arlot, S., and M. Lerasle. 2016. Choice of V for V-fold cross-validation in least-squares density estimation. Journal of Machine Learning Research 17:1–50.

Bachoc, F., H. Leeb, and B. M. Pötscher. 2019. Valid confidence intervals for post-model-selection predictors. Annals of Statistics 47:1475–1504.

Banner, K. M., and M. D. Higgs. 2017. Considerations for assessing model averaging of regression coefficients. Ecological Applications 27:78–93.

Bengio, Y., and Y. Grandvalet. 2004. No unbiased estimator of the variance of K-fold cross-validation. Journal of Machine Learning Research 5:1089–1105.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao. 2013. Valid post-selection inference. Annals of Statistics 41:802–837.

Bernardo, M. 1999. Nested hypothesis testing: the Bayesian reference criterion. Pages 101–130 *in* J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors. Bayesian statistics. Volume 6. Oxford University Press, Oxford, UK.

Bolker, B. M. 2008. Ecological models and data in R. Princeton University Press, Princeton, New Jersey, USA.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Boca Raton, Florida, USA.

Breiman, L., and P. Spector. 1992. Submodel selection and evaluation in regression. The X-random case. International Statistical Review 60:291.

Burman, P. 1989. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika 76:503–514.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer-Verlag, New York, New York, USA.

Cawley, G. C., and N. L. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11:2079–2107.

Charkhi, A., and G. Claeskens. 2018. Asymptotic post-selection inference for the Akaike information criterion. Biometrika 105:645–664.

Claeskens, G., and N. L. Hjort. 2008. Model Selection and model averaging. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, New York, New York, USA.

Dormann, C. F., et al. 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs 88:485–504.

Efron, B. 2004. The estimation of prediction error. Journal of the American Statistical Association 99:619–632.

Efron, B. 2014. Estimation and accuracy after model selection. Journal of the American Statistical Association 109:991–1007.

Efron, B., and R. Tibshirani. 1994. An introduction to the bootstrap. Chapman and Hall/CRC, London, UK.

Gabry, J., D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. 2019. Visualization in Bayesian workflow. Journal of the Royal Statistical Society: Series A (Statistics in Society) 182:389–402.

Garthwaite, P. H., and E. Mubwandarikwa. 2010. Selection of weights for weighted model averaging. Australian and New Zealand Journal of Statistics 52:363–382.

Gelman, A., J. Hwang, and A. Vehtari. 2014. Understanding predictive information criteria for Bayesian models. Statistics and Computing 24:997–1016.

Gneiting, T., and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102:359–378.

Goodrich, B., J. Gabry, I. Ali, and S. Brilleman. 2020. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1. https://cran.r-project.org/package=rstanarm

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning data mining, inference, and prediction. Second edition. Springer Series in Statistics, New York, New York, USA.

Hjort, N. L., and G. Claeskens. 2003. Frequentist model average estimators. Journal of the American Statistical Association 98:879–899.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. Statistical Science 14:382–401.

Hooten, M. B., N. T. Hobbs, and A. M. Ellison. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3–28.

Hurvich, C. M., and C. L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning. Springer, New York, New York, USA.

Johnson, R. W. 1996. Fitting percentage of body fat to simple body measurements. Journal of Statistics Education 4. https://doi.org/10.1080/10691898.1996.11910505

Kabaila, P., and C. Wijethunga. 2019. On confidence intervals centred on bootstrap smoothed estimators. Stat 8: e233.

Kuhn, M. 2008. Building predictive models in R using the caret Package. Journal of Statistical Software 28:1–26.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor. 2016. Exact post-selection inference, with application to the lasso. Annals of Statistics 44:907–927.

McCulloch, R. E. 1989. Local model influence. Journal of the American Statistical Association 84:473–478.

McLeod, A., C. Xu, and Y. Lai. 2020. bestglm: Best subset GLM and regression utilities. R package version 0.37.3. https://cran.r-project.org/package=bestglm

Piironen, J., M. Paasiniemi, and A. Vehtari. 2020. Projective inference in high-dimensional problems: Prediction and feature selection. Electronic Journal of Statistics 14:2155–2197.

Piironen, J., and A. Vehtari. 2017. Comparison of Bayesian predictive methods for model selection. Statistics and Computing 27:711–735.

R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org

Richards, S. A. 2005. Testing ecological theory using the information-theoretic approach: Examples and cautionary results. Ecology 86:2805–2814.

Richards, S. A. 2008. Dealing with overdispersed count data in applied ecology. Journal of Applied Ecology 45:218–227.

Richards, S. A., M. J. Whittingham, and P. A. Stephens. 2011. Model selection and model averaging in behavioural ecology: The utility of the IT-AIC framework. Behavioral Ecology and Sociobiology 65:77–89.

Roberts, D. R., et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929.

Southwood, T. R. E. 1978. Ecological methods with particular reference to the study of insect populations. Second edition. Chapman and Hall, London, UK.

Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2019. blockCV: An R package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models. Methods in Ecology and Evolution 10:225–232.

Vehtari, A., A. Gelman, and J. Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing 27:1413–1432.

Vehtari, A., and J. Lampinen. 2004. Model selection via predictive explanatory power. Technical report. Helsinki University of Technology, Espoo, Finland.

Vehtari, A., and J. Ojanen. 2012. A survey of Bayesian predictive methods for model assessment, selection and comparison. Statistics Surveys 6:142–228.

Wilson, J. A. 2004. Habitat quality, competition, and recruitment processes in two marine gobies. Ph.D. dissertation. Department of Zoology, University of Florida, Gainesville, Florida, USA.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.3475/suppinfo