4 Comparing Ecological Models

MEVIN B. HOOTEN AND
EVAN G. COOCH

Model selection based on information theory is a relatively new paradigm in the biological and statistical sciences and is quite different from the usual methods based on null hypothesis testing. —Burnham and Anderson (2002:3)

Why Compare Models?

Statistical models provide a reliable mechanism for learning about unknown aspects of the natural world. By their very nature, however, statistical models are placeholders for true data-generating processes. Because true data-generating processes are unknown, statistical models represent our mathematical understanding of them while accounting for inherent randomness in how the underlying ecological process operates and how the data we may observe arise. In designing statistical models, we may have multiple perspectives about the mechanisms that give rise to the data. Therefore, a critical component of the scientific process involves the assessment of model performance, often in terms of predictive ability. We value statistical models that provide accurate and precise predictions because they excel at mimicking the data-generating mechanisms. We refer to an assessment of the predictive ability of a statistical model as *validation*.

We are also generally interested in model interpretation, which is focused on the question of which variables are more or less important in predicting the response. In a high-dimensional problem, it is likely that some subset of predictor variables is not strongly associated with the response variable or, in the extreme, not associated at all (such that the estimates for these variables are zero). This interest in variable selection is often intrinsically linked to questions about relative scoring among a set of candidate models fit to a set of data.

In this chapter, we introduce the concept of model scoring for parametric statistical models, how we calculate model scores, and what we do with the scores. We place the concept of scoring in a broader discussion about parsimony and its utility for prediction. We begin with likelihood-based methods and then shift to Bayesian methods, providing examples throughout based on a generalized linear model (GLM) for avian species richness. We also discuss the relationship between model scoring and variable selection.

Scoring Models

Deviance

Given that prediction is an important indicator by which to compare models, we often rely on a quantitative metric (i.e., a score) for assessing the predictive ability of models. Prediction is a form of learning about unobserved random quantities in nature. In statistics, prediction typically refers to learning

about unobserved data (e.g., at future times or new locations). For example, suppose there are two sets of data, one you collect ($\mathbf{y}$, an $n \times 1$ vector) and one you do not collect ($\mathbf{y}_u$) (i.e., data that are unobserved). Statistical prediction involves learning about $\mathbf{y}_u$ given $\mathbf{y}$. A point prediction results in our single best understanding of the unobserved data $\hat{\mathbf{y}}_u$ given the observed data ($\mathbf{y}$) and statistical model M. Then, to assess our prediction, we might consider a score that measures the distance between our prediction and truth (Gneiting 2011).

Numerous issues arise when scoring statistical models based on predictive ability. First, we typically do not know the unobserved data ($\mathbf{y}_u$), so a score cannot be calculated. Second, if we did have access to the unobserved data, the score would depend on the way we measured distance. We can address the first issue by collecting two data sets—one for fitting the model and one for validating the model. If the validation data set is large enough, it will provide an accurate representation of the predictive ability. The second issue is impossible to resolve without setting some ground rules. Thus, statisticians have traditionally recommended scoring functions that are based on distances inherent to the type of statistical model being used for inference. Such scores are referred to as *proper scores* (Gneiting and Raftery 2007).

The *deviance* is a proper score (also referred to as the "logarithmic score"; Gneiting and Raftery 2007). It is proper because it involves the likelihood associated with the chosen statistical model (i.e., hypothesized mechanism that gives rise to the data). The deviance is usually expressed as $D(\mathbf{y}) = -2\log f(\mathbf{y} \mid \boldsymbol{\beta})$, which involves a function f representing the likelihood evaluated for the data based on the model parameters $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)'$. We use the -2 multiplier in the deviance to be consistent with historical literature, and it implies that smaller scores indicate better predictive ability. There are other types of proper scoring functions, but because the deviance is one of the most commonly used scores, we focus on it throughout.

Validation

It is tempting to use the within-sample data $\mathbf{y}$ to score a model based on the deviance $D(\mathbf{y})$. However, the score will be "optimistic" about the predictive ability of the model because we learned about the parameters in the model using the same set of data (Hastie et al. 2009). "Optimism" is a term commonly used by statisticians to refer to an artificially inflated estimate of true predictive ability. This idea is easily understood by considering a data set with n data points, to which we wish to fit models containing as many as p predictor variables. Using a familiar least-squares approach, we seek to minimize residual sums of squares (RSS), a commonly used objective function in linear regression. However, the RSS will decrease monotonically as the number of parameters increases (often characterized by an increase in the calculated R^2 for the model). While this may seem like a positive outcome, there are two important problems. First, a low RSS (or high R^2) indicates the model has low error with respect to the within-sample data (sometimes referred to as the "training" data), when our interest is in choosing a model that has a low error when predicting out-of-sample data (validation, or "test" data). Coefficient estimates will be unbiased, and, if n is much larger than p, coefficient estimates will have low variance. However, as $p \to n$, there will be a substantial increase in variance. In the extreme, where $p \geq n$, the variance of the coefficient is infinite, even though $R^2 \approx 1$. Second, if the criterion for model selection is based solely on minimizing RSS in the training data (in the least squares context; equivalently, minimizing deviance in a likelihood framework for linear models), then the model containing all p parameters would always be selected. In fact, we want to select a model with good ability to predict the response outside the sample (i.e., we seek low test error).

Formally, the preceding relates to what is known as the *bias-variance tradeoff*. As a model becomes complex (more parameters), bias decreases, but variance of the estimates of parameter coefficients in-

creases. Following Hastie et al. (2009), we illustrate the basis for this relationship by proposing that a response variable y can be modeled as $y = g(\mathbf{x}) + \varepsilon$. The corresponding expected prediction error is written as $E((y - \hat{g}(\mathbf{x}))^2)$. If $\hat{g}(\mathbf{x})$ is the prediction based on the within-sample data, then

$$E((y - \hat{g}(\mathbf{x}))^2) = \sigma^2 + (E(\hat{g}(\mathbf{x})) - g(\mathbf{x})))^2 + \text{var}(\hat{g}(\mathbf{x})) \\ = \text{irreducible error} + \text{bias}^2 + \text{variance}$$

The first term, irreducible error, represents the uncertainty associated with the true relationship that cannot be reduced by any model. In effect, the irreducible error is a constant in the expression. Thus, for a given prediction error, there is an explicit trade-off between minimizing the variance and minimizing the bias (i.e., if one goes down, the other goes up).

Out-of-sample validation helps control for optimism by using separate procedures for fitting and scoring models. Out-of-sample validation corresponds to calculating the score for the out-of-sample data. In the case where two data sets (i.e., training and validation data) are available, the deviance can be calculated for each model using plug-in values for the parameters based on the point estimates $\hat{\boldsymbol{\beta}}$ from a model fit to $\mathbf{y}$. Thus, for a set of models $M_1, \ldots, M_l, \ldots, M_L$, we calculate $D_l(\mathbf{y}_u)$ and compare to assess predictive performance (lower is better).

A completely independent second data set is not often available to use for validation. In that case, cross-validation can be useful to recycle within-sample data for scoring. In cross-validation, the data set is split into two parts—a temporary validation data set $\mathbf{y}_k$ and a temporary training data set $\mathbf{y}_{-k}$ (where the $-k$ subscript refers to the remaining set of data from $\mathbf{y}$ after the kth subset is held out). For each "fold" of training data $\mathbf{y}_{-k}$, we fit model l and calculate the score $\mathrm{D}_l(\mathbf{y}_k)$ for the validation data set. After we have iterated through all K folds, we compute the joint score as $\sum_{k=1}^{K} D_l(\mathbf{y}_k)$ and compare among the L models to assess predictive ability.

Cross-validation is an appealing method because it automatically accounts for optimism and can be used in almost any setting without requiring a separate set of validation data. However, it is based on a finite set of data and includes a circular procedure because, presumably, the best predicting model identified in the comparison would then be fit to the entire data set for final inference and/or additional prediction. Furthermore, cross-validation can be computationally intensive if the associated algorithms require substantial computing resources. In modern computing environments, the computational burden is less of an issue than it was decades ago, but the cross-validation procedure (i.e., fitting the model for a set of folds and models) can require slightly more of an overhead programming investment.

SPECIES RICHNESS: CROSS-VALIDATION

As a case study, we consider continental U.S. bird species richness (Fig. 4.1) as a function of state-level covariates throughout this chapter. Suppose that we wish to model the bird counts (y_i) by U.S. state based on a set of state-level covariates x_i, for $i = 1, \ldots, n$ where $n = 49$ continental states in the United States (including Washington, DC) and the covariates are: state area (sq. km/1,000), average temperature (average degrees F), and average precipitation (average inches per year). Because the data y_i are nonnegative integers, a reasonable starting place for a data model for y_i is the Poisson distribution such that

$$y_i \sim \text{Pois}(\lambda_i). \qquad (4.1)$$

We link the mean richness (λ_i; also known as "intensity") to the covariates ($\boldsymbol{x}_i$) and regression coefficients ($\beta_0, \ldots, \beta_p$) using a log link function

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} \qquad (4.2)$$

for a set of covariates ($x_{j,i}, j = 1, \ldots, p$). It is common to see the regression part of the model written as $\log(\lambda_i) = \beta_0 + \boldsymbol{x}_i' \boldsymbol{\beta}$ or $\log(\lambda_i) = \boldsymbol{x}_i' \boldsymbol{\beta}$, depending on whether the intercept is included in $\boldsymbol{\beta}$ (in the latter case, the first element of vector $\boldsymbol{x}_i$ is 1).

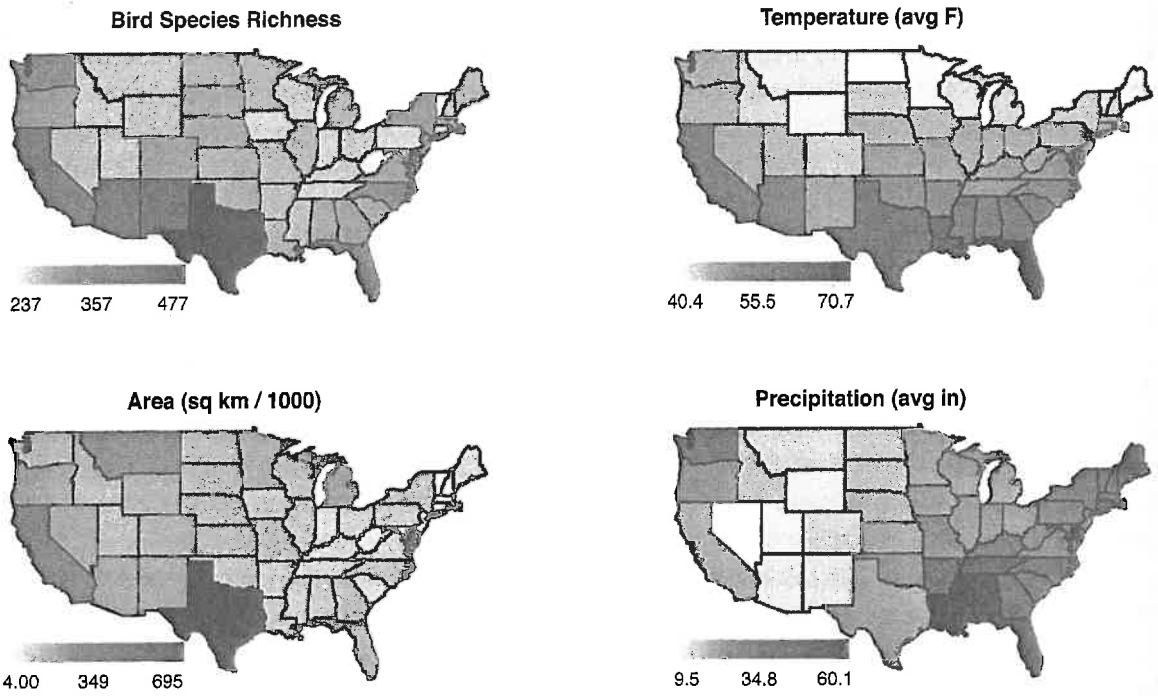

Fig. 4.1. Bird species richness in the continental United States and covariates: state area, average annual temperature, average annual precipitation.

The set of models we seek to compare are:

1. null model with only an intercept (no covariates)
2. intercept and area as covariate
3. intercept and temperature as covariate
4. intercept and precipitation as covariate
5. intercept and area and temperature as covariates

We excluded models with both state area and precipitation because they are strongly negatively correlated ($r = -0.63$). We also excluded models containing both temperature and precipitation because of moderate collinearity ($r = 0.48$). Substantial multicollinearity among covariates can cause regression models to be unstable (i.e., "irregular," more on this in what follows) and result in misleading inference (Christensen 2002; Kutner et al. 2004). Furthermore, we scaled the covariates to have mean zero and variance one before conducting all analyses. Standardizing covariates like this can reduce collinearity in some models and also puts the regression coefficients on the same scale so that they can be more easily compared.

In this case, because we had 49 observations, we used 7-fold cross-validation, breaking the data up into seven subsets. For each fold, we fit the models to 6/7ths of the data and computed the validation score (i.e., deviance) for the remaining 1/7th of the data (summing over all folds). The resulting deviance score, calculated using cross-validation $\sum_{k=1}^{7} D_l(y_k)$, for each of our five models $l = 1, \ldots, 5$) was 762.8, 597.1, 687.8, 755.4, and 551.5, respectively. The cross-validation scores indicated that models 5 and 2, both containing the "area" covariate, perform best for prediction because they have the lowest cross-validation scores.

Information Criteria

The inherent challenges associated with scoring models based on out-of-sample validation and cross-validation inspired several developments that con-

trol for optimism based only on within-sample data. In the maximum likelihood paradigm for specifying and fitting statistical models to data, two scoring approaches have been popularized: Akaike's information criterion (AIC; e.g., Akaike 1983; Burnham and Anderson 2002) and Bayes information criterion (BIC; e.g., Schwartz 1978; Link and Barker 2006). They are both similar in that they depend on the deviance as a primary component of the score, and they account for optimism in the predictive ability by penalizing the deviance based on attributes of the model or data collection process. Because smaller deviance indicates a better score, AIC and BIC penalize the score by adding a positive term to the deviance. For AIC, we calculate the score as $D(\mathbf{y}) + 2p$, where p is the number of unknown parameters $\boldsymbol{\beta}$ in the model. The score for BIC is calculated similarly as $D(\mathbf{y}) + \log(n)p$, with n corresponding to the dimension of the data set $\mathbf{y}$ (i.e., the sample size). Note that as $\log(n)$ becomes larger than 2, the penalty will have more influence on the score in BIC than AIC.

Within-sample scores are referred to as *information criteria* because the derivation of their penalties corresponds to certain aspects of information theory. *Information theory* arose from early work in signal processing and seeks to account for the information content in data. Given that statistics allows us to model data using probability distributions, information theory is concerned with how close the probability distribution we used to model the data is to the truth (Burnham and Anderson 2002). While it is impossible to calculate the distance between our model and the truth when the truth is unknown, the penalty used in AIC allows us to compare among a set of models to assess which is closest to the truth (with the lowest score indicating the closest). Conveniently, a separate derivation of the AIC penalty showed that it can also identify the best predicting model in certain circumstances (Stone 1977). Critically, the AIC penalty ($2p$) is a function of the number of unknown model parameters. Thus, complex models are penalized more than simpler ones. The concept of Occam's razor indicates that there is a sweet spot in model complexity that provides the best out-of-sample predictive ability (Madigan and Raftery 1994), with highly parameterized models being poorer predictors in limited data situations. Thus, it is often said that information criteria seek to balance model fit (to within-sample data) with parsimony (reducing model complexity to control for optimism in predictive ability).

The BIC score was derived with a different goal in mind than that of AIC. Under certain conditions, BIC identifies the data-generating model out of a set of models that includes the truth. It is also naturally a good score to use for calculating weights for model averaging, again, under certain conditions (Link and Barker 2006). Both AIC and BIC tend to rank models similarly when there are large gaps in the model performance, but AIC will select more complex models in general when the differences among models are small.

SPECIES RICHNESS: INFORMATION CRITERIA

Recall that AIC is defined as $\text{AIC} = D(\mathbf{y}) + 2p$ (based on the plug-in point estimates $\hat{\boldsymbol{\beta}}$), where p is the number of model parameters (p is equal to 1, 2, 2, 2, 3 for our five models, respectively). Similarly, BIC is defined as $\text{BIC} = D(\mathbf{y}) + \log(n)p$, where $n = 49$ for our case study. For our models, the deviance is calculated as

$$D(\mathbf{y}) = -2\sum_{i=1}^{n} \log(\text{Pois}(y_i | \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \cdots + \hat{\beta}_p x_{p,i}))), \tag{4.3}$$

where "Pois" stands for the Poisson probability mass function. We calculated AIC and BIC for each of our five models.

The results in Table 4.1 indicate that AIC and BIC are similar and agree on the ranking of the models. The information criteria also agree with the results of the cross-validation, in that models 5 and 2 are the top two models for our data. Across all models, the intercept was fairly consistent and the estimated coefficients for the "area" and "temp" predictor variables were positive while that for "precip"

Table 4.1. Likelihood-based information criteria and coefficient point estimates.

Model	AIC[1]	BIC[2]	Intercept	Area	Temperature	Precipitation
1	741.1	743.0	5.761	–	–	–
2	571.2	575.0	5.755	0.100	–	–
3	669.2	673.0	5.758	–	0.069	–
4	706.1	709.8	5.759	–	–	0.049
5	526.7	532.4	5.754	0.092	0.055	–

[1]Akaike information criterion

[2]Bayes information criterion

was negative. These results imply that increases in state area and average temperature predict higher bird species richness, while an increase in average precipitation predicts lower bird species richness.

Regularization

The concept of penalizing complex models to account for optimism and improve predictive ability is much more general than the way in which it is used in AIC and BIC. *Regularization* is a type of penalization that allows users to choose a penalty that suits their goals and meshes well with their perspective about how the world works. A general scoring expression is $D(\mathrm{y})+a\sum_{j=1}^{k}|\beta_j|^b$, where a and b are *regularization parameters* that affect the type and strength of penalty. Because the user can set b, there are infinitely many regularization forms, but the two most commonly used are the "ridge" penalty ($b=2$, equation 4a; Hoerl and Kennard 1976) and the "lasso" penalty ($b=1$, equation 4b; Tibshirani 1996):

$$D(\mathrm{y})+a\sum_{j=1}^{k}|\beta_j|^2 \quad \text{ridge penalty,} \qquad (4.4a)$$

$$D(\mathrm{y})+a\sum_{j=1}^{k}|\beta_j| \quad \text{lasso penalty.} \qquad (4.4b)$$

It is intuitive to view these penalties geometrically. For example, in the case where a model has two parameters, we can view the penalty as a shape in two-dimensional parameter space. Consider all possible values that the parameters β_1 and β_2 can assume in the space depicted in Fig. 4.2. When the model is fit to a particular data set without any penalization, the point estimate $\hat{\beta}$ will fall somewhere in this space (the lower right quadrant in the example shown in Fig. 4.2). When penalized, the estimates will be "shrunk" toward zero by some amount controlled by the penalty.

The ridge penalty ($b=2$) is represented as a circle with its radius being a function of the regularization parameter a. The lasso penalty ($b=1$) is represented as a diamond where the area is a function of the regularization parameter a. For more than two parameters, the constraint shapes become higher dimensional (e.g., spheres and boxes for $p=3$). Note that the lasso constraint has corners whereas the ridge constraint is smooth. These features play an important role in the penalization. The penalized parameter estimates are snapped back to a point on the shape (i.e., where the error ellipses intersect with the shape of the particular constraint function).

The ridge and lasso coefficient estimates represent a compromise between the model fit to the available data and a penalty to account for optimistic predictive ability. The strength of the penalty is induced by the regularization parameter a. As a increases, the shapes shrink in size (i.e., distance between the edge and the origin decreases), taking the penalized parameter estimates with them. Notice that as the regularization parameter a increases past a certain point, the lasso-penalized estimate will shrink to exactly zero for one of the parameters, whereas the ridge-penalized estimate

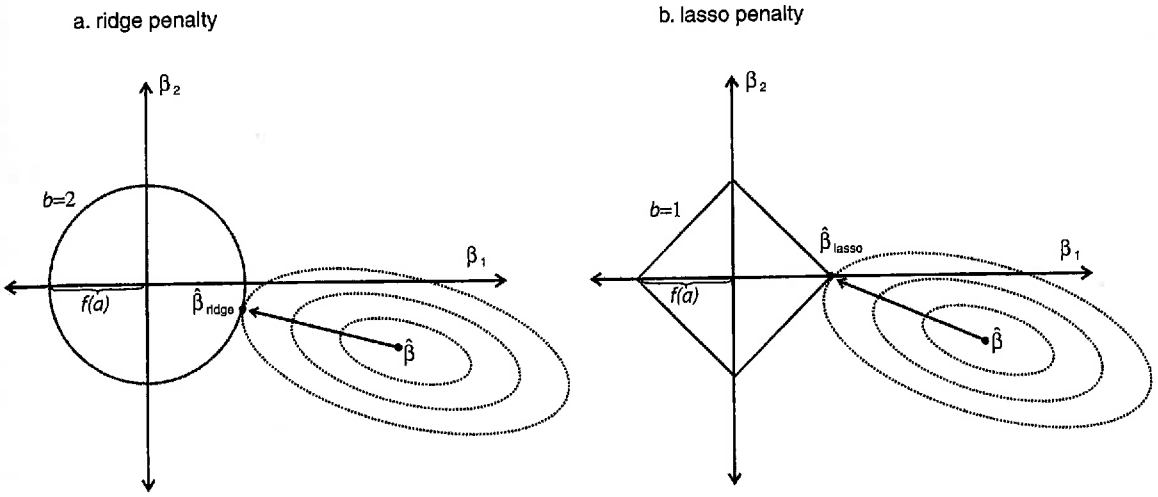

Fig. 4.2. Ridge penalty (gray circle, $b=2$) and lasso penalty (gray diamond, $b=1$) with three point estimates in parameter space for a model with two parameters (β_1, β_2): the unpenalized estimate ($\hat{\boldsymbol{\beta}}$), the ridge estimate ($\hat{\boldsymbol{\beta}}_{ridge}$), and the lasso estimate ($\hat{\boldsymbol{\beta}}_{lasso}$). The arrows indicate the shrinkage induced by the penalty in each case. The ellipses represent contours of the distribution for the unpenalized estimate ($\boldsymbol{\beta}$).

will shrink toward zero but at an asymptotic rate never reaching zero exactly until $a \to \infty$. These trajectories are a result of simple geometry. The sharp corners of the lasso penalty imply that the lasso estimates will more likely fall on a point of the diamond shape, which means that one of the two parameters will be estimated as zero (effectively removing that effect from the model, as in Fig. 4.2, where $\beta_{lasso,2}=0$). By contrast, the ridge penalty will shrink all parameters in the model, but not to zero unless $a \to \infty$.

The application of these differences in shrinkage trajectory represents the user perspective about the world. In the case of lasso, when certain parameters are set to zero by the penalty, they are effectively removed from the model, making the model discretely less complex. By contrast, the ridge estimates leave all parameters in the model, but reduce their influence appropriately. Some have argued that the ridge penalty better mimics the real world because everything is affected by everything else, even if only by an infinitesimal amount, although this may complicate the interpretability of the model. However, lasso regularization has other beneficial properties, such as retaining sparsity in the parameter space (by forcing some parameters to be zero) and it has become popular (Tibshirani 1996).

Regularization can also help alleviate the effects of multicollinearity on inference. When covariates are highly correlated (i.e., collinear), the associated coefficient estimates will often oppose each other (i.e., one gets large and the other gets small) because they are effectively fighting over the same type of variability in the data. In extreme collinearity cases, the parameter estimates can oppose each other strongly. Regularization shrinks the parameter estimates toward zero, thereby reducing the effects of collinearity. The resulting regularized estimates are technically biased, but have much lower variability. Ridge regression was developed for precisely this purpose. The term "regularization" is so named because it induces regularity in models (Hoerl and Kennard 1976). Because regular models have fewer parameters than data and do not have highly collinear predictor variables, regularization helps with both cases.

The catch with regularization is that the user has to choose the parameters a and b. The shape parameter b is often chosen based on the goals of the study

and the desired type of shrinkage, but the strength parameter a is not as easy to set. In principle, it would be most satisfying to formally estimate a along with the other model parameters β. However, the within-sample data do not carry enough information to estimate a by themselves. Thus, a is typically set based on how well it improves the predictive ability of the model for out-of-sample data using the same validation or cross-validation techniques described in the previous section. We illustrate the cross-validation approach to regularization and finding an optimal value for a in what follows. The shape parameter b can also be chosen using cross-validation alone, or it can be selected *a priori* depending on whether the user wants some parameter estimates to be set to zero if necessary.

SPECIES RICHNESS: REGULARIZATION

Regularization allows for the comparison of an infinite set of models because we can include all covariates and let the penalty shrink them toward zero based on cross-validation. We fitted both ridge and lasso Poisson regression to the bird richness data set using values of a ranging from 0 to 100. We included a simulated covariate ("sim") that is strongly correlated with the "area" covariate to demonstrate the differences among coefficient estimate trajectories for the real versus simulated covariates. The resulting trajectories and cross-validation scores are shown in Fig. 4.3. Both types of regularization, ridge and lasso, shrink the coefficient estimate for the simulated covariate ("sim") to zero faster than the others. The simulated covariate is shrunk exactly to zero by lasso, immediately resulting in the optimal model for prediction. With the ridge penalty, the coefficient estimate for the simulated covariate actually changes sign (from negative to positive) but ends up near zero in the optimal model for prediction. The regularization results in larger effects for the real covariates for both ridge and lasso. Between the two types of penalties, the resulting optimal score for the ridge penalty was better (542.7) than the score for lasso (544.2).

Scoring Bayesian Models

Posterior Predictions

Bayesian statistics are similar to likelihood-based statistics in that a parametric probability distribution is chosen as a model for the data. The difference is that parameters are treated as unobserved random variables in Bayesian models, as opposed to fixed variables in non-Bayesian models (Hobbs and Hooten 2015). We use conditional probability statements to find the probability distribution of unknown variables (i.e., parameters and predictions) given known variables (i.e., data). Thus, if we treat our model parameters β as random variables with distribution $f(\beta)$ before the data are observed, our Bayesian goal is to find the posterior distribution $f(\beta|\mathbf{y})$ after the data are observed. Using conditional probability, we find that the posterior distribution is $f(\beta|\mathbf{y}) = f(\mathbf{y}|\beta)f(\beta)/f(\mathbf{y})$, which is a function of likelihood $f(\mathbf{y}|\beta)$, the prior $f(\beta)$, and the marginal distribution of the data $f(\mathbf{y})$ in the denominator (Hobbs and Hooten 2015). The marginal distribution of the data is often the crux in solving for the posterior distribution because it usually involves a complicated integral or sum. Therefore, we use numerical approaches such as Markov chain Monte Carlo (MCMC) algorithms for approximating the posterior distribution and associated quantities (Gelfand and Smith 1990).

The Bayesian mechanism for prediction is the posterior predictive distribution $f(\mathbf{y}_u|\mathbf{y})$. Bayesian point predictions can be calculated by $\hat{\mathbf{y}}_u = \sum_{t=1}^{T} \mathbf{y}_u^{(t)}/T$ (i.e., the mean of the posterior predictive distribution) using posterior predictive samples $\mathbf{y}_u^{(t)}$ arising from the MCMC model-fitting algorithm. However, the posterior predictive distribution provides much more information about the unobserved data as well, such as the uncertainty in our predictions. Thus, while it is tempting to use only the point predictions, we can obtain a much deeper understanding of what we know, and do not know, about the things we seek to predict using additional characteristics from the posterior predictive distribution.

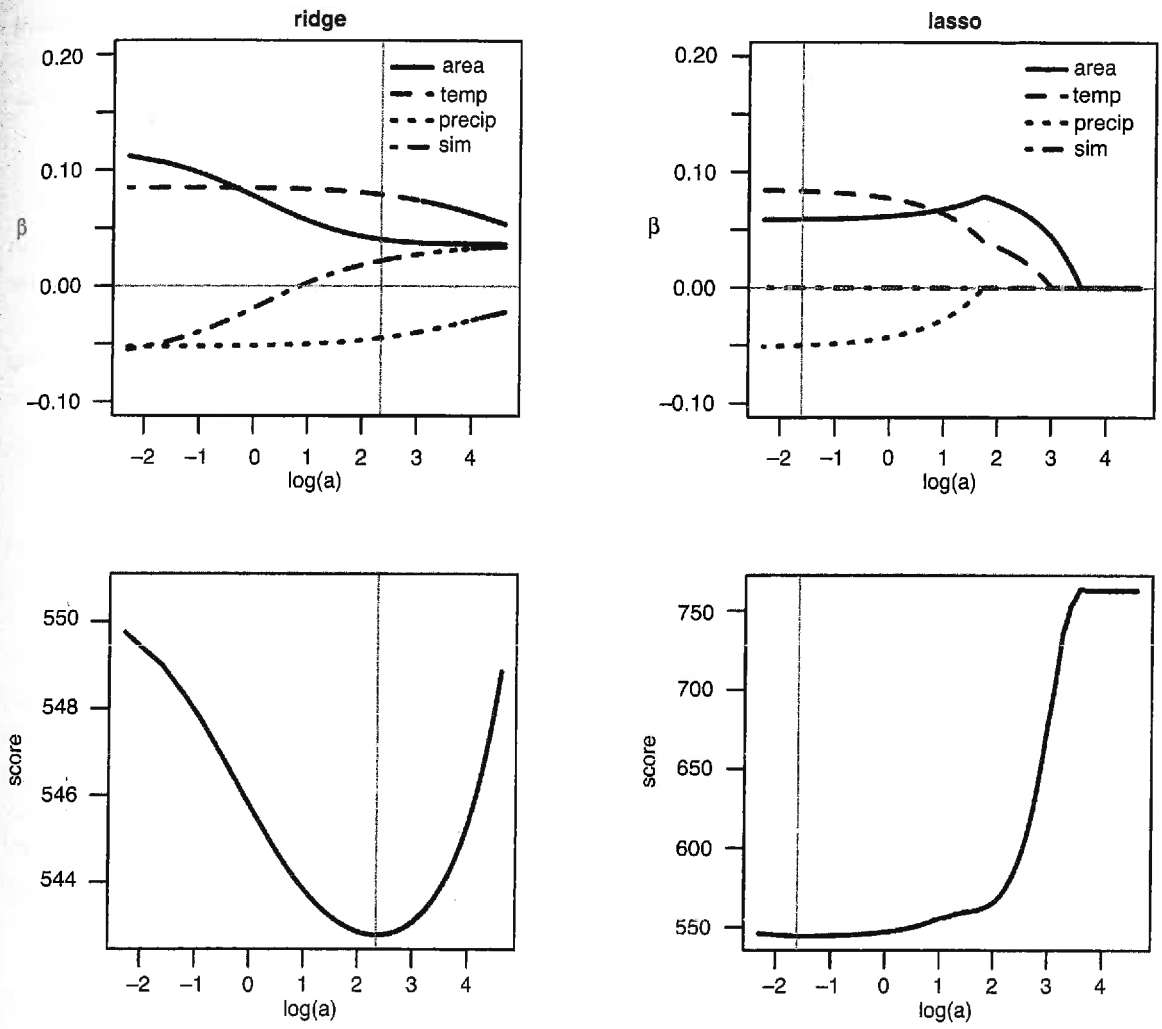

Fig. 4.3. Poisson regression coefficient estimate ($\hat{\beta}$) trajectories (top panels) based on the ridge penalty (left panels; $b=2$) and lasso penalty (right panels; $b=1$); regularization score trajectories (bottom panels) based on cross-validation. Vertical gray lines represent the values for the strength parameter a based on the optimal scores.

Scoring Bayesian Models

In principle, the same concept of scoring described in previous sections can be applied to Bayesian point predictions (Vehtari and Ojanen 2012). In this case, the deviance can be calculated using the same likelihood as in the non-Bayesian models but with Bayesian point estimates as plug-in values for the parameters β. Alternatively, we can leverage one of the key benefits that comes for free with MCMC: namely, the ability to obtain inference for any function of model components (e.g., data, predictions, and parameters). Therefore, a natural Bayesian score would be the mean posterior predictive deviance $\bar{D}(\mathbf{y}_u)$, which can be calculated using MCMC samples as $\sum_{t=1}^{T} D(\mathbf{y}_u^{(t)})/T$, where t corresponds to a MCMC iteration and T is the total number of MCMC iterations. The posterior predictive distribution provides the Bayesian mechanism for obtaining predictions from a model. Thus, because the deviance can be calculated using predictions, we treat it as a derived quantity (i.e., a statistic that does not affect

model fit, but that is a function of predictions or parameters in the model). MCMC makes it easy to obtain an estimate of this statistic, which becomes the score for our model. Calculating this new score using Bayesian methods, we can compare models as before based on out-of-sample data or cross-validation to account for optimism.

Bayesian models may require more time to fit than non-Bayesian models. Although Bayesian models can provide richer forms of inference, cross-validation may become infeasible for very large data sets and/or complex Bayesian models. Thus, within-sample scoring methods for Bayesian models have been developed. The most commonly used within-sample score for Bayesian models is the *deviance information criterion* (DIC; Spiegelhalter et al. 2002; Celeux et al. 2006). DIC takes the same form as AIC, with the deviance based on a plug-in Bayesian point estimate for $\boldsymbol{\beta}$, plus a penalty. The DIC penalty is $2p_D$, with p_D representing the effective number of parameters. It is not possible to count parameters discretely in Bayesian models because the prior provides some information about model parameters. Instead, we can think of a measure for model complexity (i.e., the optimism) as the difference in score between the deviance that accounts for the uncertainty in model parameters ($\bar{D}$) and the deviance based on only the plug-in parameter estimates ($\hat{D}$), resulting in $p_D = \bar{D} - \hat{D}$. As it turns out, for simple models, p_D is close to the number of parameters p when priors are less informative.

An alternative Bayesian score based on within-sample data uses the posterior predictive distribution directly (Richardson 2002; Watanabe 2010). The so-called Watanabe-Akaike information criterion (WAIC; Watanabe 2013) substitutes in the logarithm of the posterior predictive density for the deviance and uses a different calculation for the penalty. Thus, WAIC is specified as $-2 \log f(\mathbf{y}|\mathbf{y}) + 2p_D$. Using the MCMC sample, WAIC can be calculated as $-2\sum_{i=1}^{n} \log \sum_{t=1}^{T} f(y_i|\beta^{(t)})/T + 2p_D$, where $p_D = \sum_{i=1}^{n} \text{var}(\log f(y_i|\beta))$ and the "var" corresponds to the variance over the posterior distribution for $\boldsymbol{\beta}$. Heuristically, WAIC balances fit with parsimony to improve predictive ability because the uncertainty will increase as the model complexity increases. Thus, as the posterior variance of the deviance increases, the model is penalized more.

SPECIES RICHNESS: BAYESIAN INFORMATION CRITERIA

Recall that DIC is defined as $\text{DIC} = \hat{D} + 2p_D$, for $p_D = \bar{D} - \hat{D}$. These different forms of deviance can be computed using MCMC output from our model using

$$\hat{D} = -2\sum_{i=1}^{n} \log(\text{Pois}(y_i|\hat{\lambda}_i)), \tag{4.5}$$

and

$$\bar{D} = -2\,\frac{\sum_{t=1}^{T}\sum_{i=1}^{n} \log\left(\text{Pois}\left(y_i|\exp(\beta_0^{(t)} + \beta_1^{(t)}x_{1,i} + \cdots + \beta_p^{(t)}x_{p,i})\right)\right)}{T}, \tag{4.6}$$

where $\hat{\lambda}_i$ is the posterior mean of λ and $\beta_j^{(t)}$ is the jth coefficient on the tth MCMC iteration (for $j = 1, \ldots, p$ and a MCMC sample of size T).

Similarly, the Watanabe-Akaike information criterion is

$$\text{WAIC} = -2\sum_{i=1}^{n} \text{lppd} + 2p_D, \tag{4.7}$$

where "lppd" stands for log posterior predictive density for y_i and can be calculated using MCMC as

$$\text{lppd}_i = \log\left(\frac{\sum_{t=1}^{T} \text{Pois}\left(y_i|\exp(\beta_0^{(t)} + \beta_1^{(t)}x_{1,i} + \cdots + \beta_p^{(t)}x_{p,i})\right)}{T}\right), \tag{4.8}$$

and where Gelman and Vehtari (2014) recommend calculating p_D as

$$p_D = \sum_{i=1}^{n}\left(\frac{\sum_{t=1}^{T}(\log(\text{Pois})_i^{(t)} - \sum_{t=1}^{T}\log(\text{Pois})_i^{(t)}/T)^2}{T}\right), \tag{4.9}$$

where

$$\log(\text{Pois})_i^{(t)} = \log(\text{Pois}(y_i | \exp(\beta_0^{(t)} + \beta_1^{(t)} x_{1,i} + \ldots + \beta_p^{(t)} x_{p,i}))).$$

To specify a Bayesian model for the bird species richness data, we used the same Poisson likelihood as previously and then specified priors for the parameters. A reasonable prior for unconstrained regression coefficients is Gaussian (because the support for β_j includes all real numbers). Thus we specify

$$\beta_j \sim N(\mu_j, \sigma_j^2) \quad \text{for} \quad j = 1, \ldots, p, \quad (4.10)$$

as priors with means $\mu_j = 0$ and variances $\sigma_j^2 = 100$. Using these priors, we fit each of our models to the bird richness data and calculated DIC and WAIC.

While the values in Table 4.2 for DIC and WAIC are different, the ordering of models remains the same and is consistent with the non-Bayesian information criteria (although that may not always be true). Again, we see that models 5 and 2 provide the best predictive ability among our set of five models.

Bayesian Regularization

Comparing Bayesian models is not limited to use with out-of-sample scoring and information criteria. The concept of regularization also naturally transfers to the Bayesian setting (Hooten and Hobbs 2015). In fact, the regularization penalty already exists in Bayesian models as the prior. To see this connection, notice that the logarithm of the numerator in conditional probability is $\log(f(\mathbf{y} | \beta)) + \log(f(\beta))$. Thus, multiplying by -2, we have the same regularization expression as in previous sections, but with the penalty equal to $-2 \log(f(\beta))$. Therefore, the penalty is a function of the prior.

In regression models, the most common prior for the coefficients is a normal distribution. If we let the prior for the intercept be $\beta_0 \sim N(0, \sigma_0^2)$ and the prior for the slope coefficients be $\beta_j \sim N(0, \sigma_\beta^2)$, then the regularization shape parameter is $b = 2$ (as in ridge regression) and the strength of the penalty a is proportional to the reciprocal of prior variance ($1/\sigma_\beta^2$; Hooten and Hobbs 2015). Thus, to induce a stronger penalty, we simply make the prior variance for the slope coefficients small, hence shrinking the posterior for the slope coefficients toward zero. To choose the optimal value for σ_β^2, we can perform cross-validation (or out-of-sample validation), as before, to improve predictive ability of the model (Watanabe 2010).

The Bayesian regularization procedure can be used with a suite of different model specifications that involve regression components (e.g., logistic regression, Poisson regression, occupancy models, capture-recapture models; Hooten and Hobbs 2015). Different regularization penalties can be imposed by using different priors. For example, using a double exponential prior for β_j instead of a Gaussian prior results in a Bayesian lasso penalty (Park and Casella 2008; Kyung et al. 2010). The only potential disadvantage is that a Bayesian regularization procedure may involve more computation than fitting the model a single time when cross-validation is applied and the regularization is tuned. Even so, when fitting many types of models on modern computers, Bayesian regularization is feasible. Alternatively, strong priors that are set in the traditional way, using preexisting scientific knowledge about the process, may be enough to facilitate a natural Bayesian regularization without requiring an iterative model-fitting procedure (Seaman et al. 2012).

Table 4.2. Bayesian information criteria.

Model	DIC[1]	WAIC[2]
1	741.2	748.4
2	571.3	577.4
3	669.2	680.1
4	706.1	720.3
5	526.7	533.8

[1] Deviance information criterion

[2] Watanabe-Akaike information criterion

Discussion

We presented several different approaches for comparing parametric statistical models in this chapter. Our focus was mainly on comparing models with respect to predictive ability, but not all of the methods are designed to be optimal for prediction in the same sense. For example, BIC is more closely related to model averaging, and model averaged predictions outperform the predictions from any one model alone. Thus, multimodel inference can refer to a comparison of models based on predictive ability or an explicit combination of models to improve desired inference.

Any probability distributions (e.g., predictive distributions) can be averaged to form a new distribution, but the weights with which to average them are not unique, and any one set of weights is optimal only under certain circumstances. Bayesian methods provide the most coherent justification for model averaging because the optimal weights have been shown to equal the posterior model probabilities, $P(M_l|\mathbf{y})$, for $l=1, \ldots, L$ (Hoeting et al. 1999; Link and Barker 2006; Hooten and Hobbs 2015). The posterior model probability blends information from the model and data with a prior understanding of model suitability. Posterior model probabilities can be calculated easily for some Bayesian models, but they are intractable for others. Thus, BIC was developed for computing the optimal model averaging weights under certain conditions (equal prior model probabilities and flat prior distributions; Schwarz 1978).

The model averaging weights associated with BIC are proportional to $e^{-BIC_l/2}$, where BIC_l is the Bayesian information criterion calculated for model M_l. Burnham and Anderson (2002) suggested replacing BIC_l with AIC_l and using it in a non-Bayesian context to model average parameter estimates and predictions, a practice that has become popular in wildlife biology and ecology. However, model averaging should be performed only for quantities that do not change in their interpretation across models (Cade 2015; Banner and Higgs 2017). Regression parameters have different interpretations among models unless the predictor variables are uncorrelated; they are interpreted conditional on the other parameters in the model. However, predictions of data always have the same interpretation in models, so they can be safely averaged for final inference (Burnham and Anderson 2002; Burnham and Anderson 2004).

To illustrate model comparison based on predictive ability, we employed a Poisson GLM for count data and compared a set of five models including a variety of covariates. Across all methods we demonstrated, state area always appeared in the best predicting model. However, all of the covariates improve predictive ability beyond the null model (i.e., intercept only). In the models we considered, the information criteria provided similar insights as cross-validation, and the non-Bayesian information criteria agreed with the Bayesian approaches. This occurred because we used relatively vague priors for the regression coefficients, and thus, the information content from the data was approximately the same across models.

The regularization approaches also provided similar results for our data. However, while lasso immediately shrunk the simulated covariate to zero, ridge regression shrunk this covariate, and the others, more slowly. In this particular case, the smoothness of the trajectory allowed the regularization procedure to find a better predicting model (out of infinite possibilities) under the ridge penalty.

The algorithms required to fit the specific models we presented in this chapter, and its relatively small example data set, yielded nearly immediate results in all cases. However, for larger data sets, the cross-validation and Bayesian approaches may require more time to implement. Still, we used readily available software to fit all models and nested model-fitting commands within "for loops" to perform cross-validation when necessary. With modern computing resources and easy parallel computing, cross-validation can be sped up substantially with minimal extra effort, so the added computational burden is

not nearly as limiting now as it was in the past. However, while cross-validation automatically accounts for optimism in scoring predictive ability, it still depends on the initial data set and is limited in representing true out-of-sample predictive ability.

Finally, Ver Hoef and Boveng (2015) argue that there are valid situations that call for the use of a single, well-designed model that best represents the scientist's understanding of the ecological mechanisms and data collection process. In such cases, the emphasis is not on predictive ability, but rather on gaining a better understanding of the model components. A model component may be a simple population mean that is unknown, such as the average biomass in a survey plot, or it could be the true animal abundance in a closed study area. In these cases, we may have no need for model comparison because the desired inference is clear and the study design can be customized to answer these questions.

Acknowledgments

The authors thank David Anderson, Ken Burnham, Steve Ellner, Tom Hobbs, Jennifer Hoeting, Bill Link, Jay Ver Hoef, and Gary White. Hooten also acknowledges support from NSF EF 1241856, which helped fund this work. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

LITERATURE CITED

Akaike, H. 1983. Information measures and model selection. International Statistical Institute 44:277–291.

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

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference, 2nd ed. Springer-Verlag. New York, NY.

Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods and Research 33:261–304.

Cade, B. S. 2015. Model averaging and muddled multimodel inferences. Ecology 96:2370–2382.

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

Christensen, R. 2002. Plane answers to complex questions. Springer, New York.

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

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.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. Elements of statistical learning: Data mining, inference, and prediction, 2nd ed. Springer. New York.

Hoerl, A. E., and R. W. Kennard. 1976. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55–67.

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

Hobbs, N. T., and M. B. Hooten. 2015. Bayesian models: A statistical primer for ecologists. Princeton University Press, Princeton, NJ.

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

Kutner, M. H., C. J. Nachtsheim, and J. Neter. 2004. Applied linear regression models. McGraw-Hill/Irwin, New York.

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

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

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.

Park, T., and G. Casella. 2008. The Bayesian lasso. Journal of the American Statistical Association 103:681–686.

Richardson, S. 2002. Discussion of the paper by Spiegelhalter et al. Journal of the Royal Statistical Society, Series B 64:626–227.

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. American Statistician 66:77–84.

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.

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. 2015. Iterating on a single model is a viable alternative to multimodel inference. Journal of Wildlife Management 79:719–729.

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.