ISSN 1085-7117, Volume 15, Number 3

Journal of Agricultural, Biological, and Environmental Statistics

September 2010
Volume 15
Number 3

An Applied Journal Published by the
International Biometric Society
and the *American Statistical Association*

Springer

Hypothesis Tests on Mixture Model Components with Applications in Ecology and Agriculture

Ling XU, Timothy HANSON, Edward J. BEDRICK, and Carla RESTREPO

Multiple comparisons are widely used to compare gross features of distributions across populations. However, often a scientific hypothesis is more easily couched in terms of more focused null and alternative statistical hypotheses. For example, among distributions exhibiting clusters of continuous measurements across strata, are there clusters of measurements similar in terms of location, spread, or weight? We propose testing such hypotheses using a sequence of nested finite mixture models. Reasonable, data-driven priors are suggested based on estimates of the sample spreads and midpoints. Formal hypothesis testing is carried out through the computation of Bayes factors. The method is illustrated on Holling's (Ecological Monographs 62:447–502, 1992) forest and prairie bird body mass data, and data on the time-to-abortion in dairy cows. Supplemental simulations are available online.

Key Words: Finite mixture model; Hierarchical mixture of experts; Multiple comparisons; Textural-discontinuity hypothesis.

1. INTRODUCTION

This article develops Bayesian hypothesis tests for comparing aspects of finite mixture models across populations. We motivate our proposed sequence of nested hypothesis tests with two examples from ecology and agriculture.

Body size is one of the most important attributes of living organisms (McMahon and Bonner 1983). The enormous variability and close relationship of body mass with a diverse array of physiological, morphological, and life-history attributes (e.g., Peters 1983;

Research supported by N.S.F. Grant NSF-HRD 0206200.

Ling Xu is Assistant Professor, James Madison University, Harrisonburg, VA, USA. Timothy Hanson (✉) is Associate Professor, Division of Biostatistics, University of Minnesota School of Public Health, A460 Mayo Building: MMC 303, 420 Delaware St. S.E., Minneapolis, MN 55455, USA (E-mail: *hanson@biostat.umn.edu*). Edward J. Bedrick is Professor, University of New Mexico, Albuquerque, NM, USA. Carla Restrepo is Associate Professor, University of Puerto Rico-Río Piedras, San Juan, Puerto Rico.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 15, Number 3, Pages 308–326
DOI: 10.1007/s13253-010-0020-z

Schmidt-Nielsen 1984) makes it a unique currency to address fundamental questions in ecology and evolution. Some of these questions target understanding patterns and processes pertaining to particular levels of biological organization whereas others target particular spatial and temporal scales (Halloway 2007; Hunt 2007). Irrespective of the question and approach, a key unresolved issue is the underlying form of the body size distribution. Traditionally, the body size distribution is assumed to be continuous and unimodal (Hutchinson and MacArthur 1959; Halloway 2007). Others suggest that it is discontinuous or multimodal (Wilson 1953; Holling 1992). These two views lead to contrasting hypotheses about the processes underlying the observed variability in body size and its relationship with other attributes. For example, continuous unimodal distributions may imply the existence of a single optimal body size (Stanley 1973; Brown, Marquet, and Taper 1993) whereas discontinuous or multimodal distributions may imply the existence of thresholds, forbidden body sizes, and multiple body size optima (Holling 1992; Allen, Forys, and Holling 1999). A second issue to consider is the availability of statistical methods that cannot only help reveal multimodal distributions in body size but also interesting differences among groups or relationships with other variables, and therefore provide explanations for the observed patterns. For example, Holling (1992) compares the distribution of body sizes of birds and mammals living in contrasting habitats and finds that they not only cluster around a limited set of body sizes, but that they also share many similarities. This suggests a critical role of landscapes in organizing animal assemblages and the possibility that these assemblages are resilient to change (textural-discontinuity hypothesis) (Holling 1992; Restrepo, Renjifo, and Marples 1997).

Figure 1 depicts histograms of the log body masses of boreal (northern) prairie and forest birds with estimated densities based on finite mixture models (see Section 5). For now, note that the distributions appear to be comprised of two or three moderately homogeneous, bell-shaped components.

The timing and causation of spontaneous abortion in dairy cows is of marked interest to herd owners and dairy managers. If the causation of the event is rooted in a pathogen introduced at a specific point in the fetal life-cycle, then the timing of abortion will tend to be similar across cows. This would necessarily invalidate the proportional hazards (PH) and accelerated failure time models (AFT), both of which have been suggested for relating the timing of abortion to herd characteristics and maternal risk factors such as parity and age (Thurmond et al. 2005; Hanson et al. 2003). The effect of covariates in the AFT model effectively warps time, i.e., stretches or contracts time, relative to baseline or other covariates. In contrast, the hierarchical mixture of experts (HME) model (Jordan and Jacobs 1994; Bishop and Svensén 2003) does not warp time, but rather models the probability that an event time arises from a Gaussian component with fixed mean and precision. HME models are simply finite mixture models where the probability of latent group, or component, membership is modeled as a function of risk factors.

Consider data on the time to spontaneous abortion among $n = 2302$ dairy cattle from six herds in the San Joaquin Valley of California (Karuppanan, Thurmond, and Gardner 1997). Figure 2 shows histograms and density estimates based on finite normal mixture models for each herd. The distributions of the time to abortion appear to be comprised of three moderately homogeneous components, but in different proportions across herds.

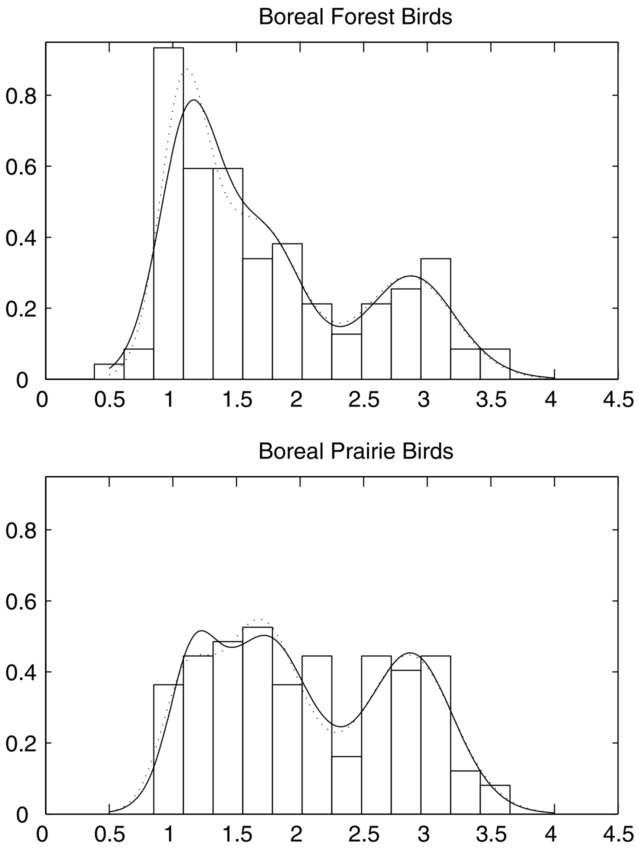

Figure 1. Histograms with M_3 (dotted) and M_2 (solid) density estimates for Holling's (1992) boreal bird data (log body mass in grams).

A visual assessment suggests that the stages during which abortion occurs are fixed across herds, but that the relative proportions falling into each stage varies across herds.

In this paper, a generalization of the one-way ANOVA model is developed for non-normal data, but where certain characteristics of the population densities are hypothesized to remain constant across groups. In the standard ANOVA, t groups are compared assuming the model

$$y_{i1}, \ldots, y_{in_i} \overset{\text{iid}}{\sim} N\big(\mu_i, \tau^{-1}\big),$$

where n_i is the sample size from the ith population $(i = 1, \ldots, t)$. The hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ holds if and only if all observations $\{y_{ij}\}$ arise from the same distribution. The standard alternative hypothesis implies that population densities differ only by location.

Inferences in the one-way ANOVA model depend crucially on the normality assumption. In many settings, the data distributions have multiple modes and skewness, and no obvious transformation to approximate normality can be found. A natural generalization of the simple ANOVA model that accommodates these features is to model densities as

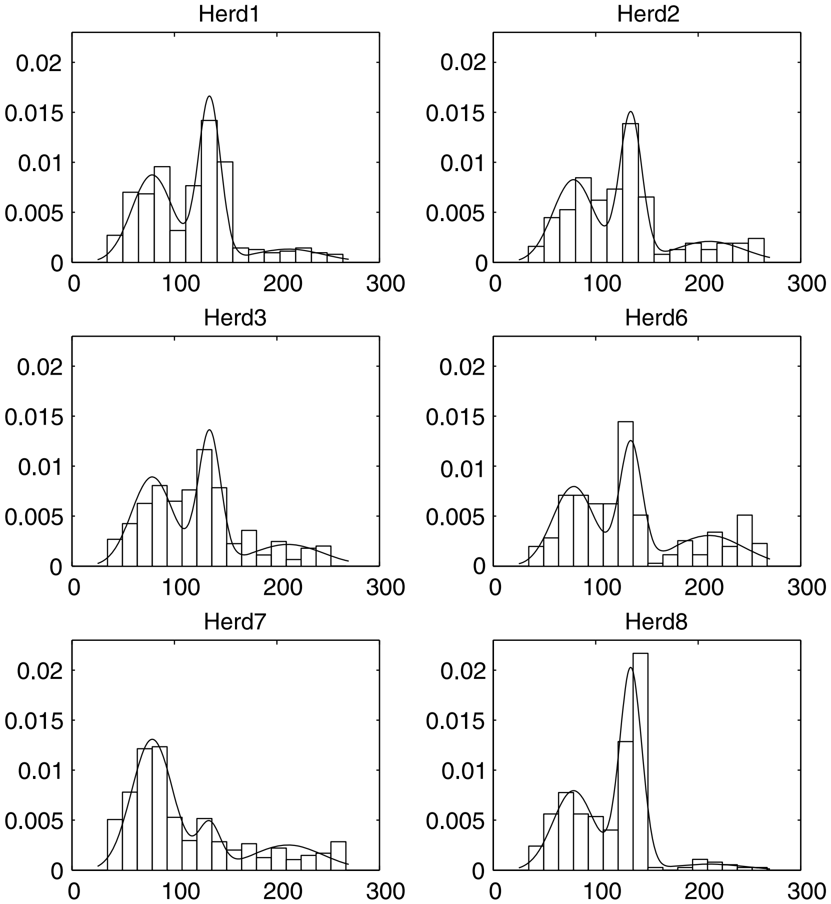

Figure 2. Histograms and M_3 density estimates for time-to-abortion data (days).

finite *mixtures* of homogeneous "clusters" or components:

$$y_{i1}, \ldots, y_{in_i} \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \omega_{ik} N\big(\mu_{ik}, \tau_{ik}^{-1}\big).$$

Relaxing the normality assumption by modeling each population density as a finite mixture of Gaussian components not only allows greater flexibility in testing whether the data arise from the same population, but further facilitates finding commonality less extreme than the hypothesis that all populations have the same density. For example, the hypothesis that there are K identical "centers of attraction" in each population, but that the distributions about these centers are possibly more dispersed or occur with different relative frequencies across populations can be formulated in the mixture model context by $H_0 : \mu_{1k} = \mu_{2k} = \cdots = \mu_{tk}$, for $k = 1, \ldots, K$. These types of hypotheses are readily tested in the Bayesian framework through the use of Bayes factors, which indicate how well one model supports the observed data relative to another. Bayes factors, which require proper priors for each competing model, are discussed in detail in Section 4.

Throughout the article, the number of components K is assumed fixed and known. Although this will rarely be the case, many methods exist for estimating K, for example, the Bayesian information criterion (BIC) (Roeder and Wasserman 1997) and the weighted gap statistic (Yan and Ye 2007). For the boreal bird data in Section 5, we find the posterior mode of K given default prior specifications using the reversible jump approach of Richardson and Green (1997), the Dirichlet process mixture approach (Escobar and West 1995), and the BIC approximation of Roeder and Wasserman (1997). These approaches often agree with each other and with what one might decide based on simply looking at histograms.

The remainder of the article is organized as follows. Section 2 defines the population model and discusses a data-driven proper prior specification. In Section 3, four nested models for testing hypotheses of interest are presented. Full conditional distributions for blocks of means, precisions and weights are given to facilitate the implementation of a Gibbs sampler. Section 4 outlines Chib's (1995) algorithm to compute Bayes factors for comparing the four models. Section 5 analyzes the boreal bird and time-to-abortion data. Comparisons with alternative model selection methods are also provided. Section 6 presents conclusions and discussion.

2. THE POPULATION MODEL AND PRIOR

Methods for computing Bayes factors using priors that are proper, but as vague as possible, have been proposed, most notably fractional Bayes factors (O'Hagan 1995) and intrinsic Bayes factors (Berger and Pericchi 1996). Both of these approaches use training sets of data to construct proper, but weakly informative priors, and are not considered here. Instead, we follow the development in Richardson and Green (1997) and consider a proper, but vague prior specification that takes into account the sample range and the midpoint of the data.

Suppose observations from each population are independent and have a distribution that is a finite mixture of K normal components. Our basic mixture model is:

$$y_{i1}, y_{i2}, \ldots, y_{in_i} \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \omega_{ik} \Phi\big(\cdot|\mu_{ik}, \tau_{ik}^{-1}\big), \tag{1}$$

for populations $i = 1, \ldots, t$, where K is known and $\{\omega_{ik}\}$ are unknown non-negative weights such that $\sum_{k=1}^{K} \omega_{ik} = 1$. Here, $\Phi(\cdot|\mu, \tau^{-1})$ is the cdf of a normal distribution with mean μ and precision τ. Let $\phi(\cdot|\mu, \tau^{-1})$ be the corresponding normal pdf.

Typically, there is little to no prior information on the component weights. A Dirichlet$(\delta \mathbf{1}_K)$ distribution, with δ small, provides a flexible, conjugate choice. Assuming $\delta = 1$, we have

$$(\omega_{i1}, \ldots, \omega_{iK})' \sim \text{Dirichlet}(\mathbf{1}_K).$$

The prior on the component means is given by

$$\mu_{i1}, \ldots, \mu_{iK} | \xi_i, \kappa_i \sim N\big(\xi_i, \kappa_i^{-1}\big), \tag{2}$$

subject to the constraint $\mu_{i1} < \mu_{i2} < \cdots < \mu_{iK}$ for identifiability and enhanced interpretability. Choosing ξ_i to be the midpoint between the sample extremes, $\xi_i = (y_{i(1)} + y_{i(n_i)})/2$ and $\kappa_i = 1/R_i^2$, where $R_i = y_{i(n_i)} - y_{i(1)}$ is the sample range, centers the prior in the middle of the data and keeps the prior flat over an interval of variation of the data. Increasing κ_i serves to shrink means towards ξ_i. Note that the prior density for $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iK})$ is $p(\boldsymbol{\mu}_i) = K! I\{\mu_{i1} < \mu_{i2} < \cdots < \mu_{iK}\} \prod_{k=1}^{K} \phi(\mu_{ik} | \xi_i, \kappa_i^{-1})$.

Chib (1995) does not enforce the order constraint on the means in a simpler normal mixture model. Neal (1999) shows that this impacts the accuracy of Bayes factor computations in modest Gibbs sampling runs. For the models we consider, the constraint is easy to incorporate into the initial and reduced Gibbs samplers used in Chib's algorithm. Furthermore, ordered means avoids the "label switching" problem and ensures interpretable models. Recently, Lee et al. (2008) suggest combining Chib's algorithm with the pivotal reordering method to compute Bayes factors in finite mixture models. We prefer imposing the constraint a priori rather than using this post-hoc approach.

A sensible default prior specification on the component precisions is useful in the absence of real prior information. To this end, Richardson and Green (1997) develop a hierarchical data-driven prior for $\{\tau_{ik}\}$ that reflects the prior belief that the component precisions should be similar, but their absolute size should be left arbitrary. This prior gives results that are pleasing to the eye for a variety of data sets, and extends to a model in which the number of components K is an unknown parameter. Their prior is given by

$$\tau_{i1}, \ldots, \tau_{iK} | \beta_i \overset{\text{iid}}{\sim} \Gamma(\alpha, \beta_i), \qquad \beta_i \sim \Gamma(g, h_i), \tag{3}$$

where α, g, and h_i are specified. This hierarchical prior induces the simple prior with pdf

$$\tau_{ij} \sim p_i(\tau) = \frac{\Gamma(\alpha + g) h_i^g \tau^{\alpha - 1}}{\Gamma(\alpha)\Gamma(g)(\tau + h_i)^{\alpha + g}}. \tag{4}$$

Richardson and Green (1997) recommend $\alpha = 2$, $g = 0.2$, and $h_i = 10/R_i^2$, which gives

$$\tau_{ij} \sim p_i(\tau) = \frac{0.24 h_i^{0.2} \tau}{(\tau + h_i)^{2.2}}. \tag{5}$$

A simple non-hierarchical $\Gamma(\alpha, \beta)$ prior on the precisions facilitates the computationally intensive approach to estimating Bayes factors in Section 4. Diebolt and Robert (1994), Carlin and Chib (1995), Chib (1995), Bishop and Svensén (2003) also consider a gamma prior. A natural approach would be to specify α and β by matching the first moments in (5) with those of $\Gamma(\alpha, \beta)$. Unfortunately, the moments of (5) do not exist. In fact, this prior is heavy-tailed, placing significant mass on very small values of $\sigma_{ij} = \tau_{ij}^{-1/2}$:

$$P(\sigma_{ij} \le 0.00001981 R_i) \approx 0.025, \qquad P(\sigma_{ij} \ge 0.387 R_i) \approx 0.025$$

Alternatively, a useful simplification of the hierarchical prior (3) is obtained by replacing β_i by its expectation under $\beta_i \sim \Gamma(g, h_i)$, $\beta_i = g/h_i = R_i^2/50$, giving the prior

$$\tau_{i1}, \ldots, \tau_{iK} \overset{\text{iid}}{\sim} \Gamma\left(2, \frac{R_i^2}{50}\right). \tag{6}$$

Here $E(\tau_{ij}) = \alpha/\beta_i = 100/R_i^2$ and so a "typical" value of σ_{ij} is $R_i/10$. The mode of τ_{ij} is $(\alpha - 1)/\beta_i = 50/R_i^2$ implying that the mode of σ_{ij} is about $0.14R_i$. By comparison, the mode of the induced prior (5) is $h_i/1.20 = 8.33/R_i^2$ implying that the mode of σ_{ij} is approximately $0.35R_i$. Under (6),

$$P(\sigma_{ij} \le 0.060R_i) \approx 0.025, \qquad P(\sigma_{ij} \ge 0.287R_i) \approx 0.025,$$

so this prior does not allow as extreme values of σ_{ij} as prior (3) and places approximately 95% probability on values of σ_{ij} within about $R_i/4$.

To allow more extreme component variation we consider an alternative simple prior with greater spread than (6). The prior

$$\tau_{i1}, \ldots, \tau_{iK} \overset{\text{iid}}{\sim} \Gamma\left(\frac{1}{2}, \frac{R_i^2}{3000}\right) \tag{7}$$

yields $E(\tau_{ij}) = 1500/R_i^2$ and a typical value of σ_{ij} of about $0.026R_i$. Thus, a typical Gaussian component "length" is about $0.1R_i$. This prior places more mass on both smaller and larger σ_{ij} than the gamma prior described above:

$$P(\sigma_{ij} \le 0.0115R_i) \approx 0.025, \qquad P(\sigma_{ij} \ge 0.824R_i) \approx 0.025.$$

Note that $E(\sigma_{ij}) = \infty$ for this prior.

Priors (5), (6) and (7) can be directly compared because all three are scale families in R_i^2. Figure 3 shows the three priors when $R_i = 1$. Priors (5) and (6) prohibit very small precisions and thus very large standard deviations. Prior (7) allows for much larger values of σ_{ij}. The induced prior (5) allows for absurdly small values of σ_{ij}, or density spikes at zero.

Richardson and Green (1997) also consider simple gamma priors of this type but recommend the full hierarchical prior (3) instead. They also discuss sensitivity of the posterior means of the τ_{ij} as $\sqrt{\beta_i/\alpha}$ is varied, and note that this sensitivity increases with the number of components K. This, in part, may simply be due to the fact that if a small number of components adequately describes a data set, additional components may only be weakly identified and thus posterior inferences will be sensitive to the prior specification. To avoid this we pick K to be as small as is plausible for a given data set.

When the number of components K is fixed at a small number, say $K = 2, 3$ or 4, the ordered component locations, weights, and precisions should be well-identified by most data and we expect the two simple priors to give similar results. We find this to be the case in the simulations and data analyses of Section 4. When K is large, or random as in Richardson and Green (1997), we expect different priors to give potentially very different results.

In summary, the most general prior specification for population i is

$$\begin{aligned}
\mu_{i1}, \ldots, \mu_{iK} &\overset{\text{iid}}{\sim} N(\xi_i, \kappa_i^{-1}) \quad \text{subject to} \quad \mu_{i1} < \mu_{i2} < \cdots < \mu_{iK}, \\
\tau_{i1}, \ldots, \tau_{iK} &\overset{\text{iid}}{\sim} \Gamma(\alpha, \beta_i), \\
(\omega_{i1}, \ldots, \omega_{iK})' &\sim \text{Dirichlet}(\delta, \delta, \ldots, \delta).
\end{aligned}$$

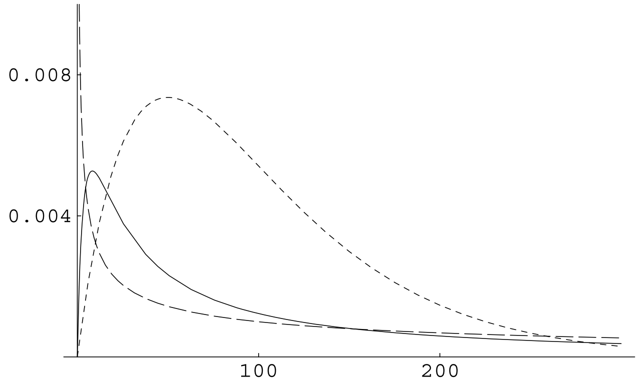

Figure 3. Precision prior densities: (solid) induced prior of Richardson and Green; (short dashed) $\Gamma(2, \frac{1}{50})$; (long dashed) $\Gamma(\frac{1}{2}, \frac{1}{3000})$.

Prior parameter values are fixed at $\delta = 1$, $\kappa_i = R_i^{-2}$, and $\xi_i = y_{i(1)} + R_i/2$ where R_i is the ith sample range. Two precision priors are considered: $(\alpha, \beta_i) = (2, R_i^2/50)$ and $(\alpha, \beta_i) = (1/2, R_i^2/3000)$.

To facilitate the implementation of a Gibbs sampler (Gelfand and Smith 1990) to fit the mixture model for a given population i, Diebolt and Robert (1994) augment the model parameters with latent allocation variables z_{ij}, one for each y_{ij}, that indicate which component $k \in \{1, \dots, K\}$ generated the observation y_{ij}. That is,

$$z_{ij} = k \quad \Leftrightarrow \quad y_{ij} \sim \Phi\big(\cdot|\mu_{ik}, \tau_{ik}^{-1}\big), \qquad j = 1, \dots, n_i.$$

Note then $P(z_{ij} = k) = \omega_{ik}$ for $k = 1, \dots, K$.

3. A NESTED SEQUENCE OF MODELS

3.1. MODEL M_1: THE FULL MODEL

The full model M_1 assumes that each of the t populations has a distinct normal mixture of K components. For the ith population,

$$y_{i1}, y_{i2}, \dots, y_{in_i} \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \omega_{ik} \Phi\big(\cdot|\mu_{ik}, \tau_{ik}^{-1}\big), \qquad i = 1, \dots, t.$$

The distributions of the allocation variables $\{z_{ij}\}$ given $(\omega_{i1}, \dots, \omega_{iK})'$ and priors on the remaining variables were specified in Section 2. Let n_{ik} be the number of observations allocated to the kth component, $n_{ik} = \sum_{j=1}^{n_i} I\{z_{ij} = k\}$. The full conditional distributions for the component precisions and for the component means are recognized as

$$\tau_{ik}|\text{ else} \sim \Gamma\left(\frac{n_{ik}}{2} + \alpha, \beta_i + \frac{1}{2} \sum_{j:z_{ij}=k} (y_{ij} - \mu_{ik})^2\right) \tag{8}$$

and

$$\mu_{ik}|\text{ else} \sim N\left(\frac{\tau_{ik}\sum_{j:z_{ij}=k} y_{ij} + \kappa_i \xi_i}{\tau_{ik}n_{ik} + \kappa_i}, \frac{1}{\tau_{ik}n_{ik} + \kappa_i}\right) I_{(\mu_{i,k-1},\mu_{i,k+1})}, \tag{9}$$

where $N(\mu, \sigma^2)I_{(a,b)}$ denotes a $N(\mu, \sigma^2)$ distribution truncated to (a, b). Here $\mu_{i0} = -\infty$ and $\mu_{i,K+1} = \infty$. The full conditional distribution on the weights is Dirichlet:

$$(\omega_{i1}, \ldots, \omega_{iK})'|\text{ else} \sim \text{Dirichlet}(\delta + n_{i1}, \ldots, \delta + n_{iK}). \tag{10}$$

Finally, the conditional distributions of the allocation variables are independent with

$$P(z_{ij} = k|\text{ else}) \propto \omega_{ik}\tau_{ik}^{\frac{1}{2}} \exp\left[-\frac{1}{2}\tau_{ik}(y_{ij} - \mu_{ik})^2\right].$$

3.2. Model M_2: Identical Means

Model M_2 restricts M_1 so that the K component means are identical across populations:

$$\mu_{1k} = \mu_{2k} = \cdots = \mu_{tk} \equiv \mu_k, \qquad k = 1, \ldots, K.$$

Model M_2 is nested within M_1 so the prior on the component means $\{\mu_k\}$ is necessarily different from the prior discussed in Section 2, but all other priors are as before. We specify

$$\mu_1, \ldots, \mu_K \overset{\text{iid}}{\sim} N\left(\xi_s, \kappa_s^{-1}\right),$$

where $\kappa_s = \min\{\kappa_1, \kappa_2, \ldots, \kappa_t\}$, $\kappa_i = R_i^{-2}$, and $\xi_s = \sum_{i=1}^t \xi_i/t$, the average midpoint of the t sample ranges. Thus, the prior is centered at the midpoint of the combined samples and covers the range of each sample.

The component means are identical across populations in this model, thus the full conditional densities for the means depend on all t samples. The full conditional distributions for these parameters are recognized as

$$\mu_k|\text{ else} \sim N\left(\frac{\sum_{i=1}^t(\tau_{ik}\sum_{j:z_{ij}=k} y_{ij}) + \kappa_s\xi_s}{\kappa_s + \sum_{i=1}^t \tau_{ik}n_{ik}}, \frac{1}{\kappa_s + \sum_{i=1}^t \tau_{ik}n_{ik}}\right) I_{(\mu_{k-1},\mu_{k+1})}, \tag{11}$$

where $\mu_0 = -\infty$ and $\mu_{K+1} = \infty$. The full conditional distributions for $\{\tau_{ik}\}$, $\{\omega_{ik}\}$ and $\{z_{ij}\}$ are the same as in model M_1, but with μ_k replacing μ_{ik}, where $i = 1, \ldots, t$, $k = 1, \ldots, K$.

3.3. Model M_3: Identical Means and Variances

Model M_3 assumes that the component means and precisions are identical across populations:

$$\mu_{1k} = \mu_{2k} = \cdots = \mu_{tk} \equiv \mu_k \qquad \text{and} \qquad \tau_{1k} = \tau_{2k} = \cdots = \tau_{tk} \equiv \tau_k, \qquad k = 1, \ldots, K.$$

This model, nested within both M_2 and M_1, assumes that populations are comprised of the same K components, but in different proportions. The prior on the component precisions $\{\tau_k\}$ is necessarily different from the precision priors discussed in Section 2. Here we use

$$\tau_1, \ldots, \tau_K \overset{\text{iid}}{\sim} \Gamma(\alpha, \beta_s),$$

where $\beta_s = \sum_{i=1}^{t} \beta_i / t$, the mean of the scale parameters under models M_1 and M_2. Recall $\beta_i = R_i^2/50$ and $\beta_i = R_i^2/3000$, thus β_s depends on the range of each sample. Under model M_3, the sample ranges should not vary significantly; therefore, β_s should be close to all $\{\beta_i\}$ and work well here.

The full conditional distributions for the component precisions are recognized as

$$\tau_k | \text{ else} \sim \Gamma\left(\alpha + \frac{1}{2}\sum_{i=1}^{t} n_{ik}, \beta_s + \frac{1}{2}\sum_{i=1}^{t}\sum_{j: z_{ij}=k} (y_{ij} - \mu_k)^2\right), \tag{12}$$

which depends on the entire data set. The full conditional distributions for $\{\mu_k\}$, $\{\omega_{ik}\}$ and $\{z_{ij}\}$ are the same as in model M_2, but with τ_k replacing τ_{ik}, where $i = 1, \ldots, t$, $k = 1, \ldots, K$.

3.4. MODEL M_4: IDENTICAL DISTRIBUTIONS

The simplest model, M_4, assumes data from all populations arise from the same distribution $y_{ij} \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \omega_k \Phi(\cdot|\mu_k, \tau_k^{-1})$. The means, precisions, and weights of the K Gaussian components are the same across populations. Compared to Model M_3, this model has the further restriction:

$$\omega_{1k} = \omega_{2k} = \cdots = \omega_{tk} \equiv \omega_k, \qquad k = 1, \ldots, K,$$

so M_4 is nested within M_1, M_2 and M_3. The prior on weights must be adjusted accordingly. Here we choose

$$\omega_1, \ldots, \omega_K \;\sim\; \text{Dirichlet}(\mathbf{1}_K).$$

The full conditional distribution on the weights is

$$(\omega_1, \omega_2, \ldots, \omega_k)' | \text{ else} \sim \text{Dirichlet}\left(1 + \sum_{i=1}^{t} n_{i1}, 1 + \sum_{i=1}^{t} n_{i2}, \ldots, 1 + \sum_{i=1}^{t} n_{iK}\right). \tag{13}$$

The full conditional distributions for $\{\mu_k\}$, $\{\tau_k\}$ and $\{z_{ij}\}$ are the same as in model M_3, but replacing ω_{ik} with ω_k, where $i = 1, \ldots, t$, $k = 1, \ldots, K$.

4. HYPOTHESIS TESTS ON MIXTURE COMPONENTS

Kadane and Lazar (2004) review a variety of criteria for model selection. We will consider Bayes factors. Han and Carlin (2001) review several methods that can be used to calculate Bayes factors, including methods due to Chib (1995), Carlin and Chib (1995), and Green (1995). They recommend Chib's approach. Of the methods we tried, including reversible jump, we found Chib's approach to have the best stability, reasonable computation cost, and moderate ease of implementation. Song and Lee (2002) give an alternative method for finite mixture models based on path sampling that could possibly be generalized to the models considered here. See also Steele, Raftery, and Emond (2006) for an importance sampling approach.

For the comparison of two models M_i and M_j on data $\mathbf{y}$, the Bayes factor is

$$B_{ij} = \frac{p_i(\mathbf{y})}{p_j(\mathbf{y})}$$

where

$$p_m(\mathbf{y}) = \int p_m(\mathbf{y}|\boldsymbol{\theta}_m)\pi_m(\boldsymbol{\theta}_m)\, d\boldsymbol{\theta}_m$$

is the marginal likelihood of the data under model M_m. Here $p_m(\mathbf{y}|\boldsymbol{\theta}_m)$ is the data model likelihood depending on parameters $\boldsymbol{\theta}_m$ and $\pi_m(\boldsymbol{\theta}_m)$ is the prior. The Bayes factor B_{ij} is the weight of evidence in favor of model M_i relative to M_j.

Unlike p-values, Bayes factors can support a model (or null hypothesis) as well as provide evidence against a model or null hypothesis. We adopt Kass and Raftery's (1995) guidelines for interpreting B_{ij}. They suggest that a Bayes factor of 1 to 3 is "not worth more than a bare mention," 3 to 20 is "positive," 20 to 150 is "strong," and greater than 150 is "very strong." Jeffreys (1961) provides a similar scale.

We consider a nested series of hypothesis tests that start with the most general model, M_1, and lead to models M_2, M_3, and M_4, respectively:

1. (Model M_1) H_1 : no constraint on model parameters.
2. (Model M_2) $H_2 : \mu_{1k} = \mu_{2k} = \cdots = \mu_{tk}$ for $k = 1, \ldots, K$.
3. (Model M_3) $H_3 : H_2$ and $\tau_{1k} = \tau_{2k} = \cdots = \tau_{tk}$ for $k = 1, \ldots, K$.
4. (Model M_4) $H_4 : H_3$ and $\omega_{1k} = \omega_{2k} = \cdots = \omega_{tk}$ for $k = 1, \ldots, K$.

In subsequent analyses, we will typically select a best model, corresponding to the model with the maximum estimated marginal probability $p_m(\mathbf{y})$. This model has a Bayes factor greater than unity when compared to every other model. However, the data may not have strong evidence in favor of this model according to Kass and Raftery's guideline.

Chib's (1995) method is a simple approach for computing the marginal probability $p_m(\mathbf{y})$ from Gibbs sampler output. For ease of exposition, we drop the model subscript m. For any $\boldsymbol{\theta}^*$, Bayes' rule on the logarithmic scale gives

$$\log p(\mathbf{y}) = \log p(\mathbf{y}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \pi(\boldsymbol{\theta}^*|\mathbf{y}). \tag{14}$$

Chib suggests choosing $\boldsymbol{\theta}^*$ to be a point with high posterior density, such as an estimate of the posterior mean or mode, to maximize computation accuracy. The first two terms of (14) are easy to compute, but the third term requires effort. Chib suggests decomposing the parameter vector into, say, j blocks of similar parameters $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_j^*)$ and running a series of j Gibbs samplers as briefly outlined below.

Each of models M_1, M_2, M_3, and M_4 has blocks of location, scale, and weight parameters, say $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\omega})$. For any of the models let $(\boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \boldsymbol{\omega}^*)$ be a point of relatively high posterior mass, for example, the posterior mean. For each model, our implementation of Chib's (1995) algorithm decomposes the ordinate as

$$\pi(\boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \boldsymbol{\omega}^*|\mathbf{y}) = p(\boldsymbol{\mu}^*|\mathbf{y})p(\boldsymbol{\tau}^*|\boldsymbol{\mu}^*, \mathbf{y})p(\boldsymbol{\omega}^*|\boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \mathbf{y}).$$

The ordinate $p(\boldsymbol{\mu}^*|\mathbf{y})$ is obtained via the 'Rao–Blackwellized' estimator suggested by Gelfand and Smith (1990). This is simply the product of Gaussian densities (9) evaluated at $\boldsymbol{\mu}^*$ for M_1, or (11) for M_2, M_3, or M_4, averaged over MCMC iterates of an initial run of the Gibbs sampler. The densities are multiplied by either $(K!)^t$ or $K!$, respectively stemming from the order constraint. The ordinate $p(\boldsymbol{\tau}^*|\boldsymbol{\mu}^*, \mathbf{y})$ is obtained from running a second, "reduced" Gibbs sampler conditioning on the fixed value $\boldsymbol{\mu} = \boldsymbol{\mu}^*$. The MCMC iterates from this reduced run are averaged over the product of gamma densities (8) evaluated at $\boldsymbol{\tau}^*$ for M_1 and M_2, or (12) for M_3 and M_4. Finally, $p(\boldsymbol{\omega}^*|\boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \mathbf{y})$ is obtained from a third run of a further reduced Gibbs sampler conditioning on both $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ and $\boldsymbol{\tau} = \boldsymbol{\tau}^*$. These iterates are averaged over the product of Dirichlet densities (10) evaluated at $\boldsymbol{\omega}^*$ for M_1, M_2, or M_3, or (13) for M_4. A complete series of nested hypothesis tests requires running $4 \times 3 = 12$ Gibbs samplers.

In addition to the Bayesian approach developed here, we consider likelihood ratio testing, and model selection based on Akaike's (1973) information criterion (AIC) and the Schwartz (1978) Bayesian information criterion (BIC). The BIC typically penalizes dimensionality more than AIC. We find in the data analyses of Section 5 that the BIC chooses the same model as Bayes factors under the prior of Section 2. More generally, the BIC is an asymptotic approximation to the logarithm of the Bayes factor, so the BIC provides a reasonable approximation to the relative evidence for two competing models (Kass and Raftery 1995).

5. EXAMPLES

5.1. BODY MASS OF BOREAL BIRDS

Two basic approaches are used to examine the possibility that body sizes do not follow a continuous unimodal distribution. Holling (1992), Restrepo, Renjifo, and Marples (1997), Marples (1998) and Stow, Allen, and Garmestani (2007) focus on quantifying the number and location of discontinuities or gaps in the distribution of body sizes. Alternatively, Havlicek and Carpenter (2001) focus on quantifying the number, location, and size of modes in body size distributions. The former approach highlights the possibility that individuals or species of certain body sizes are not favored under a particular set of conditions, therefore creating gaps in the distribution of body sizes. We consider the second approach, emphasizing "centers of attraction," that is, optimal body sizes characterized by cluster means $\mu_{i1}, \ldots, \mu_{iK}$ in assemblage i. Moreover, it is not our intent to either confirm or reject Holling's textural-discontinuity hypothesis or other hypotheses that may explain discontinuous distributions in body size (Holling 1992; Allen et al. 2006), but rather to illustrate that fitting finite mixture models with component restrictions can be a useful tool for shedding more light on the competing hypotheses.

We analyzed data on $n_1 = 106$ boreal prairie birds found east of the Alberta shortgrass prairie and $n_2 = 101$ boreal forest birds found east of the Manitoba–Ontario border in pure or mixed conifer stands (Holling 1992; Appendices 1 and 3). We examined several approaches to choosing K, including reversible jump MCMC (Richardson and Green 1997), a Dirichlet process mixture model (Escobar and West 1995), and BIC (Roeder and

Table 1. Posterior probability of numbers of components K and modes m.

		Posterior probability							
		Forest birds				Prairie birds			
Approach		1	2	3	≥4	1	2	3	≥4
Reversible jump	K	0.00	0.09	0.71	0.20	0.00	0.75	0.21	0.04
	m	0.00	0.11	0.80	0.09	0.01	0.94	0.05	0.00
DP mixture	K	0.00	0.95	0.05	0.00	0.02	0.83	0.15	0.00
	m	0.00	0.79	0.21	0.00	0.08	0.71	0.21	0.00
BIC	K	0.00	0.07	0.93	0.00	0.01	0.95	0.04	0.00
	m	0.00	0.15	0.85	0.00	0.02	0.97	0.01	0.00

Table 2. Likelihood-based Summaries for Birds. L is log-likelihood evaluated at MLE; d is dimensionality of model.

	$K = 2$				$K = 3$			
Model	d	$-2L$	AIC	BIC	d	$-2L$	AIC	BIC
M_1	10	402.5	422.5	455.8	16	366.0	398.0	451.4
M_2	8	407.8	423.7	450.4	13	368.5	394.5	437.8
M_3	6	409.5	421.5	441.5	10	372.9	392.9	426.3
M_4	5	416.1	426.1	442.7	8	384.4	400.4	427.1

Wasserman 1997). Each method was calibrated to place most prior mass on one mode (Xu 2005). Table 1 shows the posterior distribution on K for the two mixtures. The number of modes m is also included. Regardless of the method, we have rather strong evidence against simple, homogeneous populations ($K = 1$). Most posterior mass is on $K = 2$ or $K = 3$ components, with enough variability across approaches and ecological strata that we chose $K = 3$ to be "as small as plausible." The posterior summaries also suggest that the mixture distributions are not unimodal, an indication that the components are fairly well-separated.

Assuming $K = 3$ components, prior (6) gives log marginal densities of -221.97, -219.75, -220.26, and -224.51 for Models $4, 3, 2, 1$. The model rankings are M_3, M_2, M_4 and M_1, with corresponding Bayes factors $B_{32} = 1.67$, $B_{24} = 5.53$, and $B_{41} = 12.68$. Kass and Raftery's guideline classify $B_{31} = 117.65$ and $B_{21} = 70.43$ as strong evidence against model M_1. Furthermore, we have strong evidence that either the means are identical, or both the means and precisions are equal, but not the weights. We note that the same model rankings were achieved using $K = 2$, but the rankings were more decisive using $K = 3$. Figure 1 shows density estimates from models M_2 and M_3.

For comparison, Table 2 gives maximum likelihood-based model selection criteria assuming $K = 2$ and $K = 3$. The EM algorithm was used to compute the maximum likelihood estimates. Although the AIC, BIC and our approach produce slightly different model rankings, all three methods select M_3 as best for $K = 2$ and $K = 3$. A step-up test at the

5% level based on the likelihood ratio statistic chooses the most general model M_1 whereas a step-down test chooses the simplest model M_4.

The strong support in our analysis for $K = 2$ or $K = 3$ components is in contrast with Holling's (1992, p. 458) results that suggest four or more body size clumps for both bird distributions. Moreover, a comparison of the two distributions based on the methods developed in this paper, indicate that there are similarities between them but also differences. A similar conclusion was reached by Holling (1992, p. 447) when he stated that "There is a striking similarity, but not identity, between the clump structure of prairie and boreal animals." Unlike Holling, however, we were able to indicate the nature of the similarities and differences. The former are given by the constant component means and precisions whereas the latter by the strata-specific weights.

Although our intent was not either to confirm or reject Holling's textural discontinuity hypothesis, our analyses seem to provide some support to it. Under this hypothesis, it is expected that the clump structure of body mass distributions should differ between animal assemblages inhabiting landscapes that vary in their structure. The constant component means and precisions are the aspects of the assemblage referred to in Section 1 as being immutable whereas the weights represent different proportions of landscape features across the prairie and forest strata, e.g., "isolated perches and trees and with the scattered shrubs typical of some parts of the prairie" and, in fact, perhaps invite a regression analysis in the form of HME.

Holling (1992) presented three additional hypotheses to explain the presence of clumps in body mass data, and to some extent our results could also support the limited-morph hypothesis (Holling 1992, p. 549). Under this hypothesis it is expected that animal sizes "cluster into a small number of clumps even if the spatial attributes of their habitats are continuously distributed." This clustering results from the fact that only a limited number of "locomotory habits" are possible for a given range of body sizes such that hovering (hummingbirds) and soaring (albatrosses) is only possible in small and large birds, respectively. The stratum-specific weights here could be interpreted in terms of the relative abundance of landscape features conducive to these types of locomotion.

5.2. ABORTION IN DAIRY COWS

As discussed in Section 1, the timing of spontaneous abortion in diary cows is of immense interest to the dairy industry. Proper assessment of abortion risk can lead to improved management strategies. Figure 2 shows distributions for the time-to-abortion in days for 2302 pregnancies in dairy cows from six central California herds, along with model M_3 fits. The herd sample sizes are $n_1 = 434$, $n_2 = 409$, $n_3 = 307$, $n_6 = 243$, $n_7 = 652$, and $n_8 = 257$. This is a subset of the time-to-abortion data analyzed by Hanson et al. (2003) and Thurmond et al. (2005) using three-component mixture models. This figure suggests $K = 3$ is appropriate.

Assuming $K = 3$, the marginal log-density ordinates ($\log\{p(\mathbf{y})\}$) for models M_4, M_3, M_2 and M_1 are −6675.33, −6580.74, −6584.27, and −6595.79, respectively. The corresponding ordinates from a subsequent run of the Gibbs samplers differed by at most 0.13. We have found, in general, that the marginal ordinate estimate is slightly more stable for

Table 3. LPML statistics for AFT and M_3 models across the six herds.

	Herd						
	1	2	3	6	7	8	Total
AFT	−2117	−2021	−1543	−1238	−3276	−1201	−11,396
M_3	−2069	−1978	−1527	−1202	−3234	−1142	−11,151

the last three models relative to the first. We also fit the models with $K = 4$ twice, with similar results and conclusions.

The model rankings based on the marginal density are M_3, M_2, M_1, and M_4. Using Bayes factors, model M_3, which has equal means and precisions but unequal weights across herds, is strongly preferred ($BF_{32} = 34.20$) to M_2 and very strongly preferred to the other two models. The BIC values (minus 23000) are 856, 851, 789, and 950 for M_1, M_2, M_3, and M_4, respectively. The corresponding AIC values (minus 23000) are 484, 596, 650, and 888. The AIC and likelihood ratio test choose the most complex model M_1, which might be expected given the large sample size. The BIC and marginal density estimates produce the same rankings and choose a more parsimonious model M_3 with 30 fewer parameters (18 versus 48).

Table 3 compares log pseudo-marginal likelihood (LPML) across the six herds, as well as the total LPML, for accelerated failure time (AFT) and M_3 models fit with vague priors in WinBUGS. The LPML, developed in Geisser and Eddy (1979), is a leave-one-out cross-validated measure of how well a model predicts the data and is relatively insensitive to prior specification. Larger values indicate better predictive ability. In terms of prediction, model M_3 clearly outperforms the AFT model (used by Hanson et al. 2003 for a superset of these data). This is not surprising because Figure 2 is almost a textbook example of M_3, whereas there is little evidence of accelerated time "warping" (i.e., stretching or compacting) across the herds.

Hanson et al. (2003) describe two windows of elevated risk of abortion, verified by field studies. An inhospitable uterine environment can lead to an initial phase of elevated abortion risk 30–60 days after conception. A second window of elevated risk occurs 80–140 days after conception, from possible exposure to pathogens from the dam followed by an incubation period. Pathogens thought to possibly lead to abortion include brucellosis, listeriosis, leptospirosis, and bovine viral diarrhea. Maternal risk factors include parity and age. The fitted densities in Figure 2 roughly confirm the two windows of elevated risk, but also indicates a third window of risk, occurring at roughly 200–250 days but with substantially lower hazard. There is evidence that herd characteristics (e.g., culling strategies, disease management) influence the relative proportions of cows experiencing the three different types of abortion hazard implied by the model. On a herd-to-herd basis, this could have profound management implications. For example, in a herd relatively free of pathogens the hazard will substantially drop after the first time-window.

6. CONCLUSIONS AND DISCUSSION

We presented four nested models that provide a meaningful framework for comparing finite mixture models across populations. The models use a practical data-driven prior, based on the work of Richardson and Green (1997), that assumes a reasonable spread within components relative to the range of the observed data. Computational methods for computing Bayes factors based on the work of Chib (1995) were developed. The approach was verified on simulated data and further illustrated with examples from ecology and agriculture.

We also compared our approach to AIC, BIC and likelihood ratio tests, each of which is based on maximum likelihood estimates for the parameters of the mixture models. The EM algorithm is easily implemented for the models we discussed, but as with a Bayesian approach, computational difficulties may arise, especially when the sample size is insufficient to inform estimation in each component. Chung, Loken, and Schaefer (2004) note that mixture likelihoods can be nearly flat, have multiple local modes, and maxima on the boundary of the parameter space. Each of these issues may adversely impact the small sample behavior of maximum likelihood methods. Putting aside computational issues and personal preferences, a Bayesian approach provides some clear advantages such as the ability to formally compare models with different numbers of components and to quantify the number of modes in body size distributions.

The hierarchical mixture of experts (HME) model and variants described in McLachlan and Peel (2000, Chapter 10) can be used to take advantage of the three well-defined stages in which spontaneous abortion appears to occur in diary cows. Another approach would be to model latent transition probabilities for passing from one stage to the next with a discrete hazards regression model or a continuation-ratio logistic model. These models would attempt to replace the herd-specific weights $\boldsymbol{\omega}_i = (\omega_{i1}, \omega_{i2}, \omega_{i3})$ in the finite mixture model with functions of herd and cow specific covariates, thus generalizing the model. In the absence of such covariates, an approach that borrows strength across herds, and thus is useful for prediction, would be to consider a hierarchical random effects model, for example,

$$\boldsymbol{\omega}_i | \mathbf{a} \overset{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{a}), \qquad \log(\mathbf{a}) | \mathbf{m}, \mathbf{V} \sim N_3(\mathbf{m}, \mathbf{V}).$$

Xu (2005) compares Holling's (1992) boreal forest mammals to boreal prairie mammals and found considerable evidence for $K = 2$ components in both populations. The marginal likelihood ranks the models in M_4, M_3, M_2, M_1, with a decisive difference between models M_1 and M_2, but only a slight difference among models M_4, M_3, and M_2. Similar to the boreal bird data, this suggests constant cluster locations, or centers of attraction, across disperse ecological strata: forest and prairie. The environmental strata could conceivably provide the weights attached to each cluster, but the cluster locations, and possibly spreads, seem rooted beyond these differences, somewhat supporting the limited-morph hypothesis. It would be of interest to fit HME models to these data, perhaps incorporating more ecological strata, including covariates of interest such as percentages of different types of flora and the availability of water and food.

A simulation study of the proposed method's characteristics based on Xu (2005) is provided online.

SUPPLEMENTAL MATERIALS

Simulation study Simulation study of the proposed method's small-sample characteristics.

[Received April 2008. Published Online March 2010.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings 2nd International Symposium Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267–281.

Allen, C. R., Forys, E. A., and Holling, C. S. (1999), "Body Mass Patterns Predict Invasions and Extinctions in Transforming Landscapes," *Ecosystems*, 2, 114–121.

Allen, C. R., Garmestani, A. S., Havlicek, T., Marquet, P. A., Peterson, G., Restrepo, C., Stow, C., and Weeks, B. (2006), "Patterns in Body Mass Distributions: Sifting Among Alternative Hypotheses," *Ecology Letters*, 9, 630–643.

Berger, J. O., and Pericchi, L. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

Bishop, C. M., and Svensén, M. (2003), "Bayesian Hierarchical Mixtures of Experts, in *The Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, eds. U. Kjaerulff and C. Meek, pp. 57–64.

Brown, J. H. P., Marquet, P. A., and Taper, M. L. (1993), "Evolution of Body Size: Consequences of an Energetic Definition of Fitness," *The American Naturalist*, 142, 573–584.

Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society, Series B*, 57, 473–484.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Chung, H., Loken, E., and Schaefer, J. L. (2004), "Difficulties in Drawing Inferences with Finite-Mixture Models: A Simple Example with a Simple Solution," *The American Statistician*, 58, 152–158.

Diebolt, J., and Robert, C. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society, Series B*, 56, 363–375.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Geisser, S., and Eddy, W. F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 589–590.

Halloway, S. R. P. (2007), "Patterns of Abundance and Morphology as Indicators of Ecosystem Status: A Meta-Analysis," *Ecological Complexity*, 4, 128–147.

Han, C., and Carlin, B. P. (2001), "Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review," *Journal of the American Statistical Association*, 96, 1122–1132.

Hanson, T., Bedrick, E., Johnson, W., and Thurmond, M. (2003), "A Mixture Model for Bovine Abortion and Fetal Survival," *Statistics in Medicine*, 22, 1725–1739.

Havlicek, T. D., and Carpenter, S. R. (2001), “Pelagic Species Size Distributions in Lakes: Are They Discontinuous?” *Limnology and Oceanography*, 46, 1021–1033.

Holling, C. S. (1992), “Cross-Scale Morphology, Geometry, and Dynamics of Ecosystems,” *Ecological Monographs*, 62, 447–502.

Hunt, G. (2007), “The Relative Importance of Directional Change, Random Walks, and Stasis in the Evolution of Fossil Lineages,” *Proceedings of the National Academy of Sciences of the United States of America*, 104, 18404–18408.

Hutchinson, G. E., and MacArthur, R. H. (1959), “A Theoretical Ecological Model of Size Distributions Among Species of Animals,” *The American Naturalist*, 93, 117–125.

Jeffreys, H. (1961), *The Theory of Probability* (3rd ed.), Oxford: Oxford University Press.

Jordan, M. I., and Jacobs, R. A. (1994), “Hierarchical Mixtures of Experts and the EM Algorithm,” *Neural Computation*, 6, 181–214.

Kadane, J. B., and Lazar, N. A. (2004), “Methods and Criteria for Model Selection,” *Journal of the American Statistical Association*, 99, 279–290.

Karuppanan, P., Thurmond, M. C., and Gardner, I. A. (1997), “Survivorship Approaches to Measuring and Comparing Cull Rates for Dairies,” *Preventive Veterinary Medicine*, 30, 171–179.

Kass, R. E., and Raftery, A. E. (1995), “Bayes Factors,” *Journal of the American Statistical Association*, 90, 775–795.

Lee, K., Mengersen, K. L., Marin, J.-M., and Robert, C. P. (2008), “Bayesian Inference on Mixtures of Distributions,” in *Proceedings of the Platinum Jubilee of the Indian Statistical Institute*, to appear. Available as arXiv:*0804.2413*.

Marples, P. C. (1998), “Testing for Evidence of Hierarchical System Structure in the Manitoban Boreal Forest,” unpublished Ph.D. dissertation, University of Florida, Gainesville, Florida, US.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

McMahon, T. A., and Bonner, J. T. (1983), *On Size and Life*, New York: Scientific American Books.

Neal, R. M. (1999), “Erroneous Results in ‘*Marginal Likelihood from the Gibbs Output*’,” unpublished manuscript available from the author’s webpage: *http://www.cs.toronto.edu/~radford/ftp/chib-letter.pdf*.

O’Hagan, A. (1995), “Fractional Bayes Factor for Model Comparison,” *Journal of the Royal Statistical Society, Series B*, 57, 99–138.

Peters, R. H. (1983), *The Ecological Implications of Body Size*, Victoria: Cambridge University Press.

Restrepo, C., Renjifo, L. M., and Marples, P. (1997), “Frugivorous Birds in Fragmented Neotropical Montane Forests: Landscape Pattern and Body Mass Distribution,” in *Tropical Forest Remnants: Ecology, Management, and Conservation of Fragmented Communities*, eds. W. F. Laurance and J. R. O. Bierregaard, Chicago: The University of Chicago Press, pp. 171–189.

Richardson, S., and Green, P. J. (1997), “On Bayesian Analysis of Mixtures with an Unknown Number of Components,” *Journal of the Royal Statistical Society, Series B*, 59, 731–792.

Roeder, K., and Wasserman, L. (1997), “Practical Bayesian Density Estimation Using Mixtures of Normals,” *Journal of the American Statistical Association*, 92, 894–902.

Schmidt-Nielsen, K. (1984), *Scaling: Why is Animal Size so Important?* Cambridge: Cambridge University Press.

Schwartz, G. (1978), “Estimating the Dimension of a Model,” *Annals of Statistics*, 6, 461–464.

Song, X.-Y., and Lee, S.-Y. (2002), “A Bayesian Model Selection Method with Applications,” *Computational Statistics and Data Analysis*, 40, 539–557.

Stanley, S. M. (1973), “An Explanation for Cope’s Rule,” *Evolution*, 27, 1–25.

Steele, R., Raftery, A., and Emond, M. (2006), “Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS),” *Journal of Computational and Graphical Statistics*, 15, 712–734.

Stow, C., Allen, C. R., and Garmestani, A. S. (2007), “Evaluating Discontinuities in Complex Systems: Toward Quantitative Measures of Resilience,” *Ecology and Society*, 12, 1–12.

Thurmond, M., Branscum, A., Johnson, W., Bedrick, E., and Hanson, T. (2005), “Predicting the Probability of Abortion in Dairy Cows: A Hierarchical Bayesian Logistic-Survival Model Using Sequential Pregnancy Data,” *Preventive Veterinary Medicine*, 68, 223–239.

Wilson, E. O. (1953), “A New Interpretation of the Frequency Curves Associated with Ant Polymorphism,” *Insectes Sociaux*, 1, 75–80.

Xu, L. (2005), “Bayesian Methods for Comparing Populations with Multimodal Distributions,” unpublished Ph.D. dissertation, University of New Mexico Department of Mathematics and Statistics.

Yan, M., and Ye, K. (2007), “Determining the Number of Clusters Using the Weighted Gap Statistic,” *Biometrics*, 63, 1031–1037.