A Guide to Bayesian Model Selection for Ecologists

M.B. Hooten[1,2,3,4,*] and N.T. Hobbs[4,5,6]

[1]U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit
[2]Department of Fish, Wildlife, and Conservation Biology, Colorado State University
[3]Department of Statistics, Colorado State University
[4]Graduate Degree Program in Ecology, Colorado State University
[5]Department of Ecosystem Science and Sustainability, Colorado State University
[6]Natural Resource Ecology Laboratory, Colorado State University

[*]Corresponding Author; Email: mevin.hooten@colostate.edu

May 15, 2014

ABSTRACT

The steady upward trend in the use of model selection and Bayesian methods in ecological research has made it clear that both approaches to inference are important for modern analysis of models and data. However, in teaching Bayesian methods and in working with our research colleagues, we have noticed a general dissatisfaction with the available literature on Bayesian model selection and multimodel inference. Students and researchers new to Bayesian methods quickly find that the published advice on model selection is often preferential in its treatment of options for analysis, frequently advocating one particular method above others. The recent appearance of many articles and textbooks on Bayesian modeling has provided welcome background on relevant approaches to model selection in the Bayesian framework, but most of these are either very narrowly focused in scope or inaccessible to ecologists. Moreover, the methodological details of Bayesian model selection approaches are spread thinly throughout the literature, appearing in journals from many different fields. Our aim with this guide is to condense the large body of literature on Bayesian approaches to model selection and multimodel inference and present it specifically for quantitative ecologists as neutrally as possible. We also bring to light a few important and fundamental concepts relating directly to model selection that seem to have gone unnoticed in the ecological literature. Throughout, we provide only a minimal discussion of philosophy, preferring instead to examine the breadth of approaches as well as their practical advantages and disadvantages. This guide serves as a reference for ecologists using Bayesian methods, so that they can better understand their options and can make an informed choice that is best aligned with their goals for inference.

KEYWORDS

Akaike Information Criterion, Bayes Factors, Cross-Validation, Deviance Information Criterion, Model Averaging, Multi-Model Inference, Regularization, Shrinkage

1 INTRODUCTION

Model selection and Bayesian statistics have become increasingly important tools in the field of ecology (Johnson and Omland, 2004; Clark, 2005; Cressie et al., 2009; Hobbs, 2009). Despite an upward trend in the use of model selection and Bayesian methods in ecological research, the intersection of these two frameworks for inference has been minimal in the literature (Figure 1). The guidance provided about model selection in the Bayesian statistical literature is unbalanced and lacks cohesion. The theory and protocol for implementing a variety of Bayesian model selection methods seem much less tangible than the information criterion approaches for maximum likelihood we have grown accustomed to in ecology. Thus, we are at a critical juncture in our field. Do we use newer statistical technology while potentially foregoing model selection because it is too complicated, or do we use more familiar statistical methods at the potential risk of letting our choice of selection procedure dictate what scientific questions we can answer with our model(s)? An awareness of available model comparison approaches in the Bayesian framework can help the ecologist choose and apply the method that is most suited to their goals for inference.

[Figure 1 Here]

1.1 Preliminary Assumptions and Notation

Our primary focus is on providing a comprehensive description of available methods for Bayesian model selection and multimodel inference that is accessible to ecologists. For a discussion of the philosophical arguments pertaining to model selection and multimodel inference we refer the interested reader to several excellent sources, including Gelman and Shalizi (2012) and Ver Hoef and Boveng (In Review), who discuss when and why one should use model selection methods. In this exposition, we assume the reader is familiar with the philosophical underpinnings and has already decided that they 1.) seek Bayesian statistical inference, 2.) would like to compare models for the purpose of improving that inference, and 3.) have already verified the model assumptions for their particular data set. This last item is critical because if the model assumptions are not met, the resulting statistical inference (including predictions and prediction uncertainty) rests on a house of cards. Reliable inference requires checking the assumptions of our models. For further details on model checking, including the evaluation of goodness-of-fit and posterior predictive p-values, see Gelman et al. (2014 a).

We also assume the reader has broad familiarity with statistical methods including least squares and maximum likelihood, as well as a basic understanding of Bayesian model building and algorithms for implementation (e.g., Markov chain Monte Carlo). Gotelli and Ellison (2012) and Bolker (2008) provide excellent background on contemporary ecological statistics, and from a Bayesian perspective see Clark (2007), Royle and Dorazio (2008), Link and Barker (2010), and Hobbs and Hooten (In Review).

We make frequent use of matrix notation and linear algebra (to avoid excessive summation notation) throughout this guide, but readers unfamiliar with these concepts will be able to glean the big-picture concepts and connections from our descriptions. In particular, we use a common Bayesian square bracket notation '$[a|b]$' (courtesy of Gelfand and Smith, 1990) to represent probability distributions, in this case, the distribution of variable 'a' given variable 'b.' We also make occasional use of the probability notation '$P(c)$' to denote the probability of item 'c.' For matrix notation, we use a standard form where matrices and vectors are bold, with matrices uppercase (e.g., $\mathbf{X}$) and vectors lowercase (e.g., $\mathbf{x}$). Matrix and vector transpose is denoted by the "prime" symbol (e.g., $\mathbf{x}'$). We use $\boldsymbol{\theta}$ generically to denote a set of model parameters, and $\mathbf{y}$ to denote a data set, typically composed of response variables. Finally, we have defined several commonly used terms in the model selection and Bayesian literature in Table 1 to aid those readers less familiar with the subject.

1.2 Overview of Topics

In this guide, we present a wealth of available perspectives on Bayesian multimodel inference and model selection. It may come as a surprise that there are many options for model selection and multimodel inference, each with its own strengths and weaknesses. It is our view that ecologists need the ability to distinguish among methods more than they need a strict set of rules to follow in how to proceed with model selection. We use the term "guide" here (in the same sense as a field guide for birds) because we have made an effort to be thorough and to remain unaffiliated in our description of these methods. Our guide is intended to be used as a conceptual aid; ecologists can use it to learn about the variety of options available and can decide how each fits in with their own research goals. For illustration, we implement several specific methods (all computer code is available in the supplemental material). However, as space does not allow us to provide specific examples of computational algorithms for every approach, we have made an effort to provide the reader with numerous references they can consult to implement these methods in the statistical software of their choice.

This paper is organized as follows. We begin by highlighting a few of most important and sometimes lesser known take home messages concerning model selection. This prelude serves as an overview containing big picture connections between the methods we describe subsequently. We then introduce a specific Bayesian ecological model as a case example. We refer to this example throughout to illustrate differences among alternative approaches. In Section 2, we describe Bayesian model averaging, for use when the goal of the researcher is to make inferences from more than one model. In Section 3, we treat out-of-sample validation, the gold standard for model selection based on predictive ability. We then turn to a topic in Section 4 that applies broadly across Bayesian and non-Bayesian statistics, the process of regularization, which we feel is essential to understanding the subsequent material (Section 5) on information criteria. Section 6 covers model-based methods for model selection. In the penultimate Section, we provide specific guidance on matching alternative methods to inferential goals. As a visual aid to the flow of the manuscript, we show section topics and sub-topics in Figure 2, providing an overview for the relationships among ideas and methods that we describe throughout the paper.

[Figure 2 Here]

1.3 Highlights

While preparing this guide, we experienced several epiphanies ourselves that had not occurred to us previously. We discovered that most of these findings have existed in the literature for quite some time (a decade, at least), but had not been brought together in a way that supports a solid understanding and intuition about model selection. Among the most important of our own epiphanies were:

- There is no general consensus among statisticians on the topic of model selection.
- Multimodel inference can be thought of from many different perspectives, including model averaging. Thus, we use the phrase “model selection” somewhat generically (including model comparison and multimodel inference) because many of the methods we describe inherently consider multiple models (sometimes infinitely many), but aren’t considered to be model averaging in the conventional sense.
- Much of the statistical community relies heavily on out-of-sample model comparison

approaches, yet in ecology we primarily favor information criterion approaches that avoid the use of out-of-sample data for model evaluation. Despite the potential advantages for model selection, out-of-sample methods have been largely ignored by ecologists because they 1.) may require additional data beyond what was already collected in the study and 2.) historically were very computationally intensive to implement.

- Cross-validation is a hybrid approach containing both out-of-sample and within-sample aspects. From a Bayesian perspective, cross-validation for model selection is considered to be an empirical Bayesian method and can be incredibly helpful for model selection.

- Neither AIC nor BIC are appropriate for Bayesian model averaging in all situations. Both AIC and BIC were designed to be used with maximum likelihood estimates and make fairly strong assumptions about *a priori* model probabilities. Whereas AIC excels at finding good predictive models, BIC was developed mainly for model averaging purposes and is good for small sets of well-justified models.

- DIC and AIC often yield quite similar results for model selection with certain classes of models, however, DIC is not ideal for all classes of models (e.g., mixture models). No theoretical justification exists in the literature for the use of DIC in model averaging. Furthermore, DIC is not a fully Bayesian model comparison criterion.

- A truly Bayesian information criterion seems to have just been discovered (i.e., WAIC), but in actuality went unnoticed for more than a decade. WAIC resolves many of the issues with DIC, but also seems to have a critical weakness for some models.

- Regularization is an umbrella concept that spans nearly all topics in model selection. When statistical optimization problems are written as regularization expressions, it becomes clear that AIC, BIC, DIC, WAIC, posterior predictive loss, ridge regression, and Lasso all fall under the same umbrella. Moreover, regularization itself has an inherently Bayesian justification. It explicitly constrains model parameters in the same way a Bayesian prior does. Thus, model selection is similar to using a strong prior, at least in spirit.
- The Bayesian framework allows one to actually build parametric mechanisms into models that perform model selection (e.g., stochastic search variable selection and reversible jump MCMC). We refer to these as model-based model selection approaches. They can be viewed as a combination of model selection and multimodel inference.

1.4 An Exemplar: The Hierarchical Bayesian Occupancy Model

Mixture models, especially zero-inflated models, comprise an important class of statistical tools in contemporary ecological research. In particular, occupancy and capture-recapture models are very commonly used in the field of wildlife ecology (Royle and Dorazio, 2008). We consider the hierarchical occupancy model as a prototypical Bayesian ecological model. The Bayesian occupancy model presents challenges for traditional model comparison methods, thus, we introduce the model here and refer back to it later to demonstrate several approaches for model selection and multimodel inference.

In essence, the occupancy model is simply a binary regression model with binary measurement error. In its application, the occupancy model can be used to learn about the true presence or absence of a species and the niche-related features of the sites while accounting for imperfect detection (MacKenzie et al., 2006). The basic occupancy model, presented for ecologists, was described by MacKenzie et al. (2002) and included implementation details from a maximum likelihood perspective. More recently, occupancy models have been extended to model temporal dynamics (e.g., MacKenzie et al., 2003), spatial autocorrelation (e.g., Johnson et al., 2013), and community dependence (e.g., Dorazio et al., 2010).

Hierarchically, a simple occupancy model with homogeneous detection probability and heterogeneous occupancy probabilities can be written as a zero-inflated binomial data model (with detection probability p) that depends on a latent Bernoulli process (z_i, presence or absence) that varies among sites ($i = 1, \dots, n$) according to probability ψ_i. The response data, y_i, are a sum of the binary detection history for each site over a set of visits or occasions (J_i); that is, $y_i = \sum_{j=1}^{J_i} y_{ij}$, where y_{ij} are binary detection observations for site i on survey occasion j. On each occasion, the species is detected (i.e., $y_{ij} = 1$) with probability p if it is truly present, otherwise it is recorded as not detected (i.e., $y_{ij} = 0$). For simplicity, we have used a specification of the occupancy model that assumes a homogeneous detection probability p and conditional independence for detection on each site visit $j = 1, \dots, J_i$. These assumptions can be relaxed by allowing for variation in detection as well as occupancy probability.

The logit link, $\log(\psi_i/(1 - \psi_i))$, is most commonly used function relating occupancy probability ψ_i to a set of site-level covariates $\mathbf{x}_i$, however there can be computational advantages to using other link functions such as the probit (Hooten et al., 2003; Dorazio and Rodriguez, 2012; Johnson et al., 2013). The probit link function allows us to reparameterize the model using a set of auxiliary variables v_i that describe a continuous latent process representing occupancy probability (Albert and Chib, 1990). The probit occupancy model is specified hierarchically as

$$y_i \sim \begin{cases} 0 & \text{if } z_i = 0 \\ \text{Binom}(J_i, p) & \text{if } z_i = 1 \end{cases} , \tag{1}$$

$$z_i \sim \begin{cases} 0 & \text{if } v_i \leq 0 \\ 1 & \text{if } v_i > 0 \end{cases} , \tag{2}$$

$$v_i \sim \text{N}(\beta_0 + \mathbf{x}_i' \boldsymbol{\beta}, 1) , \tag{3}$$

$$p \sim \text{Beta}(1, 1) , \tag{4}$$

$$\beta_0 \sim \text{N}(\mu_0, \sigma_0^2) , \tag{5}$$

$$\boldsymbol{\beta} \sim \text{N}(\boldsymbol{\mu}, \sigma_\beta^2 \mathbf{I}) , \tag{6}$$

where the probit link function itself (i.e., Φ, the standard normal cumulative distribution function) only comes into play when we condition z_i on the regression coefficients β_0 and $\boldsymbol{\beta}$ directly; then we obtain $z_i \sim \text{Bernoulli}(\Phi(\beta_0 + \mathbf{x}_i' \boldsymbol{\beta}))$. The advantages of this probit occupancy model are primarily computational. The implicit probit link function allows us to create a fully Gibbs MCMC algorithm that requires no Metropolis-Hastings updates or tuning (Dorazio and Rodriguez, 2012; Johnson et al., 2013). We use the probit occupancy model presented in (1)–(6) as a basis for demonstrating the model selection procedures that follow, making modifications to it as needed.

2 MODEL AVERAGING

From here forward, assume that we are dealing with a set of models $\mathcal{M}$ =$\{M_1$,. . ., M_l,. . ., $M_L\}$ that are built using expert scientific judgement and are not obviously inappropriate in terms of assumptions. Model averaging allows us to combine the strengths of several models for improved inference. It has been argued (e.g., Kass and Raftery, 1995; Link and Barker, 2006) that Bayesian model averaging (BMA) is the proper way to obtain multimodel inference under the Bayesian statistical paradigm because it provides a valid probability-based mechanism for considering multiple models in the presence of process and parameter uncertainty. Hoeting et al. (1999) provided an excellent overview of BMA, complete with implementation details for selected model classes.

An important and often overlooked aspect of model averaging is that BMA was not designed as a method for model selection, but rather as a method for combining posterior distributions. Whereas many of the methods in the following Sections are based heavily on finding models that excel at out-of-sample predictive performance (e.g., AIC and DIC), BMA is intended for within-sample model combination. Thus, in what follows, we provide some insight about how BMA fits into the larger suite of model selection methods and refer the interested reader to the literature cited herein for details.

At the heart of BMA is the average posterior distribution of a quantity of interest ($g \equiv g(\boldsymbol{\theta}, \tilde{\mathbf{y}})$, typically a function of either an unknown parameter or set of data or both)

$$[g|\mathbf{y}] = \sum_{l=1}^{L} [g|\mathbf{y}, M_l] P(M_l|\mathbf{y}) \ , \tag{7}$$

where $[g|\mathbf{y}, M_l]$ is the posterior distribution of g under individual model M_l and $P(M_l|\mathbf{y})$ is the posterior probability of model M_l. The posterior model probability $P(M_l|\mathbf{y})$ is the workhorse of the BMA procedure, providing the weight of evidence in the average (7) for one model over others. Thus, we have a natural and proper Bayesian framework for multimodel inference as long as we can find the required quantities in (7). Furthermore, BMA performed on a set of models $\mathcal{M}$ yields better inference about g than any one of the models alone (Madigan and Raftery, 1994), thus we have a compelling reason to use it.

2.1 The Utility of the Marginal Data Distribution

Recall the classical expression for Bayes rule assuming a single model

$$[\boldsymbol{\theta}|\mathbf{y}] = \frac{[\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{[\mathbf{y}]} \ , \tag{8}$$

where $[\boldsymbol{\theta}]$ is the prior distribution for the parameters. The denominator $[\mathbf{y}]$, which we typically avoid finding analytically, corresponds to the aforementioned marginal data distribution for the given model; it will be large for the same set of data if the model represents them well and small if it doesn't. The marginal data distribution $[\mathbf{y}]$ is a natural model discrimination measure by itself and is fundamental in computing the posterior model probabilities $P(M_l|\mathbf{y})$. To show this, we generalize the notation to include information concerning the individual model each $[\mathbf{y}]$ is associated with. Therefore, let $[\mathbf{y}|M_l]$ be the marginal data distribution for model l. Then, the posterior model probability can be written as

$$P(M_l|\mathbf{y}) = \frac{[\mathbf{y}|M_l]P(M_l)}{\sum_{j=1}^{L}[\mathbf{y}|M_j]P(M_j)} \ , \tag{9}$$

where $P(M_l)$ is the assumed prior model probability which is commonly set to $1/L$. The use of equal prior model probabilities explicitly assumes that there may be no reason to prefer one model over another. The alternative is to set the $P(M_l)$ such that they represent an *a priori* understanding of differences among model importance as long as the sum of prior model probabilities over all models in the set equals 1. To obtain the necessary marginal data distribution for model l we need to integrate over the parameters in the joint distribution of the data $\mathbf{y}$, the model M_l, and the parameters $\boldsymbol{\theta}$ so that

$$[\mathbf{y}|M_l] = \int [\mathbf{y}|\boldsymbol{\theta}, M_l][\boldsymbol{\theta}]d\boldsymbol{\theta} \ . \tag{10}$$

Note that this (10) is the same expression typically appearing in the denominator of Bayes rule (8).

2.2 Bayes Factors

Assuming that we can find the posterior distribution for the quantity of interest $[g|\mathbf{y}, M_l]$ for all models in $\mathcal{M}$, we need only compute the posterior model weights to find the averaged posterior distribution (7). As it happens, solving the integral in the marginal data distribution (10) is often non-trivial, which is why most Bayesian studies use MCMC to avoid calculating it directly. The sum in the denominator of the posterior model probability (9) can also become intractable as the number of models L grows. Thus, despite its attractiveness and rigor, the challenge with BMA is in its implementation.

Consider the ratio of posterior probabilities for two models, say M_l and $M_{l'}$. Using a bit of algebra it is easy to show that the ratio (i.e., the posterior odds) is

$$\begin{aligned}
\frac{P(M_l|\mathbf{y})}{P(M_{l'}|\mathbf{y})} &= \frac{[\mathbf{y}|M_l]P(M_l)}{\sum_{j=1}^{L}[\mathbf{y}|M_j]P(M_j)} \Big/ \frac{[\mathbf{y}|M_{l'}]P(M_{l'})}{\sum_{j=1}^{L}[\mathbf{y}|M_j]P(M_j)} \\
&= \frac{[\mathbf{y}|M_l]}{[\mathbf{y}|M_{l'}]}\frac{P(M_l)}{P(M_{l'})} \\
&= B_{l,l'}\frac{P(M_l)}{P(M_{l'})} \qquad (11)
\end{aligned}$$

which, after the data $\mathbf{y}$ have been observed, can be written as a constant multiplier of the ratio of prior model probabilities (i.e., the prior odds). The multiplier $B_{l,l'}$ in (11) is known as the Bayes factor and is only a function of the marginal data distributions from each model (Kass and Raftery, 1995). Thus, the posterior evidence in favor of one model over another is found by updating the prior evidence with the data. Similar to the various rules of thumb for comparing models using information criteria, there have been several suggested rules of thumb in the literature for Bayes factors (e.g., $B_{l,l'} > 10$ implies strong evidence in favor of model M_l over model $M_{l'}$ according to Jeffreys (1961)).

The utility of the marginal data distribution for model averaging becomes clear because the posterior probability of any model M_l,

$$P(M_l|\mathbf{y}) = \frac{B_{l,l'}P(M_l)}{\sum_{j=1}^{L} B_{j,l'}P(M_j)}\ , \qquad (12)$$

is obtained by dividing the numerator and denominator in the posterior model probability (9) by $[\mathbf{y}|M_{l'}]$ (Link and Barker, 2006). Thus, if we have the marginal data distributions $[\mathbf{y}|M_l]$ for all models being considered, then we have the Bayes factors $B_{l,l'}$, and if we have the Bayes factors we can compute the exact Bayesian model weights for performing model averaging. Various methods exist for calculating the necessary quantities in Bayesian model averaging (e.g., Congdon, 2006), some of which we will describe in what follows (Sections 4.1.4 and 5.2). Finally, we note that one must be cautious in Bayesian model averaging when improper priors (i.e., prior distributions that do not integrate to 1) are used for parameters, as the Bayes factors are undefined in those settings (Spiegelhalter and Smith, 1982).

2.3 Willow Tit Occupancy: BMA

Royle and Dorazio (2008) describe a data set involving occupancy sampling of Swiss breeding birds as part of the Swiss Survey of Common Breeding Birds (collected by the Swiss Monitoring Haufige Brutvogel, and originally provided by Hans Schmid and Marc Kery). Thanks to Royle and Dorazio (2008), these data have become a standard textbook example used to demonstrate Bayesian occupancy models and can be found at the URL: http://www.mbr-pwrc.usgs.gov/pubanalysis/roylebook/. We use a subset of data consisting of the first 200 quadrats throughout Switzerland where surveys were conducted for up to three sampling occasions. We focus on the same species considered by Royle and Dorazio (2008), the willow tit (*Parus montanus*), a relatively common passerine in Europe that resembles the chickadee of North America in appearance. Royle and Dorazio (2008) analyzed a binary form of the data at each site and occasion (i.e., detected / non-detected) along with covariate information on elevation and forest cover (which we standardize to have mean zero and standard deviation equal to one). Further details concerning data collection methods for this study are described by Kery and Schmidt (2004).

Existing life history information concerning the environmental niche of the willow tit suggests that forest cover and elevation are important features. To demonstrate Bayesian model averaging (as well as the methods that follow) applied to the occupancy model, we constructed a set of 4 distinct candidate models to learn about the niche preferences of this species. Each occupancy model contains a homogeneous detection probability and an occupancy probability that 1.) is homogeneous, containing only an intercept (i.e., β_0; the null model, M_1), 2.) contains an intercept and elevation as a covariate (M_2), 3.) contains an intercept and forest as a covariate (M_3), and 4.) contains an intercept and both elevation and forest as covariates (M_4).

Assuming that we seek to use within-sample data to combine models, we can utilize Bayesian model averaging to obtain improved inference concerning the niche preferences of willow tit in Switzerland. Using the computational approaches described in Section 5 (i.e., reversible-jump MCMC), we calculated the posterior model probabilities for the four models described above (Table 2). Assuming equal prior probabilities for this example (i.e., $M_l = 1/4$ for $l = 1, \dots, 4$), we find that the two models containing the elevation covariate dominate the model averaged inference with posterior model probabilities of $P(M_2|\mathbf{y}) = 0.52$ and $P(M_4|\mathbf{y}) = 0.48$. Given our equal prior model probabilities, the Bayes factor for model M_2 over M_4 is computed as $P(M_2|\mathbf{y})/P(M_4|\mathbf{y}) = 1.08$.

[Table 2 Here]

We demonstrate the differences between posterior means for coefficients among all models considered in Table 3 as well as the model averaged posterior means. Notice that the BMA posterior mean for the elevation coefficient falls between the values resulting from the two models containing that covariate (i.e., M_2 and M_4), while the BMA posterior mean for the forest coefficient shrinks toward zero. This shrinkage of β_1 is caused by the very small posterior model probability for M_2 (i.e., the model with only forest as a covariate), thus down weighting the estimate resulting from that model because it carries little weight in the Bayesian model average.

[Table 3 Here]

Following the line of reasoning provided by Madigan and Raftery (1994) it is common to consider BMA for only the two models containing the elevation covariate because the others have negligible posterior model probabilities. Thus, if one desired BMA inference based on the Occam's window principle (i.e., considering only models carrying substantial weight in the averaging), one would rerun the analysis using only the two top models in this scenario. We return to Bayesian model averaging in Section 5, describing various approaches for computation.

3 MODEL VALIDATION

In this Section, assume again that we are considering a set of models $\mathcal{M}$. But now suppose we are interested in evaluating each model's performance relative to some predefined characteristic. Predictive ability is by far the most commonly sought model characteristic in the literature on model selection and thus we highlight it here. Alternatively, other methods have been developed for selection based on estimation inference (i.e., inference that seeks to improve our understanding of model parameters rather than predictions; Bondell and Reich, 2013).

3.1 Out-of-Sample Validation

If we are interested in prediction as our main characteristic of model utility, then it is sensible to evaluate the model in terms of *real* predictive ability; that is, we seek a model whose predictions are close to out-of-sample data (with closeness measured using a score function). Out-of-sample data are observations that are not used to fit the model but that we can use to compare with model predictions. In the machine learning literature, out-of-sample data are often referred to as "validation" data, whereas within-sample data are commonly referred to as "training" data (Hastie et al., 2009).

The essential idea in out-of-sample validation is that two data sets are collected; one to fit (or train) the model ($\mathbf{y}$) and one to validate the model ($\mathbf{y}_{\text{oos}}$). A large out-of-sample data set will provide the best information about the predictive performance of a model, but is obviously more intensive to collect. Thus, some trade-off between within-sample and out-of-sample data set size is necessary. For large single data sets such as those derived from web searches or financial data it is common to split the data set into two pieces, one for training and another for validation. If the original data set is large enough, the resulting decrease in inferential power due to splitting it up is negligible. In historical ecological studies it was less common to have such large data sets, at least in terms of response variables. However, with remote sensing and newer automated data collection methods such as global positioning system (GPS) telemetry devices, large ecological data sets are more common than ever. Thus, out-of-sample validation methods are becoming more realistic for ecological analyses.

Out-of-sample validation relies on the ability to compute a similarity statistic or scoring rule to obtain a measure of closeness between our out-of-sample data $\mathbf{y}_{\text{oos}}$ and the predictions $\hat{\mathbf{y}}_{\text{oos}}$ (e.g., Bernardo 1979; Czado et al., 2009; Gneiting and Raftery, 2007; Gneiting, 2011). One of the most commonly used scoring rules is the mean squared prediction error (MSPE)

$$\text{MSPE} = \sum_{i=1}^{n_{\text{oos}}} \frac{(y_{i,\text{oos}} - \hat{y}_{i,\text{oos}})^2}{n_{\text{oos}}} \; , \tag{13}$$

or its square root (RMSPE). The prediction, $\hat{y}_{i,\text{oos}}$ in MSPE, is obtained without using the out-of-sample observation $y_{i,\text{oos}}$. The out-of-sample validation procedure can be applied independently for each model in a discrete set of models $\mathcal{M}$ and the predictive scores (e.g., RMSPE_l for model M_l) can be compared to assess which model is best overall at prediction or how the models rank in terms of predictive ability.

The MSPE is a popular scoring rule because it has important properties when used with certain models. In general, Bernardo and Smith (1994) recommend logarithmic scoring rules that are both “local” and “proper.” In essence, these scoring rule characteristics guarantee that the predictive score adheres to the chosen model and data (Vehtari and Ojanen, 2012; Gelman et al., 2014 b). We describe a more general approach for scoring models based on out-of-sample data in what follows.

The practice of evaluating models based only on point estimates of parameters or predictions does not naturally incorporate our uncertainty pertaining to those quantities. One of the primary advantages of Bayesian inference is the ability to account for various sources of uncertainty, thus we now describe a method for model validation that appropriately accommodates uncertainty. In doing so, it is critical to recall how prediction works from the Bayesian perspective. In general, data that have not been observed are considered to be random quantities, thus we treat them like all other random quantities in the Bayesian setting and seek their posterior distribution. The posterior distribution for predictions is called the “posterior predictive distribution” and can be found using the integral

$$[\mathbf{y}_{\text{oos}}|\mathbf{y}] = \int [\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \ . \tag{14}$$

One option for the point prediction itself ($\hat{\mathbf{y}}_{\text{oos}}$) could be the posterior predictive mean, which technically requires another integral. That is,

$$\hat{\mathbf{y}}_{\text{oos}} = \text{E}(\mathbf{y}_{\text{oos}}|\mathbf{y}) = \int\int \mathbf{y}_{\text{oos}}[\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}d\mathbf{y}_{\text{oos}} \ , \tag{15}$$

which can be easily approximated as long as the out of sample data $\mathbf{y}_{\text{oos}}$ can be sampled from the distribution $[\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}]$ within an MCMC algorithm. If this condition is met, one can use composition sampling (Tanner, 1996) and Monte Carlo integration to approximate the point prediction by

$$\hat{\mathbf{y}}_{\text{oos}} \approx \frac{\sum_{t=1}^{T} \mathbf{y}_{\text{oos}}^{(t)}}{T} \ , \tag{16}$$

where $\hat{\mathbf{y}}_{\text{oos}}^{(t)}$ is the t^{th} MCMC sample (out of T total MCMC samples) of the predicted out-of-sample data. That is, we draw $\mathbf{y}_{\text{oos}}^{(t)}$ as a sample from $[\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}^{(t)}]$ at every MCMC iteration t for $t = 1, \ldots, T$ and then average them.

The procedure we have just described provides a way to obtain Bayesian point predictions, but it does not directly accommodate uncertainty pertaining to a score function. As it turns out, the log predictive density $\log[\mathbf{y}_{\text{oos}}|\mathbf{y}]$ is a local and proper scoring function that is appropriate for Bayesian model validation (Gelman et al., 2014 b). In the situation where we have actual out-of-sample data $\mathbf{y}_{\text{oos}}$, then we could just compute

$$\log\left(\frac{\sum_{t=1}^{T}[\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}^{(t)}]}{T}\right) \ , \tag{17}$$

using MCMC samples $\boldsymbol{\theta}^{(t)}$, as a Monte Carlo integral representation of the score function

$$\log[\mathbf{y}_{\text{oos}}|\mathbf{y}] = \log \int [\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \; . \tag{18}$$

This score can then be used to rank all models in the set $\mathcal{M}$ and find the one that yields the best predictions. Out-of-sample validation is almost as efficient as simply fitting the individual models because it only requires the additional calculation of $[\mathbf{y}_{\text{oos}}|\mathbf{y}, \boldsymbol{\theta}^{(t)}]$ on each MCMC iteration which is a low-order operation. Thus, for large ecological data sets, the out-of-sample validation approach is a very reasonable way to find good predictive models. However, as the out-of-sample size reduces, this validation procedure becomes less stable and thus more sensitive to the set of out-of-sample data.

3.2 Cross-Validation

The concept of cross-validation was developed as a way to increase the stability of validation based on out-of-sample data for smaller sample sizes. Cross-validation is similar to out-of-sample validation in that we exclude a subset of the data ($\mathbf{y}_k$) from the fitting procedure so that the model is unaware of it, and then compute the score based on the excluded data. The problem with choosing a single subset of the data to leave out is that you can only assess predictive ability for those measurements. Thus, it is common to leave out all of the data, but only in small subsets sequentially.

K-fold cross-validation involves grouping the data evenly (or approximately even) into K groups and then using each set of left out data $\mathbf{y}_k$ to compare with the model predictions based on the remaining data ($\mathbf{y}_{-k}$). We then iterate through all groups of data

$\mathbf{y}_k$ for $k = 1, \ldots, K$ and compute component scores which are summed to yield the full cross-validation score for the whole data set

$$\sum_{k=1}^{K} \log \left(\frac{\sum_{t=1}^{T} [\mathbf{y}_k | \mathbf{y}_{-k}, \boldsymbol{\theta}^{(t)}]}{T} \right) . \tag{19}$$

In the case where $K = n$ (n is the sample size), the procedure is often referred to as leave-one-out cross-validation. Leave-one-out cross-validation may be preferable when the sample size is small and there are few observations to use as training data, though the resulting estimate of prediction error becomes less stable as $K \to n$.

In general, the major disadvantage of K-fold cross-validation for Bayesian models is that we are required to refit each statistical model K times to obtain the complete set of out-of-sample predictions. Acquiring $K \times L$ individual model fits may be reasonable for simple models, but for more complicated models that take longer to fit, a K-fold increase in required computing time may not be reasonable. However, despite these challenges, when true predictive ability is the main criterion of interest, cross-validation is still very appealing for model comparison. In fact, it underlies several parsimony-based model comparison methods.

3.3 Conditional Predictive Ordinates

To improve computational tractability for large data and model sets, one could consider the posterior predictive distribution for within-sample data. That is, instead of cross-validation, simply compute the aforementioned predictive score based on the predictive distributions of the data $[y_i|\mathbf{y}]$ for $i = 1, \ldots, n$. The problem with this approach is that the predictive performance of the model will be overestimated because the data are used twice (i.e., once for model fitting and another time for model validation). The overestimation of predictive performance is referred to as "optimism" in the statistics literature and we return to this concept in Section 4.

As a potential remedy, consider the leave-one-out predictive distribution for each observation in a data set

$$[y_i|\mathbf{y}_{-i}] = \int [y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}_{-i}]d\boldsymbol{\theta} \ . \tag{20}$$

This quantity (20) is referred to as the conditional predictive ordinate (CPO$_i$; Geisser, 1993) and represents the probability (or density) of the observation y_i when the model is fit without that observation. Thus, large CPO$_i$ values correspond to very likely observations under the current model, whereas small CPO$_i$ indicates outliers and/or high-leverage observations (Pettit 1990). In principle, the computation of CPO would require a true cross-validation involving an n-fold iterative model fitting scheme. Fortunately, CPO can be approximated easily within an MCMC algorithm for model fitting as the harmonic mean of the predictive distributions evaluated at the MCMC values for the parameters $\boldsymbol{\theta}$,

$$\mathrm{CPO}_i \approx \frac{T}{\sum_{t=1}^{T}[y_i|\boldsymbol{\theta}^{(t)}]^{-1}} \ , \tag{21}$$

where $t = 1, \ldots, T$ represent the MCMC iterations. A summary statistic of these individual CPO values, such as $-\sum_i \log(\mathrm{CPO}_i)$, then provides an overall measure of predictive performance. Notice the similarity in expressions for the sum of the logged CPO values and the log predictive score (19) described in the previous Section. In terms of appropriateness for model selection, the CPO involves a harmonic mean, which yields a numerically unstable estimator in practice, but software can often be constructed to flag problematic cases (Held et al., 2010).

3.4 Willow Tit Occupancy: Model Validation

Suppose that we are now interested in comparing the 4 occupancy models we introduced in Section 2 in terms of their predictive ability. We do not have an auxiliary source of out-of-sample data to use for model validation, but we can employ Bayesian cross-validation and also compute the $-\sum_i \log(\text{CPO}_i)$ statistic based on (21) to compare the information about predictive ability using each of these methods.

We used 10-fold Bayesian cross-validation (i.e., $K = 10$) due to the moderate sample size and computed the scoring function discussed in (19) as

$$-2\sum_{k=1}^{10} \log\left(\frac{\sum_{t=1}^{T} \text{Binom}(\mathbf{y}_k|\mathbf{J}_k, p^{(t)}\mathbf{z}_k^{(t)})}{T}\right) , \tag{22}$$

where, $p^{(t)}$ and $\mathbf{z}_k^{(t)}$ are MCMC samples arising from model fits not including observations $\mathbf{y}_k$ and the negative two is multiplied merely for convenience (so that small scores are better and to compare with other model selection criteria later). Thus, the inner sum in (22) is over the MCMC iterations from a single fold of the validation procedure and the outer sum is over the K folds. We obtained 160,000 MCMC iterations to fit each model (in each fold), discarding the first 16,000 as burn-in. To illustrate the computational gains achieved using contemporary parallel programming methods we performed the cross-validation using both non-parallel and parallel algorithms. The non-parallel algorithm (i.e., a single loop over the K folds) required approximately 1 hour, whereas the parallel algorithm required over an order of magnitude less computing time at approximately 5.7 minutes. Similarly, it required 1.4 minutes to compute the CPO statistics in parallel (but 5.7 minutes in sequence). All computation was performed on a desktop workstation with two 2.93 GHz 6-Core processors and 32 GB of RAM; we note that new laptops have individual processors that are substantially faster, but parallel computing is still more efficient on the desktop we used with its many cores. All MCMC algorithms were coded natively in R (R Core Team, 2013) and the R package 'snowfall' (Knaus, 2013) was used for parallel computing.

In Table 4 we can see that the Bayesian cross-validation score generally agrees with CPO in that the two models with elevation as a covariate (i.e., M_2 and M_4) out-perform the null model (M_1) and model with only an intercept and forest as a covariate (M_3; note also that lower scores are better). The null model performs the worst based on the cross-validation score, while the two models with elevation are nearly equivalent in terms of prediction. CPO indicates that the null model may be slightly better at prediction than the model with only forest as a covariate (i.e., M_3), however, given that cross-validation evaluates predictive performance based on out-of-sample data, we might be skeptical of these CPO results for the worst performing models. This potential discrepancy between cross-validation and CPO is part of the sacrifice we make when computation time is limited.

4 STATISTICAL REGULARIZATION AND INFORMATION CRITERIA

The assessment of a set of models in terms of their predictive ability has been a central theme in the development of information criteria. However, information criteria involve specific approaches to model selection that fall under the much broader umbrella of statistical regularization. This concept of regularization, though used on a daily basis in ecology, does not appear to be widely recognized. However, regularization reveals numerous theoretical and practical connections among model selection and multimodel inference paradigms. Specifically, regularization links Bayesian and non-Bayesian approaches to model selection and here we describe how this linkage occurs. We begin by presenting the basic regularization concept, showing how it has been used traditionally in the non-Bayesian context (Section 4.1). We then describe how regularization is inherently Bayesian (Section 4.2) and highlight a few explicitly Bayesian approaches for doing it (e.g., the Bayesian Lasso in Section 4.2.2).

The term "regularization" refers to the use of an external regulator that constrains the results of an optimization problem (note that the term "regulator" is borrowed here from physics but is not commonly used in statistics, though it is perhaps more intuitive). In statistical terminology, the optimization problem could be a likelihood that needs maximizing or a posterior distribution that needs exploring (perhaps via MCMC). In the broader decision theoretic context, we might refer to a negative log-likelihood more generically as a loss function; that is, a function that expresses the "loss" incurred by inadequately estimating parameters of interest. In certain cases, the loss function may have too much freedom to be useful for inference and thus an external constraint can help make it useful.

In placing this concept of regularization in a formal statistical framework for decision making, or parameter estimation, consider the generic expression

$$L(\mathbf{y}, \boldsymbol{\theta}) + r(\boldsymbol{\theta}, \boldsymbol{\gamma}) \ , \tag{23}$$

where $L(\mathbf{y}, \boldsymbol{\theta})$ represents the loss, a function of both knowns ($\mathbf{y}$) and unknowns ($\boldsymbol{\theta}$) and, though it is related, should not to be confused with a likelihood (which we label $[\mathbf{y}|\boldsymbol{\theta}]$). The function $r(\boldsymbol{\theta}, \boldsymbol{\gamma})$ in (23) represents the regulator or constraint on the unknowns $\boldsymbol{\theta}$. The regulator function r may also depend on some other variables $\boldsymbol{\gamma}$ that may or may not be related to the loss function or its components. There are other ways to express the loss and regulator relationship, but the expression in (23) is perhaps the most common. Statistical inference can now be obtained by minimizing the joint function (23) with respect to $\boldsymbol{\theta}$, and perhaps $\boldsymbol{\gamma}$, if not already known. The primary advantage of regularization is that it can yield improved inference, often reducing the variance of estimates and increasing the accuracy of predictions. Though not often discussed in the ecological literature, this concept of regularization is quite common in many areas of statistics and machine learning (Hastie et al., 2009). As we will see in the next sections, regularization also underlies the dominant model selection approaches used in ecology and has direct ties with Bayesian statistics.

4.1 Traditional Regulator: The Penalty

To make the concept of regularization more concrete, we place it in the context of classical non-Bayesian regression modeling. That is, consider the linear model

$$y_i \sim \mathrm{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \ , \tag{24}$$

for $i = 1, \ldots, n$, where the "unknowns" are the regression coefficients β_0 and $\boldsymbol{\beta}$. For now, assume the error variance σ^2 is known, but note that it need not be in general. If our goal is to find estimates of β_0 and $\boldsymbol{\beta}$, then the loss function for this optimization problem is proportional to the negative log-likelihood $L(\mathbf{y}, \beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2$. Now consider the regulator function $\gamma_1 \sum_{j=1}^{p} |\beta_j|^{\gamma_2}$, called the "penalty" in the statistical literature, such that the optimization problem from (23) becomes

$$\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \gamma_1 \sum_{j=1}^{p} |\beta_j|^{\gamma_2} \ , \tag{25}$$

where p corresponds to the dimension of $\boldsymbol{\beta}$ (i.e., the number of covariates in the model), γ_1 is often referred to as the penalization or bandwidth parameter (in the statistics literature, λ is often used instead of γ_2; we avoid the λ notation here to reduce any confusion with the leading eigenvalue of a Leslie matrix in demographic modeling), and the exponent γ_2 is the chosen degree of the "norm." Note that the penalty is commonly written using norm notation, that is, $||\boldsymbol{\beta}||_{\gamma_2} \equiv \sum_{j=1}^{p} |\beta_j|^{\gamma_2}$ (referred to as the L_{γ_2} norm for a specific value of γ_2). The parameters γ_1 and γ_2 control the amount and type of regularization that occurs in the estimation problem. Although the parameters γ_1 and γ_2 are sometimes chosen only implicitly, based on adherence to a particular philosophical underpinning, there seems to be greater variety in the rationale and practical choices for γ_1 than for γ_2. We discuss commonly used choices for γ_2 next.

4.1.1 Ridge Regression

So-called "ridge regression" is a direct application of the above optimization problem (25) where the parameter $\gamma_2 = 2$ is used in the penalty term. In this case, we seek to minimize

$$\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \gamma_1 \sum_{j=1}^{p} \beta_j^2 \tag{26}$$

with respect to the regression coefficients β_0 and $\boldsymbol{\beta}$ given a certain value for the penalty parameter γ_1. If $\gamma_1 = 0$ then the negative log-likelihood is not penalized and the resulting estimated coefficients will be the maximum likelihood estimates (MLEs). However, as γ_1 increases, it will "shrink" the estimated coefficients $\boldsymbol{\beta}$ toward zero when (26) is minimized as a trade-off between maximizing the likelihood and meeting the constraint. This is why regularization methods in the maximum likelihood setting are commonly referred to as "penalized" or "shrinkage" methods. The shrinkage of $\boldsymbol{\beta}$ can be incredibly useful in parameter estimation and prediction.

In parameter estimation, shrinkage induces an increasing bias in $\hat{\boldsymbol{\beta}}$ with increasing γ_1 but simultaneously reduces the variance of $\hat{\boldsymbol{\beta}}$. Thus, in ridge regression, we accept a small amount of bias in our estimation of $\boldsymbol{\beta}$ in return for a potentially large reduction in variance. The reduction in variance of $\hat{\boldsymbol{\beta}}$ also decreases prediction error, providing improved prediction accuracy. More complex models provide an excellent fit to within-sample data but are poor predictors of out-of-sample data. Shrinking model parameters toward zero reduces effective model complexity thereby improving our ability to predict out-of-sample data.

These features of ridge regression are undoubtedly desirable, but may overshadow one of the most useful aspects of the regularization: alleviation of the effect of multicollinearity in the covariates (e.g., Graham, 2003). When columns of our “design matrix” $\mathbf{X}$ are correlated with each other, the associated coefficients $\boldsymbol{\beta}$ have to compete for the overall effect on the response variables $\mathbf{y}$. This competition causes the coefficient estimates $\hat{\boldsymbol{\beta}}$ to offset each other, forcing some to be very large (positive) and some very small (negative). In cases where significant multicollinearity exists, the penalty term in the optimization problem will shrink these exaggerated parameter estimates back to reasonable values. Thus, in ridge regression, we can use the “full” model including all the variables in $\mathbf{X}$ at once, regardless of how much they are correlated with each other. The alternative approach is to construct a finite model set where no single model contains any two covariates that are correlated beyond a certain threshold (e.g., correlation coefficient $\rho = 0.6$, as advocated by Burnham and Anderson, 2002). This latter approach is a type of discrete regularization, rather than a continuous one such as ridge regression.

There are a few practical considerations in the proper application of regularization methods for regression models. First, notice that we have separated the intercept β_0 from the rest of the regression coefficients $\boldsymbol{\beta}$ in (25). We isolate β_0 because we do not wish to shrink the general mean of the regression model to zero, rather, only the coefficients that interact with covariates. Second, it is advisable to standardize the covariates in $\mathbf{X}$ prior to analysis (i.e., subtract the mean and divide by the standard deviation). This standardization of covariates allows us to use a single penalty parameter γ_1 rather than one for each coefficient β_j so that they do not need to be shrunk differentially. The third consideration is the choice of γ_1, which we discuss in the next section.

4.1.2 Lasso: Least Absolute Shrinkage and Selection Operator

Continuing with the linear regression example (25) used in the previous section, now consider a different regulator function where we set $\gamma_2 = 1$ such that

$$\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \gamma_1 \sum_{j=1}^{p} |\beta_j| \ . \tag{27}$$

This new penalty term ($\gamma_1 \sum_{j=1}^{p} |\beta_j|$) is commonly referred as the “Lasso” or L_1 penalty and induces a markedly different constraint on the optimization problem. The acronym ‘Lasso’ stands for Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996) because the use of an L_1 norm penalty implies a sum of absolute coefficient values. While the L_2 penalty in ridge regression shrinks $\boldsymbol{\beta}$ toward zero nonlinearly (with increasing γ_1), the L_1 Lasso penalty shrinks the coefficients linearly in such a way that they eventually can equal zero exactly in the optimization. Thus, Lasso drops covariates from the model by setting their coefficients to zero. This absolute variable selection concept seems quite familiar to many ecologists who learned about model selection from a traditional perspective. This heuristic familiarity has made the Lasso approach very popular (Dahlgren, 2010).

To summarize, we have now seen that both γ_1 and γ_2 in (25) play important roles in statistical regularization. Given that γ_1 controls the amount of shrinkage induced, it acts as a type of *scale* parameter, while γ_2 controls the form of the shrinkage and could be thought of as a *shape* parameter. For now, we suspect that the choice of γ_2 is more a result of personal preference based on desired inference, but what about γ_1? How should we choose the amount of shrinkage?

Heuristically, we seek inference concerning model parameters that is based on a balance between model fit and predictive ability. Thus, we could treat γ_1 as we do any other model parameter and estimate it simultaneously with the others. The problems with this approach are manifold, but relate to the same basic concept: within-sample data versus out-of-sample data. Even if there is enough information in the data to actually estimate an "extra" model parameter, the fact that within-sample data are being used to learn about γ_1 limits its utility as a regulator. Recall from our discussion of cross-validation, that there are trade-offs in using the same set of data to both fit and validate (i.e., select) models. The primary trade-off is that predictive performance can only truly be assessed using out-of-sample data. Thus, it seems most reasonable to estimate model parameters based on within-sample data and choose regulator parameters based on out-of-sample data.

A strategy employed in many machine learning studies is to optimize the regularized loss function (23) given the within-sample data $\mathbf{y}$ for the first term and use an iterative cross-validation approach to choose γ_1 based on predictive ability of out-of-sample data. In practice, a strategy for the regression model would involve first optimizing (25) using $\gamma_1 = 0$ assigning a cross-validation score, and then incrementally increasing γ_1 over a range of values yielding a set of predictive scores. Given a sufficiently fine range of values for γ_1, we would then choose the regularized model yielding the best predictive score. In the case of ridge regression, our inference would consist of a full set of coefficient estimates $\hat{\boldsymbol{\beta}}$ that are properly shrunk to provide the best predictions of out-of-sample data. For Lasso, we would obtain a subset of non-zero coefficient estimates that have been shrunk according to the L_1 penalty, and the remaining coefficients would be zero (i.e., no longer in the final model). In either case, we will obtain a justifiably parsimonious model that is better at prediction than the unpenalized full model. Another advantage is that we did not have to do prior variable elimination based on highly collinear covariate pairs.

[Figure 3 Here]

Despite the many advantages to classical regularization, there are also several disadvantages. Aside from the somewhat *ad hoc* and subjective feel of the procedure, these methods are based on optimization and they yield point estimates for the model parameters of interest, but learning about the uncertainty of $\hat{\boldsymbol{\beta}}$ is not necessarily trivial or even possible in some cases. Finally, because we may want to rely on out-of-sample data to choose appropriate regulator parameters ($\boldsymbol{\gamma}$), this can dramatically increase the computational requirements of cross-validation-based regularization.

4.1.3 Akaike's Information Criterion

Continuing in a non-Bayesian context, we now explain how information criteria fit into the regularization concept. Statistical regularization is appealing for the reasons discussed in the previous section, but for many ecologists, the increased computational burden and need to select regulator parameters can be daunting. Enter the information criterion approach to statistical regularization. The general idea behind information criteria is that we choose a scoring function *a priori* that will be used to "score" each of the models based on the balance of fit using the within-sample data and parsimony (or overall predictive ability; Gneiting, 2011). Not surprisingly, most commonly used information criteria take the same form as the previously introduced regularization expression (23). For example, in the linear regression class of models, Akaike's Information Criterion (AIC) takes the form of (25) with regulator parameters $\gamma_1 = 2$ and $\gamma_2 = 0$ such that the penalty is $2\sum_{j=1}^{p}|\beta_j|^0 = 2p$.

The L_0 norm used in AIC implies that the shrinkage is only based on the number of parameters rather than the parameter values themselves. This implication is useful because each model in the model set can be fit independently and then *post hoc* scored using AIC (lower AIC implying better predictive ability of the model). However, we must be careful to avoid inducing obvious bias in the estimates by choosing a model set such that no single model contains correlated covariates because the penalty cannot provide feedback to the estimation of the parameters themselves.

AIC provides the same regularization as leave-one-out cross-validation under certain conditions (Stone, 1977). We find this a very appealing result on first glance because it could dramatically reduce the computational burden in finding a good predictive model. However, upon closer inspection, we find that the result only holds in linear Gaussian settings (i.e., regression models with additive normal errors) and under the assumption that the "true" model is in the model set being considered. This latter assumption (i.e., truth in the model set) seems to conflict with one of the main advantages of AIC extolled by proponents. Still, empirically, AIC seems to perform well in situations where it can be used (Hastie et al., 2009). For Bayesians, AIC (being a function of maximum likelihood estimates) does not appear to have a clear Bayesian interpretation, at least outside of a few contrived situations (as we discuss later in Section 4.2).

The use of an information criterion like AIC requires a compromise: We trade the continuous aspects of model selection using more general regulators (e.g., ridge regression, Lasso) for the reduction in computational burden achieved by avoiding cross-validation.

4.1.4 Bayesian Information Criterion

The so-called Bayesian Information Criterion (BIC; Schwarz, 1978) arises from a different motivation than does AIC and many other regularization methods. AIC is an information criterion that seeks to provide a measure of predictive ability, whereas BIC is distinctly concerned with multimodel inference (Link and Barker, 2006; Gelman et al., 2014 b).

Recall the marginal data distribution $[\mathbf{y}|M_l]$ for model M_l from Section 2 on Bayesian model averaging (10). The marginal data distribution is critical for computing Bayes factors and model probabilities in the Bayesian paradigm. In a maximum likelihood setting, if we consider the loss function to be $-2\log[\mathbf{y}|\hat{\boldsymbol{\theta}}]$, as is assumed with AIC, then we can approximate the marginal data distribution using a Laplace approximation (Ripley, 1996) such that for model M_l

$$\begin{aligned} \text{BIC} &= -2\log[\mathbf{y}|\hat{\boldsymbol{\theta}}, M_l] + log(n)p \\ &\approx -2\log[\mathbf{y}|M_l] \;, \end{aligned} \tag{28}$$

where $\log(n)$ is the natural logarithm of the sample size (or dimension of $\mathbf{y}$) and p is the number of "free" parameters, as before. Note that, for the linear regression model (24), this definition of BIC still retains the general regularization form of (25), but with regulator parameters $\gamma_1 = \log(n)$ and $\gamma_2 = 0$.

The utility of BIC in multimodel inference arises when we exponentiate negative one-half times the BIC (28); normalizing this quantity over all models in the model set $\mathcal{M}$ provides an approximation to the Bayesian model weights (9) described previously. Unfortunately, this approximation only holds when equal prior model weights (i.e.,

$P(M_l) = 1/L$ for $l = 1, \dots, L$) are assumed. Furthermore, because of its reliance on maximum likelihood parameter estimates, BIC does not appear to be inherently Bayesian (despite its name). Finally, BIC can only be used to approximate posterior model probabilities when the Bayes factors are well defined, which is not the case if improper priors are used in the models.

From a classical perspective, there is no clear choice, nor consensus, among statisticians, between AIC and BIC for model selection purposes (Hastie et al., 2009). Each form of automatic regulator has advantages and disadvantages. For example, BIC can be shown to be a consistent model selector (i.e., the oracle property). That is, when the "true" model is in the model set and the data set is sufficiently large, BIC will select the true model, while AIC will select models that are too large in general. On the other hand, for smaller sample sizes, BIC may indicate models that are too parsimonious because $log(n) > 2$ implies more shrinkage from BIC than AIC. Furthermore, BIC is motivated from a model averaging rather than prediction perspective, and thus it may be more justified for approximating Bayesian model weights than for model selection.

4.2 Bayesian Regulator: The Prior

The previous section describes regularization from a classical perspective, where we penalize a statistical optimization problem in such a way that it yields a better predictive model. As we hinted at earlier, the fact that the classical regularization approach seems to "work" is encouraging, but its lack of formality brings up a set of new questions (e.g., What type of regulator function to use? How much shrinkage is too much?). Furthermore, on the surface, the classical regularization methods do not appear to be able to accommodate uncertainty about the parameters or regulator function. For ecologists using Bayesian models, what is the analog to regularization in the Bayesian setting?

4.2.1 Natural Bayesian Shrinkage

The analog to regularization in the Bayesian setting is simply the Bayesian model itself! To see this, consider the linear regression example (24) used in the previous section, but now, we specify priors for the unknown model parameters $\boldsymbol{\beta}$ such that the model itself is specified as

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \\ \boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}, \sigma_\beta^2 \mathbf{I}) \ , \end{aligned} \tag{29}$$

where, for illustrative purposes, we assume the intercept β_0 and variance parameter σ^2 are fixed and known for now. The posterior distribution for $\boldsymbol{\beta}$ is then easily shown to be

$$\begin{aligned} [\boldsymbol{\beta}|\mathbf{y}] &\propto [\mathbf{y}|\boldsymbol{\beta}][\boldsymbol{\beta}] \\ &\propto \prod_{i=1}^{n} \mathrm{N}(y_i|\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \prod_{j=1}^{p} \mathrm{N}(\beta_j|\mu_j, \sigma_\beta^2) \\ &\propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right) \exp\left(-\frac{1}{2}\frac{\sum_{j=1}^{p}(\beta_j - \mu_j)^2}{\sigma_\beta^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sigma^2} + \frac{\sum_{j=1}^{p}(\beta_j - \mu_j)^2}{\sigma_\beta^2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \frac{\sigma^2}{\sigma_\beta^2}\sum_{j=1}^{p}(\beta_j - \mu_j)^2\right)\right) \ . \end{aligned} \tag{30}$$

If we let $\mu_j = 0$ for all $j = 1, \ldots, p$, and reparameterize the ratio of variances such that $\gamma_1 = \sigma^2/\sigma^2_\beta$ in the last expression of (30), then we arrive at the exact same regularization expression used in ridge regression (26) in the inner parentheses of our posterior distribution for $\boldsymbol{\beta}$ (30). Thus, by reducing our prior variance for the regression coefficients, we increase the effective regulator parameter γ_1 and induce the same sort of shrinkage on $\boldsymbol{\beta}$ as in ridge regression, but in a formal Bayesian probability framework. In fact, one could say that we are always doing a form of regularization in Bayesian statistics because the prior acts as the regulator. Given that the Bayesian posterior provides a rigorous framework for regularization, it could be argued that other classical forms of regularization are inherently Bayesian, or at least Bayesian in spirit.

Regardless of the interpretation of the regulator, as a non-Bayesian penalty or as a Bayesian prior, we can enjoy the same benefits of regularization from either perspective. However, the Bayesian perspective makes it clear that we are constraining the model parameters with "prior" information such that it assists us in finding a better predictive model. We are often taught that the Bayesian prior should either be chosen objectively as to minimize the influence on the posterior, or retrospectively, to best represent existing prior knowledge about the parameters. However, the only rule for specifying prior information in a Bayesian model is to not use the within-sample data to choose the prior. The reason for this rule is that it maintains the acyclicity in the Bayesian "graph." Bayesian models are often referred to as directed acyclic graphs because of their conditional specifications such that the data depend on the parameters and the parameters depend on either other parameters or fixed quantities. The acyclic nature of the Bayesian graph guarantees that we can use valid probability statements to learn about the unknown quantities. Interestingly, this rule of “don’t use the data twice” is commonly broken, and the model is referred to as empirical Bayesian in that setting. Empirical Bayesian methods seem to perform well, as does classical regularization, but have much weaker theoretical foundations than fully Bayesian methods. It seems clear that to fit a rigorous Bayesian model we should not use the within-sample data in the likelihood and the prior, but there is no such rule about the use of out-of-sample data to inform the prior. Thus, we could think of the three ways to specify valid priors as 1.) objectively, 2.) retrospectively, and 3.) prospectively. The term “prospective” in this sense implies the use of future data, perhaps collected at the same time as the within-sample data but not used until after (rather than before) the likelihood is specified. This third approach to specifying priors opens up the door for Bayesian cross-validation.

For example, the Bayesian cross-validation procedure for regularization of the regression model might proceed as follows: Specify the model as in (29), fit it for each of the K sets of hold-out data using a vague prior for $\boldsymbol{\beta}$ with mean zero and obtain a predictive score as described in Section 3.2. Choose an incrementally smaller prior variance σ^2_β and repeat the model fitting and cross-validation scoring process. Continue this procedure, using smaller and smaller prior variances until an optimal predictive model is identified (typically via a small score function). Finally, fit the optimal predictive Bayesian regression model using the full data set to obtain desired inference.

The problem arises in the last step of this cross-validation procedure. Once we use the prior (i.e., penalty or regulator) that has been informed by an aggregate of hold-out data, we technically cannot put all of the hold-out data back into the model to fit one last time for final inference in a fully Bayesian paradigm. In this case, the options are: 1.) use the data twice in this way and accept that the procedure is empirical Bayesian, or 2.) use two completely separate datasets, one for training ($\mathbf{y}$) and another for validating ($\mathbf{y}_{\text{oos}}$). Of course, the second option is not always preferable when analyzing data that have already been collected, but in larger data sets or when setting up new studies, collecting two independent datasets for two different purposes allows for fully rigorous Bayesian inference and model selection.

4.2.2 Bayesian Lasso

The previous section illustrates how the standard Bayesian regression model with a Gaussian prior on the coefficients provides a natural mechanism to perform statistical regularization similar to ridge regression, but how can we manipulate the regulator function? The answer is simple in the regression case: We only need to find a prior with the same form as the desired regulator function. For example, to construct a Bayesian regularization that has a penalty similar to the Lasso penalty, we need only find a prior containing an L_1 norm on the parameters. In this case, the Laplace distribution contains the L_1 norm that will impose a Lasso penalty as a prior. That is, consider the same regression data model, but with a new prior for $\boldsymbol{\beta}$ such that

$$\begin{aligned} y_i &\sim \mathrm{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \\ \beta_j &\sim \mathrm{Laplace}(\mu = 0, \sigma^2_\beta) \propto \exp\left(-\frac{|\beta_j|}{\sqrt{\sigma^2_\beta}} \right) , \end{aligned} \tag{31}$$

for $j = 1, \ldots, p$ where β_j are independent *a priori*. Park and Casella (2008) propose a similar prior for $\boldsymbol{\beta}$, as well as more standard priors for β_0 and σ^2 and dub it "The Bayesian

Lasso.” In fact, they go a step further and carefully specify a prior for a transformation of the regulator parameter that enables them to construct a fully conjugate MCMC algorithm for fitting the model. Unlike in a Metropolis-Hastings MCMC algorithm, the resulting Gibbs sampler requires no tuning of any parameters (Kyung et al. 2010). Thus, it is nearly as computationally efficient to fit the Bayesian Lasso regression model as it is the standard Bayesian regression model. Of course, Bayesian cross-validation could also be used in this scenario and would likely yield better out-of-sample predictive performance, but would also require substantially more computational effort.

Finally, after seeing the connection between Bayesian priors and regulator functions, one might wonder what sort of prior yields an AIC penalty? Following the same approach described in the Bayesian Lasso (31), it appears that the implicit AIC prior for each coefficient is $[\beta_j] \propto \exp(-|\beta_j|^0)$, such that the joint prior distribution for $\boldsymbol{\beta}$ is $[\boldsymbol{\beta}] \propto \exp(-p)$.

4.3 Willow Tit Occupancy: Bayesian Regularization

In applying Bayesian regularization to the willow tit occupancy model, we first remind the reader that the model already contains a natural regularization mechanism: the prior for $\boldsymbol{\beta}$. Recall the process component of the hierarchical occupancy model from (3)

$$v_i \sim \mathrm{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, 1) \; , \tag{32}$$

and prior from (6)

$$\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}_\beta, \sigma^2\mathbf{I}) \; . \tag{33}$$

Notice that if we standarize the covariates to have mean zero and variance one then we can reasonably set the prior mean $\boldsymbol{\mu}_\beta = \mathbf{0}$. In this case, the full-conditional distribution for $\boldsymbol{\beta}$ becomes

$$[\boldsymbol{\beta}|\cdot] \propto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{n}(v_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \frac{1}{\sigma_\beta^2}\sum_{j=1}^{p}\beta_j^2\right)\right) \tag{34}$$

as was demonstrated for the regression model (30). Thus, this full-conditional distribution for $\boldsymbol{\beta}$ has the same form as the general regularization expression (25) and the hyperparameter σ_β^2 serves as the regulator or shrinkage parameter, where $\gamma_1 = 1/\sigma_\beta^2$. In other words, the smaller we make the prior variance, the stronger the penalty in the regularization. The strategy is to explore the space of σ_β^2 for an optimal value that provides the best predictive model according to the score function of choice. To find the optimal penalty, we can explore the space of σ_β^2 using a grid search (i.e., try a range of n_β total values for σ_β^2) and compare scores based on cross-validation. This cross-validation approach requires $K \times n_\beta$ separate model fits, resulting in a potentially unreasonable amount of required computational time. For example, a 10-fold cross-validation, at 1.4 minutes per model fit and $n_\beta = 24$ dimensional grid search would require 5.6 hours to implement. However, using 24 processors in parallel, the required time could be reduced to under an hour on a high-performance desktop workstation. The three easy ways to reduce computation time are to 1.) use more processors (e.g., a high-performance computing facility), 2.) decrease the number of folds in the cross-validation (e.g., an n-fold cross-validation for the above example would require almost 5 days in sequence, but only a few hours in parallel) and 3.) use a lower resolution grid search. The latter will require fewer model fits on the same machine, but will reduce the accuracy of the optimization.

We wouldn't expect Bayesian regularization to dramatically increase predictive ability for the simple willow tit occupancy model because the two covariates (elevation and forest) are relatively uncorrelated (i.e., correlation≈ 0.12) and the sample size ($n = 200$) is large relative to the number of unknown parameters. However, to demonstrate the regularization approach, we use the full model for the willow tit data with one intercept and two regression coefficients associated with the occupancy probability (M_4). We then perform a grid search over 24 values for σ^2_β, implying a prior that ranges from precise ($\sigma^2_\beta = 0.01$) to vague ($\sigma^2_\beta = 2.25$).

We used the log posterior predictive score for 10-fold cross-validation introduced earlier (22). The complete 10-fold cross-validation at each value of σ^2_β, with model fits based on 160,000 MCMC iterations (discarding 16,000 as burn-in), took approximately 24 minutes with parallel computing.

We found that the optimal prior variance for prediction occurs at $\sigma^2_\beta = 1.02$; this is less than half of the variance we would typically use in a vague prior scenario for the occupancy model. In Figure 3 we see the posterior means for $\boldsymbol{\beta}$ taper toward zero as σ^2_β decreases. At the optimal level of regularization, the predictive score was 478.4, yielding a model that predicts as well as M_2 (the elevation only model) but uses both covariates. Notice also that the cross-validation score function increases more sharply away from the optimum as σ^2_β decreases toward zero. This effect indicates that the null model (i.e., occurring at $\sigma^2_\beta = 0$) performs substantially worse than the full model (i.e., occurring at $\sigma^2_\beta = 2.25$), a result similar to that found in the former cross-validation of the discrete model set (Table 4).

[Figure 3 Here]

4.4 Deviance Information Criterion

We have seen that a natural framework for regularization in the Bayesian context already exists and can be used in conjunction with out-of-sample data to help select an appropriate penalty. However, the classical information criteria were developed, at least in part, to alleviate the need for cross-validation and seem to perform quite well in many settings. Is there a Bayesian equivalent?

Spiegelhalter et al. (2002) proposed the Deviance Information Criterion (DIC), which has a similar form as other information criteria, in that it contains a loss function plus a penalty or regulator function. The loss function is chosen to be the deviance

$$D(\boldsymbol{\theta}) = -2\log[\mathbf{y}|\boldsymbol{\theta}] \ , \tag{35}$$

as in most other information criteria, but in order to be similar to AIC or BIC the penalty needs to incorporate the number of free parameters as a measure of model complexity. Recall that, even in the simplest Bayesian models, most parameters are constrained in some way by their priors. Furthermore, in hierarchical Bayesian models, we may have numerous latent state variables that are technically unknown but are also highly constrained by both the likelihood and prior. Thus, one crucial issue in the development of a truly Bayesian criterion is the specification of an "effective" number of parameters, say p_D. A further complication is that maximum likelihood point estimates are used to compute AIC and BIC, but this concept of maximum likelihood is only meaningful under certain situations in the Bayesian context. Thus, we can use a Bayesian point estimate, the posterior mean, in lieu of the MLE in DIC:

$$\begin{aligned}\text{DIC} &= -2\log[\mathbf{y}|E(\boldsymbol{\theta}|\mathbf{y})] + 2p_D \\ &= \hat{D} + 2p_D \ , \qquad (36)\end{aligned}$$

where, the deviance evaluated at the posterior mean for $\boldsymbol{\theta}$ is commonly written as $\hat{D}$.

To arrive at a measure of model complexity, Spiegelhalter et al. (2002) consider the difference in the deviance calculated two different ways: 1.) posterior mean deviance and 2.) deviance computed at the posterior mean of the parameters. That is, the effective number of parameters was originally defined as

$$p_D = \bar{D} - \hat{D} \ , \qquad (37)$$

such that the posterior mean deviance is

$$\begin{aligned}\bar{D} &= E_{\boldsymbol{\theta}|\mathbf{y}}(-2\log[\mathbf{y}|\boldsymbol{\theta}]) \\ &= \int -2\log[\mathbf{y}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \ . \qquad (38)\end{aligned}$$

In the case of linear regression, with vague priors on the regression coefficients, the effective number of parameters p_D approaches the number of coefficients p. Thus, the popularity of DIC has been a result of its similarity to AIC, its simplicity, and its ease of calculation using MCMC samples. There are only two quantities that need to be computed for DIC: The deviance evaluated at the posterior mean of the parameter set $\hat{D}$, which is as trivial as the deviance calculation in AIC, and the posterior mean deviance, which can be embedded into an MCMC algorithm with one or two lines of code.

For many Bayesian models (which we describe in the next Section), DIC can be used for ranking models and finding those that should predict better than others, just as AIC would. DIC addresses the issue of model complexity and in many cases yields results quite similar to AIC. A common question is whether DIC can be used for Bayesian model averaging? That is, if one follows the AIC-based guidance of Burnham and Anderson (2002), and calculates $w_j = e^{-\Delta \mathrm{DIC}_j/2}/\sum_l e^{-\Delta \mathrm{DIC}_l/2}$, where $\Delta\mathrm{DIC}_j$ represents the difference of DIC for model j and the minimum DIC across all models in the model set, do these weights w_j approximate posterior model probabilities? Despite the fact that this approach is used occasionally, the answer has not been justified in the literature. Link and Barker (2006) make a strong case for the use of BIC to approximate posterior model probabilities and perform a small set of empirical comparisons between AIC, BIC, and DIC model weighting schemes, but the theoretical foundation for Bayesian model averaging using DIC is much weaker.

4.4.1 Modified DIC

Despite its convenience, DIC has several limitations, notable among them are the potential for poorly estimating model complexity (p_D), inappropriateness with mixture models, and the lack of a direct connection with predictive ability. We elaborate on some of these these issues with conventional DIC before discussing some attractive alternatives.

There have been many alternative specifications for the effective number of parameters p_D (37), which is sometimes referred to as model complexity, or degrees of freedom, in the statistical literature. For example, Plummer (2002) suggests that a more appropriate measure of model complexity can be computed by averaging

$$\log\left(\frac{[\tilde{\mathbf{y}}^{(1,k)}|\boldsymbol{\theta}^{(1,k)}]}{[\tilde{\mathbf{y}}^{(2,k)}|\boldsymbol{\theta}^{(2,k)}]}\right) \tag{39}$$

over all MCMC samples (i.e., $k = 1, \ldots, K$), where $\tilde{\mathbf{y}}^{(1,k)}$ and $\tilde{\mathbf{y}}^{(2,k)}$ are two independent posterior predictive realizations of the data arising from two different chains (for $\boldsymbol{\theta}^{(1,k)}$ and $\boldsymbol{\theta}^{(2,k)}$) based on separate model fits. This version of model complexity (39) arises as an estimate of the expected Kullback-Leibler divergence between predictive distributions at two values for $\boldsymbol{\theta}$ (Plummer, 2002). Unfortunately, Plummer (2008) later indicates that the average of (39) may only be an appropriate penalty when the sample size is very large (i.e., $n \rightarrow \infty$). Plummer (2008) also recommends an alternative estimator for model complexity with better properties, but its calculation requires n separate model fits, which puts it on par with cross-validation, thus reducing the appeal of DIC in terms of computational efficiency. Overall, it appears that DIC (36) is most appropriate as a model selection criterion in linear models with independent data (conditional on $\boldsymbol{\theta}$) where the p_D is much smaller than n. Thus, DIC is good for comparing Bayesian versions of the same classes of models that AIC is good for comparing.

Several others have suggested that DIC is not appropriate for model selection with mixture models or missing data models (e.g., Spiegelhalter et al. 2002; Celeux et al. 2006; Plummer 2008). Zero-inflated models comprise the largest and most heavily used class of models in wildlife ecology (i.e., capture-recapture and occupancy models) and are a form of mixture model (Martin et al. 2005). The original version of DIC is thus not suitable for comparing zero-inflated models. Celeux et al. (2006) provide several suggestions that could be used as an alternative to the standard DIC for mixture models, but ultimately they do not recommend any of them as a gold standard. However, one of these modified versions of DIC was also discussed earlier by Richardson (2002) and lacked a theoretical justification until recently (Watanabe, 2010). Celeux et al. (2006) numbered this information criterion DIC$_3$, and we discuss it next.

4.5 Watanabe-Akaike Information Criterion

Aside from the aforementioned caveats, DIC is a useful information criterion in the parametric Bayesian modeling context when prediction is of primary importance. However, DIC does not best represent the actual Bayesian predictive procedure. To arrive at predictions, the Bayesian approach is to find and summarize the posterior predictive distribution (14). In computing DIC (36) the posterior predictive distribution is not needed. This seems to be a mismatch between the type of inference desired and the tool used to obtain it.

Along the same lines of reasoning we used in the previous Section on out-of-sample validation, for Bayesian model comparison based on predictive ability, we should seek a statistic that considers the log posterior predictive distribution for new data $\tilde{\mathbf{y}}$

$$\log[\tilde{\mathbf{y}}|\mathbf{y}] = \log \int [\tilde{\mathbf{y}}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \ . \tag{40}$$

The quantity in (40) is stochastic because $\tilde{\mathbf{y}}$ is assumed to be unknown (but not so in true out-of-sample validation scenarios; hence the change in notation from $\mathbf{y}_{\text{oos}}$ to $\tilde{\mathbf{y}}$), therefore a common technique in the development of most information criteria is to then consider the mean of (40) over $\tilde{\mathbf{y}}$

$$E_{\tilde{\mathbf{y}}}(\log[\tilde{\mathbf{y}}|\mathbf{y}]) = \int \log \int [\tilde{\mathbf{y}}|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta}[\tilde{\mathbf{y}}]d\tilde{\mathbf{y}} \ , \tag{41}$$

which is impossible to compute directly because the true distribution of the new data $[\tilde{\mathbf{y}}]$ is unknown. Thus, in finding an estimator of mean log posterior predictive score, Richardson (2002), Celeux et al. (2006), and Watanabe (2010) propose the log point-wise predictive score

$$\log \prod_{i=1}^{n} [y_i|\mathbf{y}] = \sum_{i=1}^{n} \log \int [y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} \ , \tag{42}$$

where Monte Carlo integration can be used to compute the integral (Gelman et al. 2014 b). There are two issues with the score in (42): 1.) the product representation of the posterior predictive distribution implies that the data are independent (conditioned on $\boldsymbol{\theta}$) and 2.) it relies completely on the observed data $\mathbf{y}$ rather than the new data $\tilde{\mathbf{y}}$. The first issue suggests that the score should not be used with models containing dependence in the data (e.g., spatial and time series models). The latter issue implies that (42) will be optimistic in its predictive score for a given model because the within-sample data are being used twice. As in DIC, the amount of optimism with this score (42) can be expressed as the effective number of parameters p_D (Watanabe, 2010). Thinking of the effective number of parameters p_D in this way is not intuitive because most ecologists have been trained to view the penalty in AIC as p, the actual number of parameters. In fact, p in that sense is really a measure of model complexity that arises naturally in the derivation of many information criteria. Thus, it is helpful to think of p_D as a measure of model complexity rather than strictly a count of the model parameters.

Gelman et al. (2014 b) present two possible estimates for p_D,

$$p_{D,1} = 2\sum_{i=1}^{n}\left(\log \mathrm{E}_{\boldsymbol{\theta}|\mathbf{y}}[y_i|\boldsymbol{\theta}] - \mathrm{E}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}])\right) , \tag{43}$$

and

$$p_{D,2} = \sum_{i=1}^{n} \mathrm{var}_{\boldsymbol{\theta}|\mathbf{y}}(\log[y_i|\boldsymbol{\theta}]) , \tag{44}$$

but prefers $p_{D,2}$ for its relationship with leave-one-out cross-validation. As with DIC, we can use Monte Carlo integration to approximate $p_{D,2}$ by computing the sum of the MCMC sample variances of $\log[y_i|\boldsymbol{\theta}^{(k)}]$ (sample variance computed over $k = 1, \ldots, K$ MCMC samples) over the observations y_i for $i = 1, \ldots, n$.

The Watanabe-Akaike Information Criterion can then be defined as -2 times the log point-wise predictive score plus the estimated optimism

$$\mathrm{WAIC} = -2\sum_{i=1}^{n}\log\int[y_i|\boldsymbol{\theta}][\boldsymbol{\theta}|\mathbf{y}]d\boldsymbol{\theta} + 2p_{D,2} , \tag{45}$$

with both elements in the sum approximated using MCMC samples at no extra computational cost beyond that required for calculating DIC (Watanabe, 2013). The addition of the estimated optimism in (45) serves as a bias correction in estimating posterior predictive accuracy similar to that of AIC and DIC, even though we have not mentioned it until now. The term "optimism," which is often used in the statistical literature, is merely another word for regulator or penalty.

This new criterion enjoys many benefits. Among them are the fact that WAIC is based on the posterior predictive distribution and is fully Bayesian, but yields the same results as DIC in linear Gaussian models with uniform priors. Furthermore, unlike DIC, WAIC is valid in both hierarchical and mixture models (Watanabe, 2013). Also, unlike DIC, the effective number of parameters calculated using $p_{D,2}$ in (44) will always be positive. In $p_{D,2}$, a parameter gets counted as a 1 if all of the learning we gain about it comes from the likelihood. Conversely, a parameter counts as a zero in the calculation of $p_{D,2}$ if the learning comes entirely from the prior. To figure out the correct proportion of each parameter to count, WAIC needs to use the data (like in DIC) to compute the optimism $p_{D,2}$. This is essential in the Bayesian context where we regularly use hierarchical structures with strong interdependencies and informative priors.

Overall, WAIC seems very appealing, however, the main disadvantage is substantial depending on the area of application: its calculation relies on an independence assumption of the data given the parameters. This assumption is regularly violated in spatial models where dependence among the data is one of the key features being modeled. Ando and Tsay (2010) provide a way to relax the independence assumption, but the resulting criterion requires numerous model fits which eliminates one of the key practical benefits of WAIC (Gelman et al., 2014 b).

4.6 Posterior Predictive Loss

In a similar spirit as that motivating WAIC, and in contrast with CPO, another approach to prediction-based model choice was presented by Laud and Ibrahim (1995) and later justified by Gelfand and Ghosh (1998). This approach, referred to as "posterior predictive loss," considers prediction from a decision theoretic perspective. Understanding this approach requires a familiarity with statistical decision theory, which we describe briefly here, referring the interested reader to more comprehensive references (e.g., Berger, 2006; Vehtari and Ojanen, 2012) for further details.

Statistical decision theory provides a rigorous framework for the decision making process in the presence of data and uncertainty (Berger, 2006). The phrase "decision making process" is quite general, encompassing decisions like choices of alternatives for management, but also including a justification for parameter estimation and prediction. In fact, behind every statistical estimator lies a set of implicit or explicit decision theoretic assumptions. A formal decision theory exists in both the classical and Bayesian realms, though Berger (2006) makes a compelling case for the completeness of the Bayesian decision theory.

In essence, a Bayesian decision theory involves three main concepts: 1) a loss function, 2) an "action" or decision, and 3) a posterior risk function. The loss function is a mathematical expression of the loss incurred if a certain decision is made and the posterior risk function is the loss averaged over the posterior distribution for the unknown quantities of interest. Thus, risk is a version of loss that has accounted for our uncertainty about the study system. The statistical literature refers to the decision minimizing the posterior risk as a "Bayes rule" (Lehmann and Casella, 1998).

For example, suppose we are interested in estimating a parameter θ given data $\mathbf{y}$. In the case of parameter estimation, the "decision" is actually just a point estimator of θ. A point estimate $\hat{\theta}$ that minimizes our risk seems desirable, thus the Bayes rule for point estimation is called a Bayes estimator. To find this Bayes estimator, we simply define a function $L(\mathbf{y}, \theta)$ that suitably represents the loss we incur for poorly estimating θ and minimize its average with respect to the posterior distribution. The value for θ that minimizes the posterior risk $\hat{\theta}$ is the resulting Bayes estimator.

As it turns out, the Bayes estimator for squared error loss (i.e., $L(\mathbf{y}, \theta) = (\theta - \hat{\theta})^2$) is the posterior mean of θ, a result that we often use for inference without putting much thought into the rationale for why we use it. Different loss functions result in different estimators. For example, the absolute loss (i.e., $L(\mathbf{y}, \theta) = |\theta - \hat{\theta}|$) results in the posterior median as the Bayes estimator and zero-one loss (i.e., $L(\mathbf{y}, \theta) = 0$ or $L(\mathbf{y}, \theta) = 1$ if $\theta = \hat{\theta}$ or $\theta \neq \hat{\theta}$, respectively) results in the posterior mode being the Bayes estimator.

Returning to the topic of model selection, Gelfand and Ghosh (1998) recommended a decision theoretic approach based on prediction rather than parameter estimation. In doing so, they proposed a loss function in terms of hypothetical replicates of the data $\tilde{y}_i$ (i.e., unobserved new data) that is a sum of two components

$$L(\tilde{y}_i, \hat{y}_i) + wL(y_i, \hat{y}_i) \ , \tag{46}$$

where $\hat{y}_i$ represents a predictive realization for the unobserved new data point $\tilde{y}_i$, and y_i represents the observed within-sample data point. In the proposed loss function (46), the w is constrained to be non-negative and expresses the relative weight given to loss for the within-sample versus new data at the same prediction $\hat{y}_i$.

Gelfand and Ghosh (1998) derived a posterior predictive risk by averaging their proposed loss function (46) over the posterior predictive distribution of $\tilde{y}_i|\mathbf{y}$. The resulting risk is then minimized with respect to the prediction $\hat{y}_i$ and summed over all observations

$i = 1, \dots, n$ to yield the model selection criterion

$$D_w = \sum_{i=1}^{n} \min_{\hat{y}_i} \int (L(\tilde{y}_i, \hat{y}_i) + wL(y_i, \hat{y}_i))[\tilde{y}_i|\mathbf{y}]d\tilde{y}_i \ , \tag{47}$$

where we would seek to find a model with the smallest D_w out of a proposed set of models given a chosen loss function $L(\cdot)$ and weight w. In practice, it can be difficult to compute the necessary integrals in (47), thus a squared error loss function is commonly used, yielding the criterion

$$D_{w,\text{sel}} = \frac{w}{w+1} \sum_{i=1}^{n} (y_i - \text{E}(\tilde{y}_i|\mathbf{y}))^2 + \sum_{i=1}^{n} \text{Var}(\tilde{y}_i|\mathbf{y}) \ . \tag{48}$$

Further, it is often assumed that the weight is very large ($w \to \infty$) thus resulting in a $D_{\infty,\text{sel}}$ criterion

$$D_{\infty,\text{sel}} = \sum_{i=1}^{n} (y_i - \text{E}(\tilde{y}_i|\mathbf{y}))^2 + \sum_{i=1}^{n} \text{Var}(\tilde{y}_i|\mathbf{y}) \ . \tag{49}$$

Note the similarity of $D_{\infty,\text{sel}}$ to the WAIC (45) and DIC (36, for large n) in that they both contain two terms in a sum, the first being a goodness-of-fit measure and the second acting as a penalty or regulator. In this case, we can see that the penalty $\sum_{i=1}^{n} \text{Var}(\tilde{y}_i|\mathbf{y})$ will increase in overfitted models where the prediction variance becomes larger with an increasing number of parameters.

For more general loss functions, such as deviance, D_w takes on a similar two component form, but the penalty is only guaranteed to be positive under certain constraints on the loss (i.e., convexity in y) and the criterion may not be suitable for mixture models. Despite this caveat, D_w does appear to be appropriate for many classes of hierarchical models because it depends directly on the posterior predictive distribution rather than the likelihood and posterior mean of the parameters alone. Also, unlike WAIC, the general form of posterior predictive loss approach appears to be suitable for correlated data models (e.g., spatial and temporal models).

Even though the posterior predictive loss approach does not technically fall into the same category as the rest of the information criteria, the form of the general loss function proposed by Gelfand and Ghosh (1998) is similar enough to the regularization expression (23), and equivalent to DIC and WAIC in certain settings, that we chose to describe it here rather than place it in its own section.

4.7 Willow Tit Occupancy: Information Criteria

In a continued assessment of predictive performance for the occupancy model set using the willow tit data, we calculated WAIC, DIC, and $D_{\infty,\text{sel}}$ for each of the 4 models previously considered (Table 5). To calculate WAIC for the occupancy model in this example, we used MCMC samples to approximate the effective number of parameters

$$p_{D,2} \approx \sum_{i=1}^{n} \frac{\sum_{t=1}^{T} \left(\log([y_i|J_i, p^{(t)} z_i^{(t)}]) - \sum_{t=1}^{T} \log([y_i|J_i, p^{(t)} z_i^{(t)}])/T \right)^2}{T}, \tag{50}$$

based on (44), where $[y_i|J_i, p^{(t)} z_i^{(t)}]$ is the binomial probability mass function and the first term in WAIC (45) is approximated as

$$-2 \sum_{i=1}^{n} \log \frac{\sum_{t=1}^{T} [y_i|J_i, p^{(t)} z_i^{(t)}]}{T} \; . \tag{51}$$

Recall that this expression (51) has the same form as the cross-validation score (22), but is based only on within-sample data.

For DIC, we used the traditional method for calculating the effective number of parameters (37) and approximated $\bar{D}$ and $\hat{D}$ by

$$\bar{D} \approx \frac{\sum_{t=1}^{T} -2\log[\mathbf{y}|\mathbf{J}, p^{(t)}\mathbf{z}^{(t)}]}{T} \tag{52}$$

$$\hat{D} \approx -2\log[\mathbf{y}|\mathbf{J}, \hat{p}\hat{\mathbf{z}}] \tag{53}$$

where $\hat{p}$ and $\hat{\mathbf{z}}$ are the posterior means for detection probability and true latent occupancy status across all sites, and $[\mathbf{y}|\mathbf{J}, p^{(t)}\mathbf{z}^{(t)}] = \prod_{i=1}^{n}[y_i|J_i, p^{(t)}z_i^{(t)}]$ is the likelihood based on the conditionally independent data for the willow tit occupancy model.

For the posterior predictive loss method, we calculated $D_{\infty,\text{sel}}$ as in (49) based on the expectation and variance approximations

$$\text{E}(\tilde{y}_i|\mathbf{y}) \approx \frac{\sum_{t=1}^{T} \tilde{y}_i^{(t)}}{T} \tag{54}$$

$$\text{Var}(\tilde{y}_i|\mathbf{y}) \approx \frac{\sum_{t=1}^{T}(\tilde{y}_i^{(t)} - \sum_{t=1}^{T} \tilde{y}_i^{(t)}/T)^2}{T} \tag{55}$$

where $\tilde{y}_i^{(t)} \sim [y_i|J_i, p^{(t)}z_i^{(t)}]$ is drawn on each MCMC iteration (for $t = 1, \ldots, T$) as a posterior predictive realization.

Of the three criteria considered in this example, recent statistical literature suggests that only WAIC is truly appropriate for the occupancy model (Gelman et al., 2014 b). However, given that DIC is commonly used to compare Bayesian occupancy models, we provide a comparison here. Furthermore, the criterion based on posterior predictive loss

($D_{\infty,\text{sel}}$) is not ideal for the occupancy model setting because the squared error loss function (49) may not be best representative for the zero-inflated binomial data model. A different loss function could be chosen, but then a derivation would be required to find a computable approximation based on MCMC samples. Still, we felt that a comparison of the methods could illuminate potential empirical differences between the approaches. If this were a real application rather than a pedagogical example, we would have only computed WAIC for this model and data set. In terms of computational time, it only required 6.1 minutes to fit the models sequentially and obtain these metrics (using 160,000 MCMC iterations for each model fit with a burn-in period of 16,000 iterations).

All of these approaches (i.e., WAIC, DIC, and $D_{\infty,\text{sel}}$) provide similar information in ranking the willow tit occupancy models by predictive ability based on within-sample data (Table 5). WAIC, DIC, and $D_{\infty,\text{sel}}$ all suggest model M_3, the model containing only the forest covariate, as the worst predictive model, with the null model next (M_1), and a virtual tie among the two models containing the elevation covariate (i.e., M_2 and M_4). This latter result is in agreement with the earlier cross-validation and CPO model comparison.

5 MODEL-BASED MODEL SELECTION

To a certain extent, the regularization methods discussed in Section 5 (especially the fully Bayesian Lasso described in Section 5.2.2) are model-based approaches to model selection. They are model-based because they contain a formal mechanism that trades off model fit for model parsimony. In Section 5.2.1, we saw that the Bayesian model itself provides a natural model reduction mechanism via the prior. In contrast to this form of continuous shrinkage induced by a strong prior on the parameters, other methods have been developed in a similar spirit that explicitly augment the overall model structure with selection components whose job it is to switch on and off various effects in the full model (O'Hara and Sillanpaa, 2009). The basic idea then is to build a model that contains all of the potential model components and then let the model decide which of them are helpful and which are not.

5.1 Indicator Variable Selection

For instructive purposes, consider again the basic linear regression model from (24)

$$y_i \sim \mathrm{N}(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \ ,$$

where, the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j, \ldots, \beta_p)'$ contains the individual coefficients corresponding to the p predictor variables of interest. A modification of the original regression model has been proposed such that $\beta_j = z_j \cdot \theta_j$ for $j = 1, \ldots, p$, where each original parameter is written as a product of a binary indicator variable z_j and a regression coefficient θ_j (e.g., George and McCulloch, 1993; Carlin and Chib, 1995; Kuo and Mallick, 1998). In general, a prior would be specified for each (z_j, θ_j) pair and the full Bayesian model could then be fit, yielding inference not only about the coefficients β_j, but also the selection indicators z_j. In this setting, if the posterior mean for a particular z_j is large (i.e., closer to one than zero) it would indicate that the j^{th} covariate is important in the model; conversely, when the posterior mean of z_j is close to zero it effectively removes the j^{th} effect from the model thereby inducing a certain parsimony.

In implementing an indicator variable selection model, one would be tempted to use independent priors for z_j and θ_j; for example, we might specify

$$z_j \sim \text{Bern}(\phi)$$

$$\theta_j \sim \text{N}(0, \tau_j^2) \ ,$$

for all $j = 1, \ldots, p$, assuming the covariates are standardized. However, an independent prior specification can cause computational problems if the prior for θ_j is too vague (i.e., the prior variance, τ_j^2, is large) because when $z_j = 0$ in an MCMC algorithm, θ_j will be sampled from its prior and the subsequent sampling of future $z_j = 1$ will rarely occur since the θ_j is likely to be far from the majority of posterior mass. Thus, to alleviate these computational problems, others (e.g., George and McCulloch, 1993; Carlin and Chib, 1995) have suggested joint priors for z_j and θ_j that include explicit dependence between the indicators and coefficients.

In Gibbs variable selection, Carlin and Chib (1995) and Dellaportas et al. (1997) suggest decomposing the joint prior distribution $[z_j, \theta_j] = [\theta_j | z_j][z_j]$. In this joint prior specification, the Bernoulli prior for z_j is retained, but the prior for θ_j conditional on z_j is written as

$$\theta_j | z_j \sim z_j \text{N}(0, \tau^2) + (1 - z_j)\text{N}(\mu_{\text{tune}}, \sigma^2_{\text{tune}}) \ , \tag{56}$$

which has the form of a mixture distribution and is often referred to as a “slab and spike” prior (Miller, 2002). The Gibbs variable selection procedure then involves choosing the tuning parameters μ_{tune} and σ^2_{tune} such that $\text{N}(\mu_{\text{tune}}, \sigma^2_{\text{tune}})$ is near the posterior so that the MCMC algorithm exhibits better mixing. Surprisingly, the seemingly informative prior

(56) does not actually influence the posterior for β_j, but rather only influences the behavior of the MCMC algorithm (Carlin and Chib, 1995).

In a similar model-based approach called "stochastic search variable selection," George and McCulloch (1993) proposed a joint prior for z_j and θ_j. However, unlike in the Gibbs variable selection, this alternative prior does influence the posterior and can be written as

$$\theta_j | z_j \sim z_j \mathrm{N}(0, c\tau^2) + (1 - z_j)\mathrm{N}(0, \tau^2) \ . \tag{57}$$

In stochastic search variable selection, both c and τ^2 are tuned such that τ^2 is quite small, providing an effective spike at zero while $c\tau^2$ is larger, creating a slab around zero. The slab then provides the prior for θ_j when the variable β_j is in the model (i.e., when $z_j = 1$). Both Gibbs and stochastic search variable selection methods require tuning to ensure well-mixed MCMC algorithms, but both can be useful for model-based model selection.

5.2 Reversible-Jump MCMC

A related model-based approach to model selection is referred to as reversible-jump Markov chain Monte Carlo (RJMCMC; Green, 1995). Normally, we reserve the names of computational approaches for algorithms only, not statistical models; however, in this case, the method really describes a model, but we retain the label RJMCMC for convention. In describing the RJMCMC approach, first recall the model set $\{M_1, \ldots, M_l, \ldots, M_L\}$ described earlier in Section 2.1. Now suppose that each of the models contain their own corresponding parameters $\boldsymbol{\theta}_l$. Note that the lengths, say p_l, of these parameter vectors $\boldsymbol{\theta}_l$ may vary. In RJMCMC, we treat the model index l as a random quantity to be modeled along with the set of all possible parameters $\boldsymbol{\theta}$. Or alternatively, we treat the number of parameters p_l as a random quantity and specify a model for it. Under certain assumptions, the posterior distribution of interest then is

$$[\boldsymbol{\theta}, l|\mathbf{y}] \propto [y|\boldsymbol{\theta}_l, l][\boldsymbol{\theta}_l|l][l] \ , \tag{58}$$

where $[\boldsymbol{\theta}_l|l]$ is the prior distribution for the parameters in model M_l and $[l]$ is the prior distribution for model M_l itself. The beauty of this specification is that it places multimodel inference directly in a fully Bayesian context.

The use of MCMC to implement this model (58) involves the usual steps: specify initial values for unknowns and then cycle through the unknowns, updating each one sequentially. The complication arises when sampling the model index l, and hence its associated parameters $\boldsymbol{\theta}_l$, because the model dimension changes depending on which model is sampled. Thus, care must be taken to account for the potentially different model dimension when accepting a Metropolis-Hastings proposal for the parameters in an MCMC algorithm. The term “reversible” derives from the fact that certain properties of the Metropolis-Hastings update must be retained to arrive at a valid posterior distribution (Green, 1995; Godsill, 2001). Specifically, if we leave one model space with a particular dimension for another of a different dimension, we need to ensure that we can revert back to the former dimension later in the Markov chain. Thus, a modified version of the Metropolis-Hastings ratio can be constructed for certain models that corrects for the transdimensional nature of the algorithm.

RJMCMC approaches have become a popular option for computing Bayes factors and

Bayesian model probabilities (e.g., Johnson and Hoeting, 2011). When prior model probabilities are assumed to be equal, the Bayes factor ($B_{l,l'}$) can be computed simply by calculating the quotient of summed number of visits to each model (M_l and $M_{l'}$) in the RJMCMC algorithm (Hastie and Green, 2012).

Due to its model-based form, RJMCMC is an appealing method for Bayesian multimodel inference but can be tricky or impossible to implement for complicated models. To that end, Barker and Link (2013) described a method that provides RJMCMC results using a *post hoc* approach that only requires one to fit the L individual models and then post-process the resulting MCMC samples using a second MCMC algorithm in the form of a Gibbs sampler. We describe this approach and apply it to the willow tit data next.

In the big picture, Godsill (2001) and O'Hara and Sillanpaa (2009) show that the RJMCMC and indicator variable selection approaches are related. The key difference is that the auxiliary variables z_j are effectively moving the model between dimensions by switching on and off model components. In doing so, Gibbs and stochastic search variable selection side-step the transdimensional complication altogether.

5.3 Willow Tit Occupancy: RJMCMC

We presented results pertaining to Bayesian model averaging earlier in Section 2. To compute those Bayesian model averaging quantities we use the RJMCMC approach described by Barker and Link (2013) which we briefly summarize here. One advantage of the Barker and Link (2013) approach is that the individual models can be fit separately and then recombined subsequently with a secondary MCMC algorithm to obtain posterior model probabilities. After the initial set of 4 occupancy models were fit individually (requiring only 5.7 minutes in sequence), the following secondary algorithm was constructed to iteratively sample the model and associated parameters.

1. Set MCMC iteration index to $k = 1$.

2. Choose initial model $M_l^{(k)}$. In our case we used $M_l^{(1)} = M_4$, the full model.

3. Select $p_l^{(k)}$, $\beta_{0,l}^{(k)}$, and $\boldsymbol{\beta}_l^{(k)}$ from the former MCMC output for model $M_l^{(k)}$.

4. If there are remaining parameters from the full model not obtained in step 3 (i.e., for models M_1, M_2, and M_3) then sample those from a known distribution (the form of which is arbitrary according to Barker and Link, 2013). We used a standard normal distribution to sample remaining parameters, $\text{N}(0, 1)$.

5. Order the parameter values from steps 3 and 4 and combine to form $\boldsymbol{\theta}$. For example, if $M_l^{(k)} = M_2$, then $\boldsymbol{\theta} \equiv (p_l^{(k)}, \beta_{0,l}^{(k)}, \beta_{1,l}^{(k)}, u_2^{(k)})'$, where $u_2^{(k)} \sim \text{N}(0, 1)$.

6. Compute the full-conditional model probability

$$P(M_l|\cdot) = \frac{[\mathbf{y}|\boldsymbol{\theta}, M_l][\boldsymbol{\theta}|M_l]P(M_l)}{\sum_{l'=1}^{4}[\mathbf{y}|\boldsymbol{\theta}, M_{l'}][\boldsymbol{\theta}|M_{l'}]P(M_{l'})} \tag{59}$$

   for each model $l = 1, \ldots, 4$.

7. Sample $M_l^{(k+1)}$ from a categorical distribution with probabilities $P(M_1|\cdot)$, $P(M_2|\cdot)$, $P(M_3|\cdot)$, and $P(M_4|\cdot)$.

8. Increment the model index $k = k + 1$ and go to step 3.

A few of the terms in step 6 of the Barker and Link (2013) algorithm need further clarification with respect to the specific model set under consideration. The likelihood term for our willow tit occupancy model simplifies to $[\mathbf{y}|\boldsymbol{\theta}, M_l] \equiv [\mathbf{y}|p_l^{(k)}, \beta_{0,l}^{(k)}, \boldsymbol{\beta}_l^{(k)}]$ which can be found by integrating $\mathbf{z}$ and $\mathbf{v}$ out of the hierarchical model such that

$$[\mathbf{y}|p_l, \beta_{0,l}, \boldsymbol{\beta}_l] = \prod_{i=1}^{n} \left(\psi_i p^{y_i}(1-p)^{J_i - y_i} I_{\{y_i>0\}}\right) + \left(1 - \psi_i + \psi_i(1-p)^{J_i}\right) I_{\{y_i=0\}} \ , \tag{60}$$

where we have omitted the MCMC indexing for clarity. In the integrated likelihood (60), $\psi_i = \mathbf{x}'_{l,i}\boldsymbol{\beta}_l$ and $I_{\{\dots\}}$ is an indicator variable that is one when the condition in the subscript is true and zero otherwise. The prior term can be factored into terms relevant for the current model being considered and terms for the remaining parameters: $[\boldsymbol{\theta}|M_l] \equiv [p_l^{(k)}][\beta_{0,l}^{(k)}][\boldsymbol{\beta}_l^{(k)}][\mathbf{u}^{(k)}]$. The last term, $[\mathbf{u}^{(k)}]$, is simply a product of independent standard normal distributions in our occupancy model.

This secondary MCMC algorithm required only seconds to run, as compared with the original model fits which required minutes. Furthermore, we found the secondary MCMC algorithm suggested by Barker and Link (2013) easier to program than the inline RJMCMC algorithm because we didn't have to modify the actually model fitting code. Obtaining the posterior model probabilities from the secondary MCMC algorithm output simply requires calculating the number of times each model $M_l^{(k)}$ is sampled out of the total number of MCMC iterations (e.g., $P(M_2|\mathbf{y}) = 83200/160000 = 0.52$).

Several other alternatives exist for implementing RJMCMC and obtaining required BMA quantities. Notable among them are techniques for regression models that exploit orthogonality properties in the design matrix allowing for a simplification in the model sampler (Clyde et al., 1996). More recently, a form of data augmentation has been proposed to generalize these methods for cases where the design matrix is non-orthogonal (Ghosh and Clyde, 2011). Overall, the suite of new approaches for model-based model selection is rapidly expanding and is making Bayesian model averaging more accessible than ever for ecologists. Still, fully automated software for performing BMA for a huge class of potential models is lacking due to the complexity of rigorously calculating the required quantities. As with many of the cutting-edge statistical methods, ecologists who wish to use them are acquiring the necessary statistical and computational skills to implement them on their own.

6 GUIDANCE

Thus far we have provided a fairly comprehensive review of methods for Bayesian model selection and multimodel inference, along with the advantages and disadvantages of each. One can use this document as a reference in deciding what type of model selection is appropriate depending on the desired statistical inference in a particular project. Assuming that the researcher desires some form of model selection or multimodel inference, and that they plan to use Bayesian methods, we provide the following set of questions and answers to help guide the researcher in finding an appropriate set of tools:

1. Is the researcher planning a new study? If so, he or she may want to consider collecting two sets of data, one for training, and another for validation. When prediction is of utmost importance, there is no substitute for out-of-sample data in model selection. It may be time for a paradigm shift in the way we design ecological studies. If predictive model selection is desired, we need to collect data that facilitates inference on both parameters and models.

2. Is the researcher using a historical data set?

(a) If the data set is large and computation time is not an overriding issue, the researcher may want to consider K-fold cross-validation for a set of candidate models or Bayesian regularization. Most Bayesian cross-validation implementations will require K separate fits of the model, thus increasing the computational time significantly. However, parallel computing is now possible on the desktop computer thanks to several user friendly software packages. So, cross-validation may not be as impractical as one might initially think.

(b) If the data set is small, n-fold cross-validation over a set of candidate models or Bayesian regularization may be more appropriate. The caveat is that leave-one-out cross-validation is not as stable as K-fold for $K < n$. Small data sets are always going to present problems for statistical inference and there is not much one can do to alleviate these issues, regardless of statistical paradigm.

3. Is the researcher wanting to do prediction-based model selection with a simple Bayesian model when computational time is limited? If so, they might want to consider using DIC. As a prediction-based information criterion, DIC performs similar to AIC in choosing parsimonious models. The caveat is that, like AIC, DIC will also choose larger models than necessary when the sample size is large. The biggest caution about DIC arises when the posterior mean of the parameters does not describe the central tendency of the posterior distribution well. Thus, DIC is not appropriate when there exist multiple modes in the posterior. Furthermore, DIC is best as a selection criterion when the number of effective parameters is much smaller than the sample size, which may not be the case in hierarchical models where the number of latent variables scales with sample size.

4. Does the researcher want to do prediction-based model selection with a hierarchical Bayesian model when computational time is limited? If so, Gelman et al. (2014 b) recommend using WAIC to select models. Unlike DIC, WAIC does not rely on posterior means of parameters, instead it uses the posterior predictive distribution and is the “most Bayesian” of all the information criteria. However, despite all the benefits of WAIC, it still only depends on within-sample data and its computationally friendly form requires an independence assumption at the data level, which is not appropriate for time series or spatial models. In these cases, posterior predictive loss provides an alternative.

5. Does the researcher desire model averaged inference on parameters or predictions? Bayes factors are the appropriate tool for doing Bayesian model averaging, but they often can only be approximated. Bayes factors can be approximated using BIC, but only under certain circumstances, and since BIC is not actually Bayesian, it has limited utility in a fully Bayesian setting. Hoeting et al. (1999) provided a good summary of methods for approximating model weights that have a formal justification. Note that, aside from BIC, none of the other information criteria have a solid foundation for Bayesian model averaging (e.g., AIC, DIC, WAIC). Bayes factors are not recommended in cases where models include improper priors (Spiegelhalter and Smith, 1982).

6. Does the researcher want a fully integrated model fitting and selection procedure? If so, a model-based approach like indicator or Gibbs variable selection, stochastic search variable selection, or RJMCMC may be warranted. Furthermore, connections exist between many model-based approaches and BMA under certain conditions.

These model-based methods perform best with some tuning of the algorithms, but when tuned, they perform quite well and seem to be more computationally efficient than cross-validation. As with information criteria, model-based model selection methods depend only on within-sample data and thus have the same set of caveats. Also, RJMCMC can be quite difficult to implement for certain models, but there are newer approaches that can be used to provide the same inference based on individual model fits (e.g., Barker and Link, 2013).

7 CONCLUSION

Ecologists are fascinated with model selection, and many have customized their research questions around likelihood methods for model selection and multimodel inference as illustrated by the recent forum on p-values and model selection in Ecology (2014, volume 95). Bayesian methods are becoming more common in ecological studies, but due to a fracturing of the literature pertaining to Bayesian model selection, it appears that many studies simply rely on conventional methods without much thought. Many Bayesian ecologists are aware of issues with certain Bayesian model selection approaches (e.g., Bolker, 2009), but are unaware of alternatives and how these alternatives may relate to each other. We have compiled and summarized the large body of literature on Bayesian model selection and multimodel inference methods in this guide so that ecologists can be better informed about their options.

What stands out to us is that, despite the seeming consensus among ecologists and wildlife biologists in how to perform model selection and multimodel inference, it is far from settled among statisticians; particularly in the Bayesian realm of inference. What also stands out is that nearly all model selection and multimodel inference methods are focused on improving predictive capabilities of models by balancing model fit and model parsimony. Prediction is often most important to the machine learning community (e.g., classification and regression trees, boosting and bagging algorithms) and related methods rely almost exclusively on out-of-sample data for model validation to improve prediction, but in the ecological and biological sciences, our scope seems to be limited to within-sample data. With an increasing ability to collect more data through, for example, better telemetry devices, remote sensing, citizen science efforts, and operations like NEON (National Ecological Observatory Network), ecologists are finally finding themselves with more data to answer scientific questions. Thus, model selection methods that rely on a separate set of validation data are now more accessible than ever for ecologists.

Cross-validation is an incredibly useful tool for model selection when only a single data set is available, a tool that is often overlooked or ignored on the grounds that it may be computationally infeasible. However, the current era of computing is seeing the most improvement in processor quantity and no longer in processor speed (Sutter, 2005). The one thing that computers are getting better at is parallel processing, and that happens to strongly favor the notion of model selection via cross-validation. A bit of extra effort spent on bookkeeping aspects of programming can make true prediction-based model selection feasible through the parallelization of a cross-validation procedure. Using the occupancy model as an example, we demonstrated that parallel programming requires relatively little extra effort to implement but can improve computational efficiency dramatically (e.g., from hours to minutes, sometimes seconds).

When it seems that fitting a single model is the computational bottleneck, we need to remember that there are several entire subfields within statistics and computer science devoted to finding more efficient ways to specify and fit models. Automated MCMC software has been a boon for science, allowing ecologists to easily specify and fit complicated Bayesian models (e.g., Kery, 2010), but a common complaint is that these software packages are slow. Fortunately, a wave of new automatic Bayesian software is becoming available (e.g., INLA, STAN, LibBi) that has shown dramatic increases in speed, but improvements can also be gained just by creating our own MCMC algorithms. This gives us the flexibility to use model reparameterizations and newer computational tricks such as variational Bayes (e.g., Omerod and Wand, 2010) and statistical emulators (e.g., Hooten et al., 2011) to speed up the model fitting process, which in turn aids in out-of-sample model selection.

Finally, as a closing thought, we feel that it is the right time for ecologists to become more open-minded about the use of strong priors. It is somewhat ironic that many popular non-Bayesian statistical methods (e.g., model selection, penalized likelihood, Lasso) depend on the implicit use of strong priors while at the same time Bayesians are warned against them. Bayesian priors provide a formal mechanism for placing constraints on models and, when used correctly, such constraints can be incredibly helpful (e.g., Moreno and Lele, 2010). Furthermore, seemingly vague priors can have a dubious effect on inference (Seaman et al., 2012) in models commonly used in ecological analyses. Yet, stronger priors can help with model selection, multicollinearity, and algorithm stability, not to mention formally incorporating existing scientific information into new analyses (e.g., Garrard et al., 2012).

ACKNOWLEDGEMENTS

The authors thank the handling editor and two anonymous reviewers as well as several colleagues, including David Anderson, Brian Brost, Frances Buderman, Alison Cartwright, Bob Dorazio, Brian Gerber, Tabitha Graves, Ephraim Hanks, Megan Higgs, Jennifer Hoeting, Devin Johnson, Shannon Kay, Bill Link, Kiona Ogle, Ann Raiho, Jay Rotella, Andy Royle, Viviana Ruiz-Gutierrez, Maria Uriarte, Jay Ver Hoef, and Perry Williams for valuable insight and early discussions about this work. Support for this work was provided by NSF 000347455 and NSF EF1241856. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

Albert, J. and S. Chib. (1990). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88: 669-679.

Ando, T. and R. Tsay. (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting*, 26: 744-763.

Barker, R.J. and W.A. Link (2013). Bayesian multimodel inference by RJMCMC: A Gibbs sampling approach. *The American Statistician*, 67: 150-156.

Berger, J.O. (2006). *Statistical Decision Theory and Bayesian Analysis.* Springer.

Bernardo, J.M. and A.F.M. Smith. (1994). *Bayesian Theory.* John Wiley & Sons.

Bolker, B. (2008). *Ecological Models and Data in R*, Princeton University Press.

Bolker, B. (2009). Learning hierarchical models: advice for the rest of us. *Ecological Applications*, 19: 588-592.

Bondell, H.D. and B.J. Reich. (2013). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, In Press.

Burnham, K.P. and D.R. Anderson. (2002). *Model Selection and Multimodel Inference, Second Edition.* Springer-Verlag.

Carlin, B.R. and S. Chib. (1995). Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57: 473-484.

Celeux, G., F. Forbes, C.P. Robert, and D.M. Titterington. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1: 651-674.

Clark, J.S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters*, 8: 2-14.

Clark, J.S. (2007). *Models for Ecological Data: An Introduction.* Princeton University Press.

Clyde, M.A., H. Desimone, and G. Parmigiani. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91: 1197-1208.

Congdon, P. (2006). Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational Statistics and Data Analysis*, 50: 346-357.

Cressie, N., C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19: 553-570.

Czado, C., T. Gneiting, and L. Held. (2009). Predictive model assessment for count data. *Biometrics*, 65: 121254-1261.

Dahlgren, J.P. (2010). Alternative regression methods are not considered in Murtaugh (2009) or by ecologists in general. *Ecology Letters*, 13: E7-E9.

Dellaportas, P., J.J. Forster, and I. Ntzoufras. (1997). On Bayesian model and variable selection using MCMC. *Technical Report: Department of Statistics, Athens University of Economics and Business*, Athens, Greece.

Dorazio, R.M., M. Kery, J.A. Royle, and M. Plattner. (2010). Models for inference in dynamic metacommunity systems. *Ecology*, 91: 2466-2475.

Dorazio, R.M. and D.T. Rodriquez. (2012). A Gibbs sampler for Bayesian analysis of site-occupancy data. *Methods in Ecology and Evolution*, 3: 1093-1098.

Garrard, G.E., M.A. McCarthy, P.A. Vesk, J.Q. Radford, and A.F. Bennett. (2012). A predictive model of avian natal dispersal distance provides prior information for investigating response to landscape change. *Journal of Animal Ecology*, 81: 14-23.

Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall, London.

Gelfand, A.E. and S.K. Ghosh. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika*, 85: 1-13.

Gelfand, A.E. and A.F.M. Smith. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85: 398-409.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. (2014 a). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC.

Gelman, A., J. Huang, and A. Vehtari. (2014 b). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, In Press.

Gelman, A. and C.R. Shalizi. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66: 8-38.

George, E.I. and R.E. McCulloch. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 85: 398-409.

Ghosh, J. and M.A. Clyde. (2011). Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association*, 106: 1041-1052.

Godsill, S.J. (2001). On the relationship between Markov Chain Monte Carlo methods for model uncertainty. *Journal of Computational and Statistical Graphics*, 10: 230-248.

Gotelli, N.J. and A.M. Ellison. (2012). *A Primer of Ecological Statistics, Second Edition.* Sinauer Associates.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106: 746-762.

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

Graham, M.H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, 84: 2809-2815.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82: 711-732.

Hastie, D.I. and P.J. Green. (2012). Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66: 309-338.

Hastie, T., R. Tibshirani, and J. Friedman. (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer.

Held, L., B. Schrodle, and H. Rue. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In Kneib, T. and Tutz, G. (eds.), *Statistical Modelling and Regression Structures Festschrift in Honour of Ludwig Fahrmeir*, 91110. Springer.

Hobbs, N.T. (2009). New tools for insight from ecological models and data. *Ecological Applications*, 19: 551-552.

Hobbs, N.T. and M.B. Hooten. (In Press). *Bayesian Models: A Statistical Primer for*

Ecologists. Princeton University Press.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14: 382-417.

Hooten, M.B., Larsen, D.R., and C.K. Wikle. (2003). Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landscape Ecology*, 18: 487-502.

Hooten, M.B., W.B. Leeds, J. Fiechter, and C.K. Wikle. (2011). Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. *Journal of Agricultural, Biological and Environmental Statistics*, 16: 475-494.

Johnson D.S. and J.A. Hoeting. (2011). Bayesian multimodel inference for geostatistical regression models. *PLoS ONE*, 6: e25677.

Johnson, D.S., P.B. Conn, M.B. Hooten, J. Ray, and B. Pond. (2013). Spatial occupancy models for large data sets. *Ecology*, 94: 801-808.

Johnson, J.B. and K.S. Omland. (2004). Model selection in ecology and evolution. *TRENDS in Ecology and Evolution*, 19: 101-108.

Kass, R.E. and A.E. Raftery. (1995). Bayes factors. *Journal of the American Statistical Association*, 90: 773-795.

Kery, M. (2010). *Introduction to WinBUGS for Ecologists.* Academic Press.

Kery, M. and H. Schmidt. (2004). Monitoring programs need to take into account imperfect species detectability. *Basic and Applied Ecology*, 5: 65-73.

Knaus, J. (2013). snowfall: Easier cluster computing (based on snow). R package version 1.84-4. URL: http://CRAN.R-project.org/package=snowfall.

Kuo, L. and B. Mallick. (1998). Variable selection for regression models. *Sankhya, Series B*, 60: 65-81.

Kyung, M., J. Gill, M. Ghosh, and G. Casella. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5: 369-412.

Laud, P. and J. Ibrahim. (1995). Predictive model selection. *Journal of the Royal Statistical Society, Series B*, 57: 247-262.

Lehmann, E.L. and G. Casella. (1998). *Theory of Point Estimation.* Springer.

Link, W.A. and R.J. Barker. (2006). Model weights and the foundations of multimodel inference. *Ecology*, 87: 2626-2635.

Link, W.A. and R.J. Barker. (2010). *Bayesian Inference: with Ecological Applications.* Academic Press.

MacKenzie, D.I., J.D. Nichols, G.B. Lachman, S. Droege, J.A. Royle, and C.A. Langtimm. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83: 2248-2255.

MacKenzie, D.I., J.D. Nichols, J.E. Hines, M.G. Knutson, and A.B. Franklin. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, 84: 2200-2255.

MacKenzie, D.I., J.D. Nichols, J.A. Royle, K.H. Pollock, L.L. Bailey, and J.E. Hines. (2006). *Occupancy Estimation and Modeling.* Elsevier.

Madigan, D. and A.E. Raftery. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89: 1535-1546.

Martin, T.G., B.A. Wintle, J.R. Rhodes, P.M. Kuhnert, S.A. Field, S.J. Low-Choy, A.J. Tyre, and H. Possingham. (2005). Zero-tolerance in ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8: 1235-1246.

Miller, A. (2002). *Subset Selection in Regression.* Chapman & Hall/CRC.

Moreno, M. and S.R. Lele. (2010). Improved estimation of site occupancy using penalized likelihood. *Ecology*, 91: 341-346.

O'Hara, R.B. and M.J. Sillanpaa. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4: 85-118.

Omerod, J.T. and M.P. Wand. (2010). Explaining variational approximations. *The American Statistician*, 64: 140-153.

Park, T. and G. Casella. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103: 681-686.

Pettit, L.I. (1990). The conditional predictive ordinate for the normal distribution. *Journal* *of the American Statistical Association*, 52: 175-184.

Plummer, M. (2002). Discussion of the paper by Spiegelhalter et al. in *Journal of the Royal Statistical Society, Series B*, 64: 620.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9: 523-539.

R Core Team. (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org.

Richardson, S. (2002). Discussion of the paper by Spiegelhalter et al. in *Journal of the Royal Statistical Society, Series B*, 64: 626-227.

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press.

Royle, J. A. and R.M. Dorazio. (2008). *Hierarchical Modeling and Inference in Ecology.* Academic Press.

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6: 461-464.

Seaman, J.W. III, J.W. Seaman Jr., and J.D. Stamey. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, 66: 77-84.

Spiegelhalter, D.J. and A.F.M. Smith. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society, Series B*, 44: 377-387.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. van der Line. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64: 583-639.

Stone, M. (1977). An asymptotic equivalence of choice of model cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B*, 36: 44-47.

Sutter, H. (2005). The free lunch is over: A fundamental turn toward concurrency in software. *Dr. Dobbs Report*, 30(3).

Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd ed.* Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58: 267-288.

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

Ver Hoef, J.M. and P.L. Boveng. (In Review). The hidden costs of multimodel inference. *Journal of Wildlife Management.*

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.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14: 867-897.

SUPPLEMENTAL MATERIAL

- Supplement: ZIP file containing data and R code (Ecological Archives).

TABLES

Table 1: Glossary

Term	Definition
AIC	Akaike's information criterion, a within-sample non-Bayesian score for prediction.
Bayes factor	The ratio of marginal data distributions pertaining to two models.
BIC	Bayesian (Schwartz) information criterion, a within-sample non-Bayesian score for model averaging.
CPO	Conditional predictive ordinate, a within-sample score for leverage.
Cross-validation	The iterative use of within-sample data to validate models in terms of out-of-sample predictive ability.
DIC	Deviance information criterion, a within-sample quasi-Bayesian score for prediction.
Effective number of parameters	p_D, a measure of model complexity as a penalty in Bayesian information criteria.
Empirical Bayesian	The use of within-sample data to inform Bayesian model components such as priors.
Out-of-sample data	An auxiliary set of data that are used for model comparison.
Posterior predictive loss	An approach for scoring models based on decision theory.
Regularization	Constraining a statistical optimization problem (i.e., penalization or shrinkage).
Regulator	constraint, optimism, penalty, or prior.
Score	A function used to evaluate models numerically, usually in terms of predictive ability.
WAIC	Watanabe-Akaike information criterion, a within-sample fully-Bayesian score for prediction.
Within-sample data	Response data typically used to fit a model, but also to calculate information criteria.

Table 2: Willow Tit Occupancy: Prior and posterior model probabilities.

Model	Covariates	$P(M_l)$	$P(M_l \vert \mathbf{y})$
M_1	NULL	0.25	0.00
M_2	ELEV	0.25	0.52
M_3	FOR	0.25	0.00
M_4	ELEV + FOR	0.25	0.48

Table 3: Willow tit occupancy posterior means for p, β_0, and $\boldsymbol{\beta}$ across all models and using BMA.

Parameter	M_1	M_2	M_3	M_4	BMA
p (detection prob.)	0.26	0.26	0.26	0.26	0.26
β_0 (intercept)	0.17	0.38	0.89	0.29	0.32
β_1 (elevation)	0.00	1.95	0.00	1.80	1.85
β_2 (forest)	0.00	0.00	1.79	0.39	0.18

Table 4: Willow tit occupancy results for cross-validation and CPO.

Model	Covariates	C-V Score	$-\sum_i \log(\mathrm{CPO}_i)$
M_1	NULL	552.4	240.2
M_2	ELEV	478.4	220.0
M_3	FOR	526.9	246.2
M_4	ELEV + FOR	478.8	220.4

Table 5: Willow tit occupancy results for WAIC, DIC, and $D_{\infty,\text{sel}}$ (posterior predictive loss).

Model	Covariates	WAIC	DIC	$D_{\infty,\text{sel}}$
M_1	NULL	481.7	462.2	288.0
M_2	ELEV	440.2	432.2	270.8
M_3	FOR	492.4	483.8	305.2
M_4	ELEV + FOR	440.7	432.9	271.2

FIGURE LEGEND

Figure 1 The results of a Web of Science search in number of articles per search string for each of the past 25 years (`http://thomsonreuters.com/web-of-science/`).

Figure 2 Overview of topics treated in this guide. These topics are grouped by their linkages to the main model selection and multimodel inference themes. Boxes represent overarching concepts, rounded boxes represent certain approaches that fall under those concepts, and ovals correspond to specific tools (gray indicates tools that are not clearly Bayesian). Arrows indicate specific types of approaches and tools that fall under the broader concepts, whereas dashed lines represent links among items if certain assumptions hold (e.g., BIC can be used for model averaging if parameters can easily be counted, priors are vague, and posterior modes are used as point estimates for parameters).

Figure 3 Willow Tit Occupancy: Bayesian Regularization. a.) Shrinkage trajectories for the posterior mean of $\boldsymbol{\beta}$ (y-axis) plotted against prior variance for $\boldsymbol{\beta}$ (x-axis). Parameter estimates yielding the best predictive model based on the two covariates occur at the vertical gray line. Note that the correlation between elevation and forest is 0.12. b.) The cross-validation score (y-axis) presented in (22) plotted against prior variance for $\boldsymbol{\beta}$ (x-axis). The optimal score (i.e., smallest; score= 478.5) for prediction occurs at the vertical gray line (i.e., minimum score occurs at $\sigma^2_\beta = 1.02$).

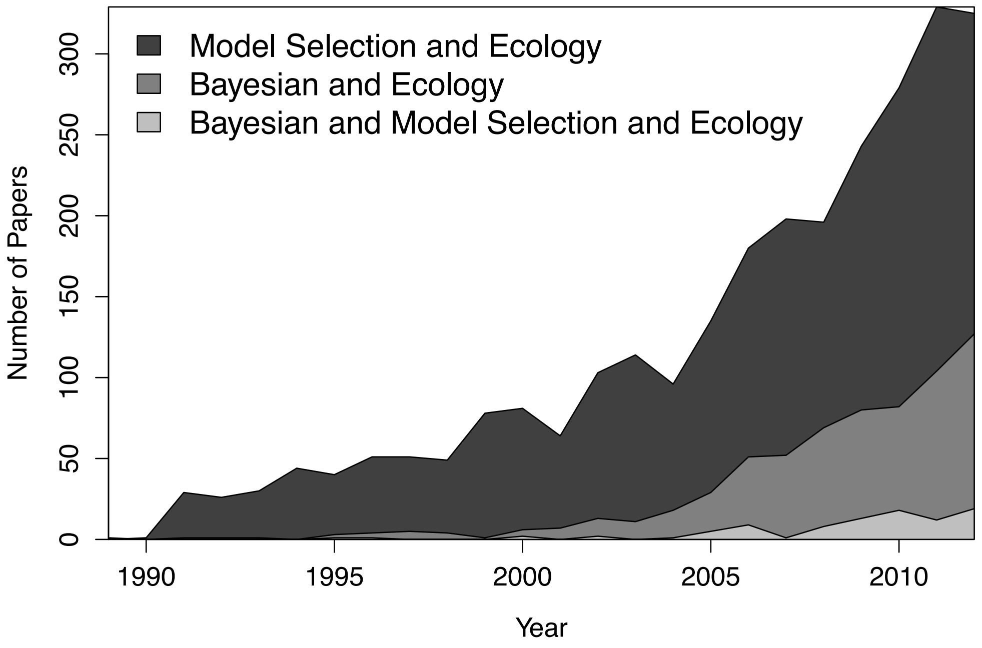
Model Selection and Ecology
Bayesian and Ecology
Bayesian and Model Selection and Ecology
Number of Papers
0
50
100
150
200
250
300
1990
1995
2000
2005
2010
Year

Figure 1:

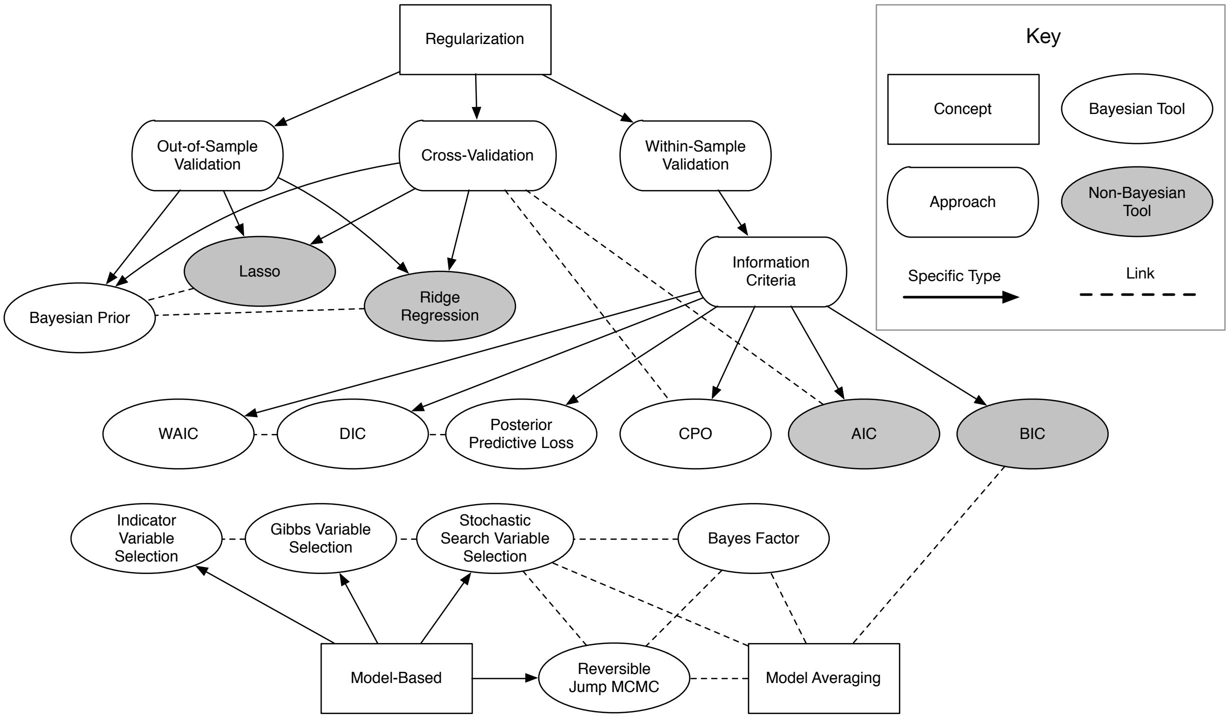
Regularization
Key
Concept
Bayesian Tool
Approach
Non-Bayesian Tool
Specific Type
Link
Out-of-Sample Validation
Cross-Validation
Within-Sample Validation
Lasso
Ridge Regression
Bayesian Prior
Information Criteria
WAIC
DIC
Posterior Predictive Loss
CPO
AIC
BIC
Indicator Variable Selection
Gibbs Variable Selection
Stochastic Search Variable Selection
Bayes Factor
Model-Based
Reversible Jump MCMC
Model Averaging

Figure 2:

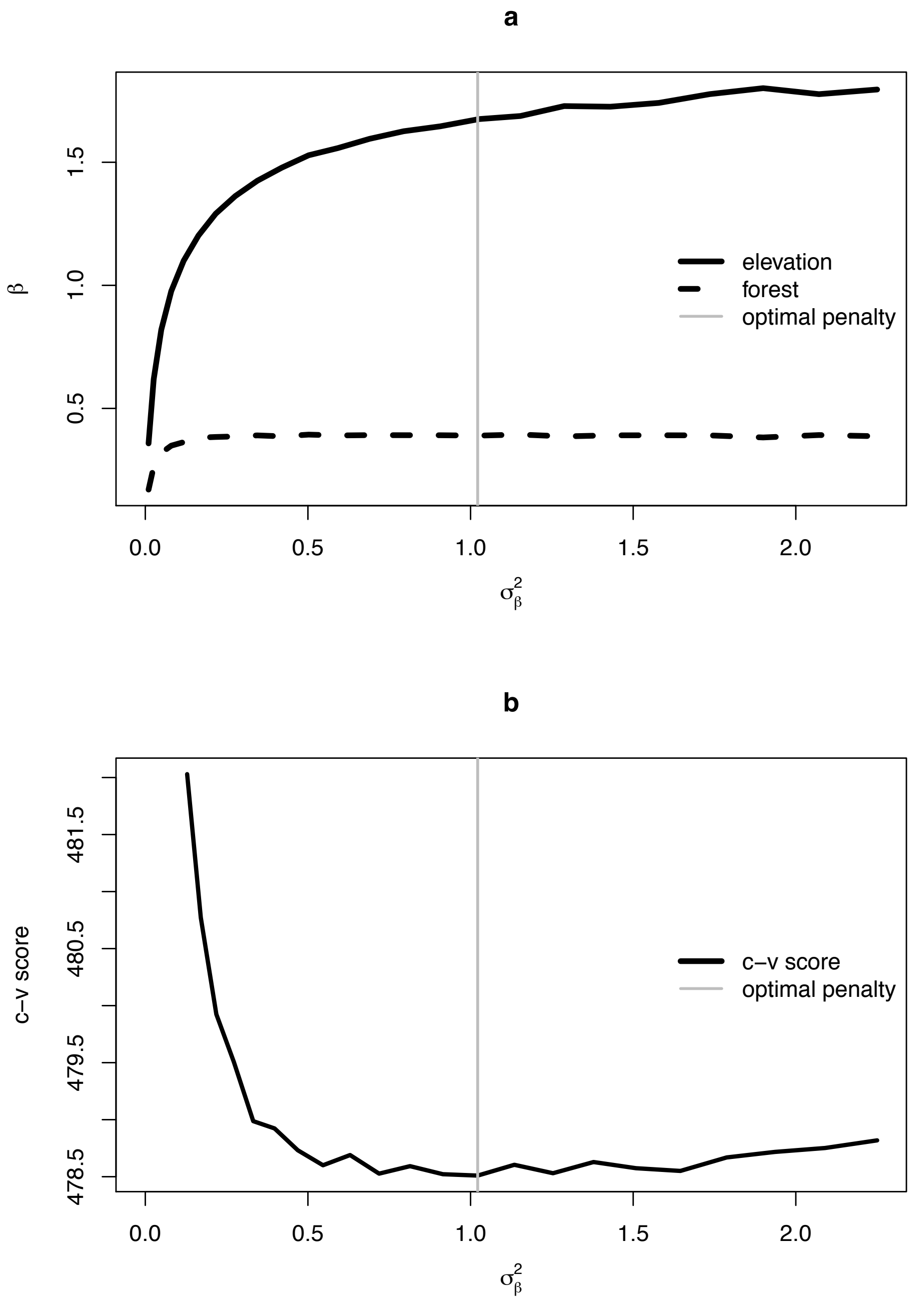
a
β
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
σ²β
elevation
forest
optimal penalty
b
c−v score
481.5
480.5
479.5
478.5
0.0
0.5
1.0
1.5
2.0
σ²β
c−v score
optimal penalty

Figure 3: