B

BAYESIAN STATISTICS

KIONA OGLE AND JARRETT J. BARBER
Arizona State University, Tempe

Bayesian statistics involves the specification of a joint probability model to describe the dependence among observable quantities (e.g., observed, or unobserved but predictable, data) and unobservable quantities (e.g., ecological model parameters, treatment effects, variance components). Inference about unobserved quantities is based on the posterior distribution (or posterior), which is the conditional probability distribution of the unobserved quantities given, or posterior to observing, the observed quantities. In principle, once a joint probability model is specified, the posterior follows via Bayes theorem, a well-known result from probability theory. In practice, except for relatively simple models, numerical methods must be used to approximate the posterior. Methods consist largely of algorithms to sample the unknown quantities from the posterior, hence effectively obtaining a histogram, whose approximation to the posterior improves as this sample size increases; inference is often reduced to summarizing this histogram with means, medians, and credible intervals. The majority of these numerical methods are broadly referred to as Markov chain Monte Carlo (MCMC) methods. MCMC has become so popular in the context of Bayesian statistics that it is often mistaken as synonymous with Bayesian, but many Bayesian statistical problems may be seen simply as an area of application of MCMC methodology with an objective to sample from a distribution of interest.

THE FUNDAMENTALS

Subjective Degrees of Belief, Probabilities, and Coherency

Aside from the well-established mathematical definition of probability, how do we interpret probability? We adopt the interpretation of probability as subjective or as a personal degree of belief about the occurrence of events. Among several other interpretations, the relative frequency interpretation is undoubtedly the most influential in statistics, wherein the probability of an event is viewed as the relative frequency of occurrence of the event in a large number of (hypothetical) repeated experiments. This view forms the basis of much of statistics as practiced over the past 80 years, including the hypothesis testing framework(s) of R. A. Fisher, J. Neyman, and E. S. Pearson. The subjective view allows us to specify probabilities for any events whatsoever, not just for those events for which the relative frequency interpretation is applicable (e.g., the event that it will rain tomorrow somewhere in the Great Plains of North America); yet it allows us to call upon frequentist, or other, notions of events to help us specify *our* probabilities.

Why adopt a Bayesian approach? Aside from more pragmatic considerations, a Bayesian approach (or a procedure consistent with a Bayesian approach) is essentially the only statistical approach to statistical inference that is coherent. Coherency refers to a calculus of (i.e., method for computing) degrees of belief that is self-consistent and noncontradictory. Coherency is often discussed in terms of gambling odds and illustrated by showing that a lack of coherency can result in a wager that ensures certain loss (or gain, depending on perspective; a so-called Dutch book). Alternatively, foundational developments of degrees of belief show that lack of coherency violates a set of

principles that many people find to be self-evident. And, if our degrees of belief and the manipulation thereof obey the coherency principle, then they coincide essentially with the modern mathematical definition of probability. Thus, in practice, we may simply use the term *probability,* and all of the properties that follow from the modern definition. In this entry, the term *quantification of uncertainty* is used frequently in place of *subjective probability* or *degree of belief.* For those who do not find coherency compelling, a Bayesian approach remains a way to implement the likelihood principle, which many people do find compelling. Finally, Bayes theorem provides a coherent mechanism for updating our prior probabilities with information in observed data.

Basic Probability Rules and Bayes Theorem

Bayesian statistics is based on a simple, uncontested probability result: Bayes theorem. To derive Bayes theorem, consider two random variables X and Y (i.e., quantities that are uncertain and whose potential values are described by a probability distribution), which, in the Bayesian context, may refer to data or to parameters. We are interested in the joint probability that X assumes value x and Y assumes value y, which we express as $P(X = x, Y = y)$, or, for convenience, $P(x, y)$. (Throughout, we use P to denote probability or probability density.) Given $P(x, y)$, then, in principle, we have its marginal distributions, $P(x)$ and $P(y)$, and its conditional distributions, $P(x \mid y)$ and $P(y \mid x)$. The marginal distribution of X, for example, describes the distribution of potential values after having summed or integrated over all possible values of Y; we might use the marginal distribution of X when we do not know about the other variable, Y. We would use the conditional distributions when we know the values of the conditioning variables, as in the distribution of x given y, $P(x \mid y)$, with y being the conditioning variable such that the conditional distribution of X depends on the particular value of $Y = y$. Using basic results, the joint probability distribution of two quantities can be written as the product of the conditional probability of one quantity given the other and the marginal probability of the other: $P(x, y) = P(x \mid y) P(y) = P(y \mid x)P(y)$.

The operational appeal of this result is that the statistical modeler may not know how to specify, or model, the joint distribution directly, but may know how to model a conditional distribution and a marginal. For example, we may be satisfied with a regression model of the distribution of tree height, Y, given tree diameter, X, and we may have a model for tree diameter. In this case, the above result ensures that our inferences are based on a valid joint probability distribution given by $P(y \mid x)P(x)$. The order of conditioning is left to the modeler, and it is important to realize that the joint model built from a specification of $P(y \mid x)$ and $P(x)$ generally is not the same as the joint model built from a specification of $P(x \mid y)$ and $P(y)$; this should not be confused with having a joint model, $P(x, y)$, for which both conditionals and both marginals are determined, in principle.

With a joint distribution specified, Bayes theorem offers a mechanism by which to implement the principle of conditional inference, i.e., inferring unknowns given knowns. Assuming we wish to infer about events involving X, given what we know, i.e., conditional on, $Y = y$, the object of our inference is the conditional distribution, $P(x \mid y)$. We simply rearrange $P(x \mid y)P(y) = P(y \mid x)P(x)$ to get Bayes theorem,

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}. \qquad (1)$$

Typically, $P(y \mid x)$ is a model for observed data y given unobserved parameters x, and $P(x)$ is a (prior) distribution, together, by the previous result, giving a full (joint) probability model for all quantities in the numerator of Equation 1. A simple application of Bayes theorem is illustrated in Box 1.

BOX 1. USING BAYES THEOREM WITH DISCRETE RANDOM VARIABLES

Consider the following example that illustrates the application of Bayes theorem to compute a posterior probability in a discrete setting. Assume you are hiking through a forest that has been affected by bark beetle infestations. For simplicity, we will define two discrete random variables: let B denote the beetle status of a tree in the forest ($B = 0$ if no evidence of beetle attack, $B = 1$ if evidence of beetle attack) and D the state of the tree ($D = 0$ if tree is living, $D = 1$ if tree is dead). You notice a dead ($D = 1$) tree in the distance, and you ask: "what is the probability that the tree was attacked by beetles?" That is, you wish to compute, for this tree, the following conditional probability: $P(B = 1 \mid D = 1)$. There is a deep ravine that prevents you from hiking to the tree to directly evaluate its beetle status. But, coincidently, you have with you a publication by a local researcher that reports the results of a census whereby he inventoried the beetle status and state of

BOX 1 *(continued)*.

thousands of trees in several nearby locations, giving the following probabilities:

$P(B = 1) = 0.3$ (i.e., 30% of the trees were attacked by beetles)

$P(D = 0|B = 0) = 0.8$ (i.e., of the trees that were not attacked, 80% were living)

$P(D = 0|B = 1) = 0.1$ (i.e., of the trees that were attacked, 10% were living)

You use this information to compute the probability of interest, $P(B = 1|D = 1)$. Using Bayes theorem, you have

$$P(B = 1|D = 1) = \frac{P(D = 1|B = 1) \cdot P(B = 1)}{P(D = 1)} \quad (1.1)$$

From the inventory publication, you know $P(B = 1)$ and you compute $P(D = 1|B = 1) = 0.9$ since $P(D = 1|B = 1) + P(D = 0|B = 1) = 1$, and you are given $P(D = 0|B = 1)$. Thus, the numerator in Equation 1.1 is equal to $0.9 \times 0.3 = 027$. You are not given $P(D = 1)$, but using basic probability rules, you compute it as

$$\begin{aligned} P(D=1) &= \sum_{b=0}^{1} P(D = 1, B = b) = \sum_{b=0}^{1} P(D = 1|B = b) \cdot P(B = b) \\ &= P(D = 1|B = 0) \cdot P(B = 0) \\ &\quad + P(D = 1|B = 1) \cdot P(B = 1) \quad (1.2) \\ &= (1 - 0.8) \times (1 - 0.3) + 0.9 \times 0.3 \\ &= 0.2 \times 0.7 + 0.9 \times 0.3 = 0.41 \end{aligned}$$

Thus, the probability of interest is $P(B = 1|D = 1) = 0.27/0.41 \cong 0.659$. And the probability that the tree was not attacked by beetles is $P(B = 0|D = 1) \cong 1 - 0.659 = 0.341$.

Here, we can think of $P(B = 1)$ as the prior probability that a tree was attacked by beetles, which is simply based on the inventory data and the overall proportion of trees that were attacked; $P(D = 1|B = 1)$ as the likelihood of a tree being dead if it was attacked by beetles; and $P(B = 1|D = 1)$ as the updated or posterior probability that the tree was attacked by beetles given that it is dead.

In the Bayesian statistical context, we denote observed quantities by *Data*, unobserved quantities by θ. In Equation 1, θ plays the role of x, *Data* the role of y. In principle, θ includes all unobserved quantities about which we wish to make inference. This may include unobservable parameters as in, say, a regression model, or unobserved but potentially observable data values that we may wish to infer (e.g., missing or future observations). Thus, Bayes theorem is used to obtain the conditional probability distribution of θ given observations *Data*:

$$P(\theta|Data) = \frac{P(Data|\theta)\ P(\theta)}{P(Data)}. \quad (2)$$

Here, $P(\theta|Data)$ is the posterior (probability) distribution of θ given *Data*, which may be a density function, a mass function, or a combination of both, depending on the nature of the components of θ. (Often, as in a regression setting, covariates are assumed known and are notationally suppressed.) Because inference proceeds conditional on *Data*, and because the likelihood function associated with $P(Data|\theta)$ also is viewed as a function of θ given *Data*, $P(Data|\theta)$ is often referred to as the likelihood. This may cause confusion among those who are familiar with the notion of a likelihood function: if we call $P(Data|\theta)$ the likelihood, this suggests conditioning on *Data*. This perspective, however, is inconsistent with the interpretation of the numerator of Equation 2 as a joint distribution of (*Data*, θ) being built from the conditional distribution of *Data* given θ. For this reason, we prefer to call $P(Data|\theta)$ the data model. Here, $P(\theta)$ is the prior (probability) distribution for θ, which quantifies our prior understanding or uncertainty about θ before observing *Data*, and $P(Data)$ is the prior predictive distribution of *Data* because this is the distribution we would use to predict *Data* before it is observed; we cannot use $P(Data|\theta)$ directly, because θ is unknown. Uncertainty about θ is incorporated into the prior predictive by marginalizing over θ with respect to $P(\theta)$, giving $P(Data) = \int P(Data|\theta)P(\theta)d\theta$ if θ is a continuous random variable.

If we want to predict unobserved data, denoted *Data′*, we use Bayes theorem in Equation 2 to obtain the posterior of all unobserved quantities, now (*Data′*, θ), conditional on observed *Data*. That is, $P(Data', \theta|Data) \propto P(Data', Data|\theta)P(\theta)$. Note that we follow the common practice of denoting unobserved data separately from other unknowns in θ, though this is done merely for convenience of interpretation, and all unknowns may be denoted by θ. With this latter note, we see the *posterior predictive* distribution simply as the marginal posterior distribution, $P(Data'|Data) = \int P(Data', \theta|Data)d\theta$. If our data model is constructed hierarchically as $P(Data', Data|\theta) = P(Data'|Data, \theta)P(Data|\theta)$, then we may write $P(Data'|Data) = \int P(Data'|Data, \theta)\ P(\theta|Data)d\theta$, which simplifies to $\int P(Data'|\theta)\ P(\theta|Data)d\theta$ if *Data′* and *Data* are independent.

Sequential processing of data is, in principle, straightforward with Bayes theorem and is one of the strengths of this approach. Assume we observe *Data* followed by *Data′* at a later time. We may obtain the posterior $P(\theta|Data)$, then, when *Data′* is available, simply use

$P(\theta|Data)$ as the prior for θ in $P(Data'|Data, \theta)$, updating the posterior to $P(\theta|Data, Data')$. That is,

$$P(\theta|Data', Data) = \frac{P(Data'|Data, \theta)P(\theta|Data)}{P(Data'|Data, \theta)}. \quad (3)$$

We recognize the denominator as the posterior predictive distribution for *Data′*, given the first data set, *Data*. The above procedure can be repeated indefinitely as data arrive. Or, we may choose to obtain the posterior at once when both data sets are available; the result is the same posterior, $P(\theta|Data, Data')$, which can be shown by applying Bayes theorem again with a bit of algebra.

Except for relatively simple applications, the integral in the denominators in Equation 2 or 3 typically is not available analytically, and, hence, an analytical solution for the posterior is not available. In practice, this generally is not an issue, because once we have observed *Data* and defined $P(Data|\theta)$ and $P(\theta)$, $P(Data)$ is a normalizing constant (or constant of proportionality) with respect to the unknown quantity, θ, and many algorithms are available to sample from distributions known up to a constant, thereby allowing us to approximate the actual distribution (see the section on numerical methods below). Thus, the joint distribution defined by the numerator of Equation 2 is often the main focus of modeling efforts. For this reason, the posterior is often expressed as

$$P(\theta|Data) \propto P(Data|\theta)\ P(\theta). \quad (4)$$

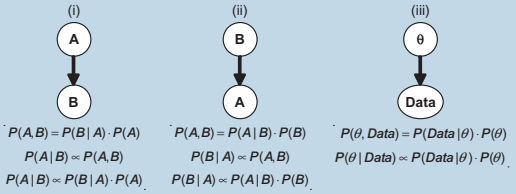
We also refer to the right-hand side of Equation 4 as the unnormalized posterior since division by $P(Data)$ is required for $P(\theta|Data)$ to integrate (θ continuous) or sum (θ discrete) to 1. Equation 4 is the full probability model for observable and unobservable quantities and is the essence of a Bayesian statistical model.

GRAPHICAL MODELS

Graphical models, especially directed acyclic graphs (DAGs), are useful for depicting and aiding the construction of probability models. In particular, DAGs are useful for building complicated, full probability models from relatively simple, conditionally independent components (Box 2). Indeed, graphical models play an important role in the development of the popular software programs WinBUGS and OpenBUGS. We introduce only enough material to write an expression for a full probability model in terms of conditionally independent model components and to aid in the construction of full conditional distributions, to which we return later.

BOX 2. GRAPHICAL MODELS TO FACILITATE CONSTRUCTION OF BAYESIAN MODELS

We can use a directed acyclic graph (DAG) to describe the conditional dependencies between model quantities. We may find it easier or more appropriate to express *B* as conditional on *A* (or depends upon *A*) as in DAG (i). Conversely, to express *A* as conditional on *B*, we would draw DAG (ii). In (i), we may refer to *A* as parameters (θ) and *B* as data (*Data*) in our model, and thus DAG (i) becomes (iii) in our typical notation. The DAG defines the joint distribution for the quantities of interest (e.g., *A* and *B*, or θ and *Data*) as the product of conditional and marginal probability distributions. In a Bayesian data analysis, we are interested in the posterior distribution of some quantities (e.g., θ) given other quantities (e.g., *Data*), and this conditional distribution, $P(\theta|Data)$, is proportional to the joint distribution described by the corresponding DAG in (iii).



The DAG consists of circular or elliptical nodes that represent different stochastic quantities in the model that are described by probability distributions (see "Graphical Models" section). Square or rectangular nodes represent fixed quantities (e.g., covariate data that we may assume to be measured without error). The nodes may be referred to as child or parent nodes (see the "Graphical Models" section) and as root, internal, or terminal nodes. A root node is a node that does not have parents, e.g., *A* in (i) and θ in (iii); a terminal node is a node that does not have any children, e.g., *A* in (ii) and *Data* in (iii); and an internal node is a node that gives rise to child nodes and that has its own parent nodes. The nodes are connected by unidirectional edges (or arrows) that indicate the conditional dependency between two nodes *A*.

A graph consists of a set of nodes, typically depicted by circles and representing stochastic quantities, including data (before they are observed) and parameters; relationships between nodes are specified with a set of directed edges or arrows. A DAG is directed because edges have single tips and is acyclic because following the arrows never allows a node to be revisited. Sometimes, squares are used to denote fixed quantities that do not receive a stochastic specification (see Box 3), like covariates in a regression, but such fixed quantities are typically not

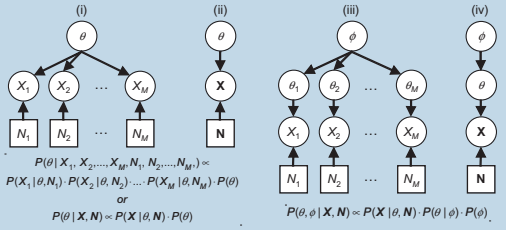
relevant to the qualitative probabilistic properties represented by the DAG and are often notationally suppressed. Also, some authors use squares to denote observed values of stochastic quantities, but circles are used here.

Let v denote a node in the graph, and let V denote the set of all nodes. Referring to our previous notation, V includes θ and *Data* (before it is observed; Box 2). Define a parent of v as any node having an arrow emanating from it and pointing to v. Denote the set of parents of v as $parents[v]$, and let V_{-v} denote all nodes except v. Now, an otherwise complicated full (joint) probability model of all stochastic quantities can be written as a product of relatively simple, conditionally independent components,

$$P(V) = \prod_{v \in V} P(v \mid parents[v]). \tag{5}$$

BOX 3. DAGS FOR THE TREE MORTALITY EXAMPLE

The Bayesian model for the beta-binomial tree mortality example can be expressed as a DAG in (i). In this example, the DAG is associated with two levels: level 1 is the stochastic data level for the X's, and level 2 is the parameter level (representing the prior for θ). In (i), each observation is explicitly indicated such that separate nodes are shown for the number of trees that survived (X_j) and the total number of trees (N_j) in each plot j, $j = 1, 2, \ldots, M$. As the model becomes more complicated, with more levels and more quantities, the DAG can become complicated. Thus, we may express the DAG in (i) in a more compact form whereby **X** and **N** denote the vectors of all data on the number of trees that survived and the total number of trees, respectively: DAG (ii).



The above model and DAGs (i) and (ii) assume the probability of mortality (θ) is the same for all plots. However, θ may vary by plot due to plot-level differences in, for example, stand density, soil properties, and exposure. Thus, we may want to include plot specific θ's such that $\theta = (\theta_1, \theta_2, \ldots, \theta_M)$. We may specify a hierarchical prior for the θ's and treat each θ_j as coming from a parent population defined by hyperparameters (e.g., ϕ), as in (iii) and (iv); see the "Hierarchical Bayesian" section. This model has three levels: level 1 is the stochastic data level, level 2 is the first parameter level (here, for θ), and level 3 is the second parameter level (representing the hyperprior for ϕ).

In the tree mortality example (see Box 3), $parents[X_i] = \{\theta\}$, and the factors corresponding to the right-hand side of $P(V)$ above are $P(X_i \mid \theta, N_i)$. Note that the N_i have no bearing on the conditional independence of the X_i, since the N_i are fixed covariates, and we may also write $P(X_i \mid \theta)$. Incidentally, if the N_i were stochastic, still, the X_i would be conditionally independent given $parents[X_i] = \{\theta, N_i\}$ in this case.

We should realize that $P(V)$ is the numerator in Bayes theorem, before any conditioning on observations is done, and we will likely not recognize its form, except in relatively simple models. In this case, we may attempt to approximate the posterior via numerical methods (discussed below), for which we need the conditional distributions of subvectors of V or of individual nodes v, which are called full conditional distributions or just full conditionals. For this, let the children of node v, denoted $children[v]$, be those nodes that are pointed to directly by the arrows emanating from v. Then the form of the full conditional for v is given by

$$\begin{aligned} P(v \mid V_{-v}) &\propto P(v, V_{-v}) \\ &\propto (\text{factors in } P(V) \text{ containing } v) \\ &= P(v \mid parents[v]) \prod_{w \in children[v]} P(w \mid parents[w]). \end{aligned} \tag{6}$$

In other words, to obtain (an expression proportional to) the full conditional of v, simply look at the right-hand side of the conditional representation of $P(V)$ in Equation 5, which, again, occurs as the numerator of Bayes theorem, and use only those factors depending on v. Note that this works for node v, which may be a vector, or for collections of nodes, but application of numerical methods (discussed below) to sample larger vectors is generally more challenging in practice.

Estimation and Testing

In principle, the posterior, or its approximating histogram (discussed below), contains all of the information we need for inference. We may compute various summaries such as means, medians, modes, quantiles, variances, and intervals or regions containing θ with specified probability (credible intervals). Still, how "good" are these summaries? To answer this question, we introduce the decision theoretic notions of a loss function and minimum expected loss as a goodness criterion.

Let θ be the quantity we wish to estimate, and let $\delta(y)$ denote the procedure, which depends on data, y, that we use to estimate θ. We call δ an estimator (of θ), and, for a particular value of y, $\delta(y)$ is an estimate of θ. Note

that δ may be referred to as a (decision) procedure, an estimator, or a rule. The objective is to find an estimator that is somehow optimal. For this, we introduce a loss function, $L(\theta, \delta(y))$, defined on the Cartesian product of the ranges of θ and $\delta(y)$, which are typically the same. The loss function assumes nonnegative values, with greater values indicating a larger discrepancy, or loss, between target, θ, and estimator, $\delta(y)$. The loss function is almost never sufficient to allow us to choose a best estimator since, in general, the loss depends on θ and y, and we don't know θ. For some θ and y, δ_1 may minimize loss, and, for other values, δ_2 may minimize loss. To address this ambiguity, we may compute an average or expected loss.

In particular, we may average $L(\theta, \delta(y))$ with respect to $P(y|\theta)$ to get frequentist risk, the objective being to find δ that minimizes frequentist risk. Still, frequentist risk depends on the unknown θ, and additional criteria, such as admissibility or minimaxity, which we do not define here, may be used to obtain an optimal estimator. Or, unbiasedness may be introduced to help find an optimal estimator, though this criterion is relatively uncommon from a decision theoretic point of view.

For a familiar example, consider squared error loss, $L = (\theta - \delta(y))^2$, the most ubiquitous loss function. (Assume θ and δ are scalar valued for simplicity of presentation.) Then, frequentist risk is commonly known as mean square error (MSE), and, if θ is the expected value of y, frequentist risk reduces to the variance of y. In some cases, $\delta(y) = y$ minimizes frequentist risk and is the best unbiased estimator of its mean, the sample average being best unbiased, in many cases, if we have a sample of y values from $P(y|\theta)$.

Alternatively, we may average $L(\theta, \delta(y))$ with respect to $P(\theta|y)$ to get posterior expected loss. Again, the optimality criterion is to find the procedure, δ, to minimize expected loss. Posterior expected loss still depends on y, but this is less of a problem than dependence on θ since we observe y. This criterion says, for each y, choose the procedure $\delta(y)$ that minimizes posterior expected loss. Such a procedure is called a Bayes rule, procedure, or estimator. (To avoid confusion, note that Bayes theorem is sometimes called Bayes rule.) It turns out that Bayes rules are usually admissible estimators, and admissible rules are Bayes rules or limits of Bayes rules. Thus, in principle, a frequentist (using frequentist risk) looking for admissible rules may adopt a Bayesian approach.

As yet another alternative, we may choose to average $L(\theta, \delta(y))$ over both y and θ (assuming we introduce a prior) to obtain integrated (frequentist) risk. It can be shown that the procedure minimizing integrated risk is the same as the Bayes rule. Thus, again, in principle, a frequentist may adopt a prior and use the criterion of posterior expected loss to find an optimal estimator in terms of integrated risk, called Bayes risk when evaluated at the Bayes rule.

The upshot of the current discussion is that common summaries of the posterior are often Bayes rules with respect to some loss function and using the posterior expected loss criterion. For example, the posterior mean is the Bayes rule under squared error loss, and the posterior median is the Bayes rule under absolute error loss, $|\theta - \delta(y)|$. Similar results hold for the posterior mode, quantiles, and other summaries of the posterior for other losses.

For testing hypotheses about θ, Bayes rules (decisions) to reject or accept a hypothesis can be framed as an estimation problem, leading again to minimizing posterior expected losses. But, roughly speaking, hypothesis testing in a Bayesian framework is a relatively delicate matter, especially with simple point null hypotheses like H_0: $\theta = \theta_0$. In some sense, the Bayes factor, which is the ratio of posterior to prior odds, may be considered a Bayesian response to the ubiquitous p-value in frequentist significance testing.

PRIORS

The prior distribution has been criticized for not being objective, but the data model and, perhaps, loss are subjective components of a Bayesian analysis as well. Likewise, the choice of the likelihood in a classical analysis is subjective. When we do have prior information about θ, then a Bayesian analysis is difficult to ignore. When we have no prior information, we may choose a noninformative or reference prior if we wish to exploit the conditional nature of Bayesian inference or if we prefer to think (and behave coherently) in terms of probabilities. In this case, or when prior information is vague, then a sensitivity analysis is appropriate, wherein we adjust the prior within a reasonable range to determine the effects on the posterior. If the posterior exhibits little sensitivity to the prior, perhaps being dominated by a large data set, then we may feel satisfied with our analysis. In cases where the posterior is sensitive to a vague prior, then we may want to work harder to obtain prior information, more or better data, or both. The literature on prior distributions is vast, and here we highlight notions of (conditionally) conjugate priors, noninformative priors, and prior and posterior propriety.

A prior is said to be conjugate with respect to some class of distributions if the form of the posterior is the same as the prior. Hence, the form of the posterior is known, and the computation of the posterior is usually a matter of computing a few parameters with simple formulae. For a large class of commonly used data models (exponential families), we can, in principle, get arbitrarily close to any prior using a mixture of conjugate priors. So, in principle, conjugate priors can be useful approximations to our true prior, and they permit useful analytical simplifications. In the case where we do not have prior information, use of conjugate priors is often for analytical simplicity or convenience. Aside from analytical conveniences, conjugate priors often aid interpretation via the "device of imaginary observations." That is, they can be interpreted as contributing prior information in the form of an imaginary sample size; this is discussed in greater detail in the simple example given below.

Only in the simplest cases, as in the example below, can we hope to use conjugacy to obtain the posterior. In more complicated models, we may not have overall conjugacy, but may still retain conditional conjugacy. In this case, a prior can be chosen for a subvector of θ so that the resulting full conditional distribution is of the same form as the prior, and, hence, the full conditional may be sampled from directly (see the numerical methods section, below). In the context of the DAG notation, this is a matter of looking at the factors of $P(V)$ that include v, ignoring the prior, $P(v)$, for the moment, then choosing $P(v)$ so that the form of the full conditional $P(v|V_{-v})$ is the same as that of $P(v)$ (generally with different parameter values, of course). For example, we may have a logistic regression model with linear predictor $\alpha + \beta x + \varepsilon$, for some unknown parameters α and β, fixed covariate x, and error $\varepsilon \sim N(0, \sigma^2)$. In this case, an inverse-gamma prior for σ^2, independent of the prior for α and β, is conditionally conjugate to the factor (containing σ^2) arising from the normal distribution in the linear predictor, yielding an inverse-gamma full conditional for σ^2. Typically, in practice, we simply consult a catalog of conjugate data model prior pairs rather than actually implement this procedure for finding a conjugate prior; the work has already been done for us in many cases.

In many cases, we do not have prior information, and we may appeal to noninformative priors. Intuitively, noninformative priors are meant to embody a lack of information about a parameter. There are various senses in which this may be true, and only Jeffreys prior is mentioned here. Jeffreys prior is obtained from the Fisher information matrix, $I^F(\theta)$, or number as $P(\theta) \propto |I^F(\theta)|^{1/2}$, where $|\mathrm{A}|$ denotes the determinant of matrix A. Under some technical conditions—often referred to as regularity conditions—which hold for many familiar distributions, including exponential families, $I^F(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log(P(y|\theta))\right]$. For example, Jeffreys (joint) prior for the mean and variance, $\theta = (\mu, \sigma^2)$, of a normal distribution is $P(\theta) \propto (\sigma^2)^{-3/2}$, which, incidentally, is not the same as the product of Jeffreys priors obtained for the mean and variance, separately, $P(\theta) \propto 1/\sigma^2$, which we may use if we are thinking that the mean and variance are *a priori* independent. Jeffreys prior is invariant to transformation. Thus, specifying Jeffreys prior for θ and then obtaining the prior for the transformation $\eta = h(\theta)$ is the same as specifying Jeffreys prior for η. Jeffreys prior represents an ad hoc method for obtaining a prior and, technically, is outside the realm of Bayesian statistics since it involves expectation over unobserved data values, y, violating the likelihood principle by not conditioning on observed data, hence, strictly speaking, violating the Bayesian paradigm.

In what sense is Jeffreys prior noninformative? Fisher's information is a commonly used measure of information about the parameters contained in the data model. A larger information number discriminates θ from $\theta + \Delta\theta$ more than a smaller information number, and similarly for $|I^F(\theta)|^{1/2}$. Thus, choosing a prior proportional to $|I^F(\theta)|^{1/2}$ is noninformative in the sense of not changing the discriminating action of this measure of information. Jeffreys prior can also be thought of as the prior that is equally noninformative for all transformations of θ.

The Jeffreys priors given in the normal examples above are improper distributions, which means that their integrals are infinite, as is often the case with Jeffreys priors. If the prior is improper, then the interpretation of a full probability model, conditional distribution, and marginal distribution no longer holds technically. However, the posterior may still be proper, which is the most important thing to check when specifying improper priors. Note also that improper priors can create problems with Bayes factors, and we recommend avoiding improper priors in this case unless much care is taken to avoid their pitfalls when computing Bayes factors.

A SIMPLE EXAMPLE

Here is a simple example to illustrate some of the concepts and equations presented above. Refer to the DAGs in Box 3 (i and ii) as we develop the example. Consider the problem of estimating the probability, θ, of a tree

dying during a severe drought. A researcher conducts a study and counts the number of trees that died, X_i, and the total number of trees, N_i, in $i = 1, 2, 3, \ldots, M$ randomly located plots within a forested region. We define the different components or levels of the model, which in this simple example means we define the data model for the data, $P(\boldsymbol{X}|\theta, \boldsymbol{N})$, and the prior for θ, $P(\theta)$. We treat $\boldsymbol{N}$ as a known, fixed covariate.

If we assume θ is the same for all plots and the number of trees that died in plot $i(X_i)$ is independent of the number that died in other plots, then the logical choice for a data model is a binomial distribution for the X_i, and we write

$$X_i|\theta \sim binomial(\theta, N_i). \tag{7}$$

In defining the model for X_i, we can ignore the factors that do not depend on θ, expressing the data model as proportional to the kernel of the binomial pmf (i.e., the factors that contain θ):

$$\begin{aligned} &P(X_i = x_i|\theta, N_i) \\ &\quad = \binom{N_i}{x_i}\theta^{x_i}(1-\theta)^{N_i-x_i} \propto \theta^{x_i}(1-\theta)^{N_i-x_i}. \end{aligned} \tag{8}$$

Next, we may assume that the X_i are conditionally independent given θ, and the complete data model or the likelihood of all data is

$$P(\boldsymbol{X}|\theta, \boldsymbol{N}) \propto \prod_{i=1}^{P}\theta^{x_i}(1-\theta)^{N_i-x_i} = \theta^{\sum_{i=1}^{P}x_i}(1-\theta)^{\sum_{i=1}^{P}(N_i-x_i)}. \tag{9}$$

Of course, this does not mean that the X_i are (unconditionally) independent, because counts of dead trees depend on θ.

Next, we define $P(\theta)$, and, in doing so, we should think about constraints on θ. In this example, θ is a probability parameter, and we should consider picking a prior that is defined for $0 \leq \theta \leq 1$. We consider a conjugate prior, and, without previous experience, we seek to identify a pdf (θ is a continuous random variable) with a kernel that matches the form of the likelihood in Equation 9, which will give a posterior of the same form as the prior (see the previous discussion about obtaining conjugate priors). In this case, the conjugate prior is the beta distribution, and we write $\theta \sim beta(\alpha, \beta)$, where α and β are the (hyper)parameters for which we assign specific values shortly. The beta pdf is

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{10}$$

Again, since we may ignore the factors that do not contain θ, we can express the prior as proportional to the kernel of the beta pdf.

Finally, we combine the data model and prior via Bayes theorem to arrive at the posterior for θ:

$$\begin{aligned} P(\theta|\boldsymbol{X}, \boldsymbol{N}) &\propto P(\boldsymbol{X}|\theta, \boldsymbol{N})P(\theta) \\ &\propto \left(\theta^{\sum_{i=1}^{P}x_i}(1-\theta)^{\sum_{i=1}^{P}(N_i-x_i)}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) \\ &= \theta^{\sum_{i=1}^{P}x_i+\alpha-1}(1-\theta)^{\sum_{i=1}^{P}(N_i-x_i)+\beta-1}. \end{aligned} \tag{11}$$

Via kernel matching, we recognize the unnormalized posterior as the kernel of a beta such that

$$\theta|\boldsymbol{X}, \boldsymbol{N} \sim beta\left(\sum_{i=1}^{P}x_i + \alpha, \sum_{i=1}^{P}(N_i - x_i) + \beta\right). \tag{12}$$

The role of α and β should now be clear, but their interpretation may vary slightly. Returning to the "device of imagining observations," if we examine Equation 12, α is equivalent to the imaginary or prior number of trees that died, and β is equivalent to the prior number of trees that survived; $\alpha + \beta$ is interpreted as the prior sample size. If we examine Equation 11 from the perspective of the likelihood, then $\alpha - 1$, $\beta - 1$, and $\alpha + \beta - 2$ may be interpreted as the prior number of trees that died, the prior number that survived, and the prior sample size, respectively.

We can also compare properties of the prior and posterior to further evaluate the role of the prior. For example, the prior and posterior means for θ, $E(\theta)$ and $E(\theta|\boldsymbol{X}, \boldsymbol{N})$, respectively, and for comparison, the maximum likelihood estimate (MLE) for θ ($\hat{\theta}_{MLE}$) are

$$\begin{aligned} E(\theta) &= \frac{\alpha}{\alpha+\beta},\ E(\theta|\boldsymbol{X}, \boldsymbol{N}) = \frac{\sum_{i=1}^{P}x_i + \alpha}{\sum_{i=1}^{P}N_i + \alpha + \beta}, \\ \hat{\theta}_{MLE} &= \frac{\sum_{i=1}^{P}x_i}{\sum_{i=1}^{P}(N_i - x_i)}. \end{aligned} \tag{13}$$

$E(\theta|\boldsymbol{X}, \boldsymbol{N})$ may be viewed as a compromise between $E(\theta)$ and $\hat{\theta}_{MLE}$, and $E(\theta|\boldsymbol{X}, \boldsymbol{N}) = \hat{\theta}_{MLE}$ if $\alpha = \beta = 0$. The beta pdf is undefined for $\alpha = \beta = 0$, resulting in an improper (Haldane) prior, but the posterior is proper as long as $\sum x_i \neq 0$ and $\sum x_i \neq \sum N_i$. If we want a proper, noninformative prior for θ, we may chose a uniform prior on the interval (0, 1), equivalent to $beta(1, 1)$. This would contribute an effective prior sample size of 2 ($\alpha = 1$ dead tree, $\beta = 1$ living), with $E(\theta) = 0.5$. The Jeffreys prior, $beta(\frac{1}{2}, \frac{1}{2})$, would contribute a prior

sample size of 1. Both priors are conjugate and contribute some amount of information, but their influence will depend on the data, $\sum N_i$ and $\sum x_i$. If $\sum x_i$ and $\sum N_i$ are "large," then the data overwhelm the prior and $E(\theta \mid \boldsymbol{X}, \boldsymbol{N})$ will be very close to $\hat{\theta}_{MLE}$.

HIERARCHICAL BAYESIAN

The simple model in Equation 4 can be extended to many levels by, for example, incorporating hierarchical parameter models. For example, in the tree mortality example presented previously, we may wish to model the probability of mortality (θ) as varying by plot i to account for plot-to-plot variability associated with differences in environmental conditions or to include information about spatial correlation in the θ_i. One approach is to model the θ_i as coming from a population of θ's, with the population distribution described by hyperparameter ϕ, introducing a third level to the DAG (see Box 3, iii and iv). Such hierarchical parameter models are often relevant for modeling data obtained from nested sampling designs or representing different levels of aggregation. Here, we specify a joint prior for $P(\theta, \phi)$, and, based on the DAG and probability rules, we can write this as $P(\theta \mid \phi)P(\phi)$.

The simple Bayesian framework can also be extended to explicitly partition measurement or observation error from ecological process uncertainty, adding another level to the model. This extension is often referred to as a hierarchical Bayesian model (although, the previous example also results in a hierarchical model) or the process sandwich. Here, we define *Process* as the underlying latent, unobservable process that generates the data, and we partition θ into θ_D ("data-related" parameters) and θ_P ("process-related" parameters) and write

$$P(\theta_D, \theta_P, Process \mid Data) \propto P(Data \mid Process, \theta_D) \times P(Process \mid \theta_P)P(\theta_D, \theta_P). \quad (14)$$

The data model is now given by $P(Data \mid Process, \theta_D)$, and for simplicity, we can think of the data as varying around the latent process plus observation error such that the expected value of the data, $E(Data)$, is equal to (some function of) *Process*, and θ_D typically contains parameters describing measurement error variances, covariances, and/or biases (e.g., instrument drift). Here, *Process* is a stochastic quantity, and $P(Process \mid \theta_P)$ is referred to as the stochastic process model; we can think of the true, latent process as varying around some expected process, $E(Process)$, plus process error. Thus, θ_P may contain parameters in the $E(Process)$ model as well as process error (co)variance terms. Note that $E(Process)$ is typically deterministic, conditional on θ_P, and may be described by different types of models ranging from, for example, a simple linear regression to more complex models such as a difference equation or matrix model of population dynamics, a differential equation describing the spread of a pathogen or invasive species, or a nonlinear, biochemical-based model of plant photosynthesis. Thus, it is via $E(Process)$ that we have the opportunity to explicitly incorporate ecological theory/models into the probabilistic Bayesian framework.

In the vast majority of applications, once we have conditioned the data on the latent process, it is fair to assume that the measurement or observation errors are independent. Likewise, one approach to modeling process errors is to assume that they are independent. However, the assumption of independence for both error components often leads to identifiability problems such that process and observation variance terms cannot be separated. One solution is to specify a tight prior for the variance terms associated with observation error, which is reasonable when existing information is available. If such information is lacking, identifiability of the different variances terms may be facilitated by incorporation of different error structures for each component. Since we cannot measure and account for all factors affecting *Process*, we cannot model it perfectly via $E(Process)$, and we are led to consider modeling the process error structure. These unobserved factors are almost invariably temporal, spatial, or biological in nature, thus leading us to incorporate temporal, spatial, or biological structure into the process errors.

NUMERICAL METHODS AND MARKOV CHAIN MONTE CARLO (MCMC)

In practice, an analytical solution for the posterior typically is unavailable because most realistic models are nonlinear or otherwise too complex to permit easy calculation of the joint posterior. But this typically is not a problem, because effective computational strategies are available for sampling from the joint posterior despite such complexities. For models that involve conjugate priors, we may obtain analytical summaries or straightforwardly simulate values of θ from the posterior. If we cannot simulate directly from the joint posterior, we can simulate directly from known full conditionals, particularly if we are able to specify conditionally conjugate priors; this is known as Gibbs sampling. However, when Gibbs does not perform well, or when we do not recognize the form of the full

conditionals, we may be able to use other methods for simulating from the posterior, such as adaption–rejection, Metropolis–Hastings, or slice sampling. Nearly all algorithms work with the normalized and/or unnormalized full conditionals for the unknowns. Consider the DAG in Box 3iii, which indicates the full joint posterior as

$$P(\theta_1, \theta_2, \ldots, \theta_M, \phi \mid X_1, X_2, \ldots, X_M, N_1, N_2, \ldots, N_M) \propto P(X_1 \mid \theta_1, N_1) \cdot P(X_2 \mid \theta_2, N_2) \cdot \ldots \cdot P(X_M \mid \theta_M, N_M) \cdot P(\theta_1 \mid \phi) \cdot P(\theta_2 \mid \phi) \cdot \ldots \cdot P(\theta_M \mid \phi) \cdot P(\phi). \quad (15)$$

If we are interested in a particular θ_j, we can work directly with its full conditional, and, since it only depends on X_j, N_j, and ϕ, we can simply write the full conditional for θ_j as

$$P(\theta_j \mid \phi, X_j, N_j) \propto P(X_j \mid \theta_j, N_j) \cdot P(\theta_j \mid \phi). \quad (16)$$

For the example in Box 3iv, we choose the binomial pmf for $P(X_j \mid \theta_j, N_j)$ and a conjugate beta prior for $P(\theta_j \mid \phi)$ with hyperparameter $\phi = (\alpha, \beta)$. Thus, the full conditional for θ_j is recognizable: it is a beta distribution, and we can use Gibbs to sample directly from it. However, we cannot identify conjugate priors for α and β based on knowledge of common distributions. And we do not recognize the full conditional for $\phi = (\alpha, \beta)$, so we cannot use Gibbs to simulate values of ϕ. In this case, we may use Metropolis–Hastings (M–H) or some other algorithm that will allow us to sample from the unnormalized full conditional.

Both the Gibbs and M–H algorithms are part of a more general class of Markov chain Monte Carlo (MCMC) algorithms. The general idea behind MCMC methods is to simulate a sequence of values, e.g., $\theta^1, \theta^2, \ldots, \theta^T$, from the joint posterior. The draws are from a Markov chain because the probability of drawing a particular value of θ at iteration t depends on the value of θ at iteration $t - 1$. The general procedure is to sample θ sequentially from a proposal distribution that may depend on the last value of θ in the sequence, on the data, or both, and the MCMC draws will eventually approximate or converge to the target (posterior) distribution in that the histogram formed from these values can be made arbitrarily close to the actual posterior with increasing T.

To illustrate an MCMC approach, a simple M–H algorithm is outlined in Box 4. The M–H algorithm is based on an accept/reject rule, and it requires us to specify a proposal distribution, Q. If we know the normalized full conditional for θ, $P(\theta \mid \phi, X, N)$, then we may use this for Q, though this may not be the optimal choice in terms of MCMC behavior. In this case, if we let $Q_t(\theta_j^* \mid \theta_j^{t-1}) = P(\theta_j^* \mid \phi^{t-1}, N_j, X_j)$, then r in Equation 2.1 (Box 4) reduces to $r = 1$, and we accept every proposed value, yielding the Gibbs algorithm, a special case of the M–H algorithm.

Key issues common to both the M–H and Gibbs samplers, and other MCMC methods, include choosing starting values (e.g., θ^0; Box 4), evaluating convergence and burn-in, and determining the length of the simulation, T. One must understand and consider these issues when implementing MCMC algorithms, but it is beyond the scope of this chapter to define these terms and discuss these issues; we refer the reader to relevant texts

BOX 4. GENERAL METROPOLIS–HASTINGS (M–H) ALGORITHM

The general M–H sampling algorithm is outlined as follows. Let θ be a vector of P elements, and denote θ_j as jth element and θ_j^0 as the starting value for θ_j in the MCMC sequence. The M–H algorithm proceeds as follows:

- For $t = 1, 2, 3, \ldots T$ iterations,
  - For $j = 1, 2, \ldots, P$ parameters,
    - Sample a proposal value θ_j^* from a proposal distribution $Q_t(\theta_j^* \mid \theta_j t-1)$.
    - Calculate the acceptance ratio as
      $$r = \frac{P(\theta_j^* \mid \phi, X, N) \cdot Q_t(\theta_j t-1 \mid \theta_j^*)}{P(\theta_j^{t-1} \mid \phi, X, N) \cdot Q_t(\theta_j^* \mid \theta_j^{t-1})} \quad (2.1)$$
    - Set $\theta_j^t = \begin{cases} \theta_j^* & \text{With probability min } (1, r) \\ \theta_j^{t-1} & \text{otherwise} \end{cases}$ (2.2)

There are four main points to note. First, the accept/reject rule results in the proposed value being accepted if it increases the posterior density relative to the previous value (i.e., when $r > 1$). If the proposed value decreases the posterior density ($r < 1$), then we keep the proposed value with probability r. Second, multiplication by the ratio of the proposal densities evaluated at θ_j^{t-1} and θ_j^* satisfies the "detailed balance condition" such that the algorithm is guaranteed to converge to the posterior distribution. Third, there are many methods for choosing the proposal. For example, Q may be chosen to be independent of the previous θ_j value, and a common choice for Q is the prior for θ_j. The random walk proposal is very common, and it is defines $\theta_j^* = \theta_j^{t-1} + \varepsilon^t$. A common choice for ε^t is $\varepsilon^t \sim N(0, \nu)$, yielding a symmetric proposal for Q, and thus the acceptance ratio reduces to the ratio of the posterior densities evaluated at θ_j^* and θ_j^{t-1}. Fourth, the posterior for θ and the proposal for θ should be expressed on the same scale. For example, if it is more convenient to specify a prior for θ, but propose on a different scale, e.g., for $f(\theta) = \log(\theta)$, then a transformation of variables must be applied to either the prior (and thus the posterior) or to the proposal so that both are defined for the same quantity. See "Further Readings" for more details about different M–H algorithms.

in the "Further Reading" section. High-level software packages are available for implementing most Bayesian models that require MCMC methods. Popular software includes OpenBUGS, its predecessor WinBUGS, and JAGS. One must still be familiar with evaluating MCMC output (i.e., convergence, burn-in, mixing, and so on), and these software packages have built-in tools to help facilitate the process.

OTHER CONSIDERATIONS

This entry has only skimmed the surface of Bayesian statistics. In practice, ecologists often deal with challenging problems that merge ecological models with one or more datasets which may vary at different scales or represent different levels of completeness, intensity, accuracy, or resolution. Implementation of such models may require that we consider "tricks" for improving MCMC behavior such as, but not limited to, reparameterization of a model or model components, which may also facilitate the interpretation of model parameters, parameter expansion in hierarchical parameter models (applied in certain situations when small hierarchical variance terms cause poor MCMC behavior), or reversible jump MCMC to accommodate models characterized by a variable dimension parameter space. Finally, we see model diagnostics as an underdeveloped field in Bayesian statistics, and classical diagnostic tools are often used. Regardless of whether one chooses a Bayesian versus classical approach, one should always consider conducting some sort of diagnostics, a common one being evaluating model goodness-of-fit by comparing, for example, posterior results for replicated data to observed data, but there are many other avenues to explore.

CONCLUSIONS

The Bayesian approach represents a different way of thinking compared to the classical or frequentist approach. For example, unknown quantities such as model parameters are treated as random variables in a similar fashion to the treatment of data. The flexibility of the Bayesian framework allows the scientific problem to drive the modeling and data analysis, whereas classical methods may often force the analysis to fit within a relatively restrictive framework.

A major challenge that ecologists face is how to integrate diverse datasets representing complex ecological phenomena with process models designed to learn about these complexities. Bayesian statistics enables such integration, which has otherwise been difficult to achieve via traditional approaches. An important aspect of the Bayesian framework is the ease with which multiple data sources can be integrated. And it also allows us to explicitly deal with common issues experienced by ecologists, such as missing data, unbalanced sampling designs, temporally or spatially misaligned data, temporal or spatial correlation, and multiple sources of uncertainty. Thus, there is tremendous potential for Bayesian methods in ecology, and the probabilistic, hierarchical modeling framework presents an excellent opportunity to integrate ecological theory and models with empirical information.

SEE ALSO THE FOLLOWING ARTICLES

Computational Ecology / Frequentist Statistics / Information Criteria in Ecology / Markov Chains / Model Fitting / Statistics in Ecology

FURTHER READING

Berger, J. O. 1985. *Statistical decision theory and Bayesian analysis,* 2nd ed. New York: Springer-Verlag.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian theory*. Chichester: John Wiley & Sons.

Casella, G., and R. L. Berger. 2002. *Statistical inference.* Pacific Grove, CA: Duxbury.

Clark, J. S., and A. E. Gelfand. 2006. *Hierarchical modelling for the environmental sciences: statistical methods and applications.* Oxford: Oxford University Press.

Gamerman, D., and H. F. Lopes. 2006. *Markov chain Monte Carlo.* Boca Raton: Chapman and Hall/CRC Press.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian data analysis.* Boca Raton: Chapman and Hall/CRC Press.

Gelman, A., and J. Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1995. *Markov chain Monte Carlo in practice.* Boca Raton, FL: Chapman and Hall/CRC.

Ogle, K., and J. J. Barber. 2008. Bayesian data-model integration in plant physiological and ecosystem ecology. *Progress in Botany* 69: 281–311.

Robert, C. P. 2001. *The Bayesian choice,* 2nd ed. New York: Springer-Verlag.

BEHAVIORAL ECOLOGY

B. D. ROITBERG

Simon Fraser University, Burnaby, British Columbia, Canada

R. G. LALONDE

University of British Columbia, Okanagan, Kelowna, Canada

Behavioral ecologists study a very wide variety of behaviors, as can be seen in almost every issue of the journals *Behavioral Ecology* and *Behavioral Ecology and Sociobiology*, but these behaviors can easily be categorized in four main behavioral classes, each of which has a high association with fitness variation. These are energy acquisition (feeding), aggression (fighting), reproduction