Journal of Data Science **12**(2014), 175-196

A Comparison of Statistical Tools for Identifying Modality in Body Mass Distributions

Ling Xu[1], Edward J. Bedrick[2], Timothy Hanson[3*] and Carla Restrepo[4]
[1]*James Madison University,* [2]*University of New Mexico,*
[3]*University of South Carolina and* [4]*University of Puerto Rico*

Abstract: The assessment of modality or "bumps" in distributions is of interest to scientists in many areas. We compare the performance of four statistical methods to test for departures from unimodality in simulations, and further illustrate the four methods using well-known ecological datasets on body mass published by Holling in 1992 to illustrate their advantages and disadvantages. Silverman's kernel density method was found to be very conservative. The excess mass test and a Bayesian mixture model approach showed agreement among the data sets, whereas Hall and York's test provided strong evidence for the existence of two or more modes in all data sets. The Bayesian mixture model also provided a way to quantify the uncertainty associated with the number of modes. This work demonstrates the inherent richness of animal body mass distributions but also the difficulties for characterizing it, and ultimately understanding the processes underlying them.

Key words: Bayesian, body-size data, excess mass test, kernel density estimate, mixture model.

1. Introduction

Scientists across many realms are interested in quantifying modality in distributions. Examples abound, for example in economics (Henderson, Parmeter and Russell, 2008), micro-array gene expression data (e.g., Dazard and Rao, 2010), and astrophysics (e.g., Escobar and West, 1995). Here, we consider ecological data, which often departs from the "ideal" normal distribution. This not only has implications for data analysis but also for the interpretation of the ecological and evolutionary problems under observation. One point in case concerns the distribution of body size in animal assemblages. Some authors have described

*Corresponding author.

these distributions as unimodal and continuous (e.g., Hutchinson and MacArthur, 1959; May, 1986) whereas others have described them as discontinuous such that species aggregate around certain body sizes leading to multimodal distributions (e.g., Oksanen *et al.*, 1979; Holling, 1992; see Figure 1). The ecological and evolutionary implications of these divergent views are profound. Whereas the former implies the existence of a single optimal body size (Stanley, 1973; Brown *et al.*, 1993) the latter implies the existence of several optima (Griffiths, 1986), raising questions about the underlying processes (e.g., Scheffer and van Nees, 2006; Allen and Holling, 2008).

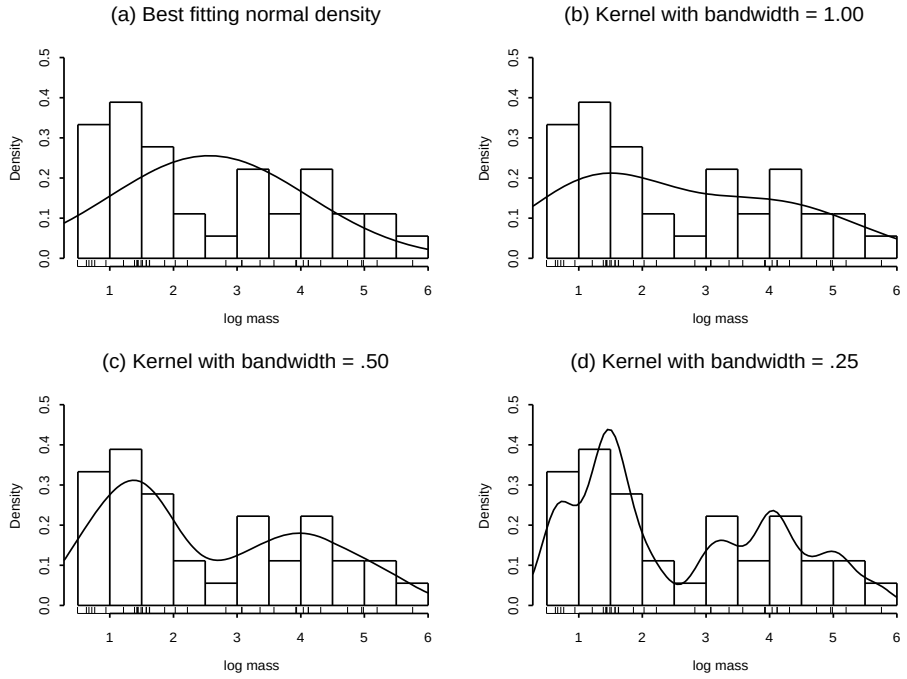

Figure 1: Distribution of body size of Holling's (1992) Boreal Forest Mammal (BFM) data. (a) Histogram and estimated density function based on the best fitting normal distribution. Probability density functions based on a normal kernel estimate in which the bandwidth is (b) 1.0, (c) 0.50, and (d) 0.25

The occurrence of multiple modes in body size is not limited to species assemblages. They have been widely documented in many animal populations as intersexual (Ipiña and Durand, 2000), intrasexual (Wright, 1968; Rüppell and Heinze, 1999), and caste (Wilson, 1953) size polymorphisms. In these systems, individuals sharing similar body sizes perform well-defined tasks and in this sense, they represent functional groups. Other examples include intra (Wright, 1968; D'Onghia *et al.*, 2000) and inter (Grant, 1986) population size differences. In all

instances, multimodality seems to be associated with gross intra and inter population heterogeneity resulting from a mixture of subpopulations, raising questions about their origin and the processes maintaining them in space and time.

Revealing and quantifying multimodality in body size data, however, has been challenging for ecologists and evolutionary biologists, as demonstrated by the diversity of methods that they have used (Allen *et al.*, 2006 and references therein) and the concerns that others have raised (Manly, 1996; Siemann and Brown, 1999). Statisticians have faced similar challenges over the last 30 years as demonstrated by the variety of methods that they have devised (Silverman, 1981; Hartigan and Hartigan, 1985; Silverman, 1986). Yet, there is little agreement on what tools might be most appropriate for revealing and quantifying multimodality, not to mention investigating the processes underlying these distributions. In this paper we: (1) review three statistical methods for detecting multiple modes, (2) introduce a Bayesian test for assessing modality, (3) compare the four approaches in simulations on a variety of density shapes, (4) evaluate and compare the performance of these methods on body mass data for species assemblages, and (5) discuss the advantages and disadvantages of each method in terms of the future development of these tools to address a variety of ecological and evolutionary questions. Although we focus on body size data, we emphasize that many other data are amenable to similar analyses.

2. Approaches

A mode of a continuous probability distribution is a location at which the probability density assumes a local maximum value. Distributions with a single mode (unimodal) are insufficient to describe many datasets, and several classes of methods have been devised to reveal more complex distributions. These include histograms, kernel density estimates and mixture models. Histograms are primarily descriptive, while kernel estimates and mixture models allow for inference using both non-parametric and parametric approaches.

2.1 Kernel Density Estimation: The Silverman and Hall and York Tests

These two tests of modality are based on kernel density estimates. Suppose we have a sample $x_1, x_2, \cdots, x_n$ from a population with an unknown density function $f(t)$. A popular non-parametric estimate for this density is the so-called kernel estimate

$$\hat{f}(t;b) = \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{t-x_i}{b}\right),$$

where b is a user defined bandwidth and $K(\cdot)$ is a user defined kernel function.

Two popular choices for the kernel function are the uniform or box kernel $K(t) = 1$ for $-0.50 \leq t \leq 0.50$ and the normal kernel $K(t) = \exp(-0.5t^2)/\sqrt{2\pi}$. With a box kernel, the estimate is a smoothed version of the sample histogram. For simplicity, we consider normal kernels. An important property of normal kernel estimates is that the number of modes or peaks in $\hat{f}(t)$ is a non-increasing function of the bandwidth. That is, larger bandwidths lead to smoother estimated densities with fewer modes (Figure 1). Givens and Hoeting (2005) discuss several methods for choosing the bandwidth.

Silverman (1981) proposed a test of modality that is based on the minimum bandwidth $b_{\min}$ that produces a unimodal kernel estimator $\hat{f}(t; b_{\min})$. The idea behind Silverman's test is that a large value of the bandwidth b is needed to smooth the data and produce a unimodal density estimate if the population density $f(t)$ has two or more modes. Thus, large values of $b_{\min}$ provide evidence that the underlying density is not unimodal. To test the null hypothesis that $f(t)$ is unimodal, Silverman suggests the p-value

$$\text{pval} = P(\ b_{\min} > b_{\min}^{\text{obs}} \mid f(t) \text{ is unimodal}),$$

where $b_{\text{obs}}^{\min}$ is the minimum bandwidth based on the sample. This p-value is estimated using a bootstrap procedure where repeated random samples $x_1^*, \cdots, x_n^*$ are selected from a distribution with density $\hat{f}(t; b_{\min}^{\text{obs}})$. For each bootstrap sample, the minimum bandwidth $b_{\min}^*$ is computed from the kernel estimate

$$\hat{f}^*(t; b) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{t - x_i^*}{b}\right).$$

The estimated p-value is the proportion $\widehat{\text{pval}}$ of bootstrap samples with $b_{\min}^* > b_{\min}^{\text{obs}}$. Given that $b_{\min}^* > b_{\min}^{\text{obs}}$ if and only if $\hat{f}^*(t; b_{\min}^{\text{obs}})$ has more than one mode, the p-value is just the proportion of bootstrap samples in which $\hat{f}^*(t; b_{\min}^{\text{obs}})$ has more than one mode. Silverman gives a simple method to sample the distribution $\hat{f}(t; b_{\min}^{\text{obs}})$. In practice, the distribution $\hat{f}(t; b_{\min}^{\text{obs}})$ is rescaled to have the same mean and variance as the observed data.

Hall and York (2001) established that the p-value for Silverman's test is too large, even in large samples. An important implication of this result is that Silverman's test may have low power and fail to detect multiple modes when they exist. To correct this, Hall and York (2001) inflate the minimum bandwidth $b_{\min}^{\text{obs}}$ in Silverman's test. For a test with size α (for example, $\alpha = 0.05$), the hypothesis of unimodality is rejected if the proportion $p^*(\alpha)$ of bootstrap samples where $\hat{f}^*(t; \lambda_\alpha b_{\min}^{\text{obs}})$ has more than one mode is less than α. Here λ_α is a decreasing function of α with $\lambda_\alpha > 1$ and $\lambda_{.05} \approx 1.13$. We note that $p^*(\alpha) \leq \widehat{\text{pval}}$ so the adjustment has the desired effect of making Silverman's test less conservative. Hall

and York (2001) also propose a procedure where λ_α is estimated using bootstrap methods. We found this modification to be unnecessary in our examples.

The proportion $p^*(\alpha)$ is not a p-value because it depends on the test size α. If we define the p-value to be the minimum value of α such that $p^*(\alpha) \leq \alpha$, then $p^*(\alpha)$ is an upper bound on the p-value when $p^*(\alpha) < \alpha$ and a lower bound on the p-value when $p^*(\alpha) > \alpha$. The p-value can be obtained by iteratively evaluating $p^*(\alpha)$ until $|p^*(\alpha) - \alpha|$ is small.

Hall, Minnotte and Zhong (2004) proposed a calibration of Silverman's test using a non-Gaussian kernel. In contrast to a Gaussian kernel, the number of modes is a nonmonotone function of the bandwidth with some non-Gaussian kernels. They used three different non-Gaussian kernels: Epanechnikov, biweight and triweight. When testing for modality, they showed that using non-Gaussian kernels does not substantially alter the conclusions of Silverman's test. We will not consider this approach here.

Hall and York's (2001) adjustment to Silverman's (1981) approach is available in the package `silvermantest` available online from www.uni-marburg.de/fb12/stoch/research/rpackage; this package is referenced in Vollmer, Holzmann and Schwaiger (2013).

2.2 The Dip and Excess Mass Tests

A difficulty with implementing Silverman's test is that the interval on which the density is estimated must be constrained to reduce the influence of isolated extreme data values producing spurious modes or bumps (see Hall and York, 2001). Two histogram-based methods that were developed to circumvent this problem are the dip (Hartigan and Hartigan, 1985) and the excess mass (Müller and Sawitski, 1991) tests, neither of which requires estimating the density. The test statistics are equivalent for one dimensional densities as considered here, so we restrict attention to the excess mass test.

For testing whether a density $f(t)$ is unimodal against the alternative hypothesis that the distribution is bimodal, the measure of excess mass for two modes is defined by

$$\mathrm{EM}_2(L) = \max_{C_1,C_2} \left[\sum_{j=1}^{2} \{\hat{p}\,(C_j) - L \text{ length}\,(C_j)\} \right],$$

where $L > 0$ is a constant, the maximum is taken over all disjoint intervals C_1 and C_2 on the line and $\hat{p}(C_j)$ is the proportion of the sample that falls in C_j (Figure 2(a)). The measure of excess mass for one mode is defined similarly: $\mathrm{EM}_1(L) = \max_C[\hat{p}(C) - L \text{ length}(C)]$.

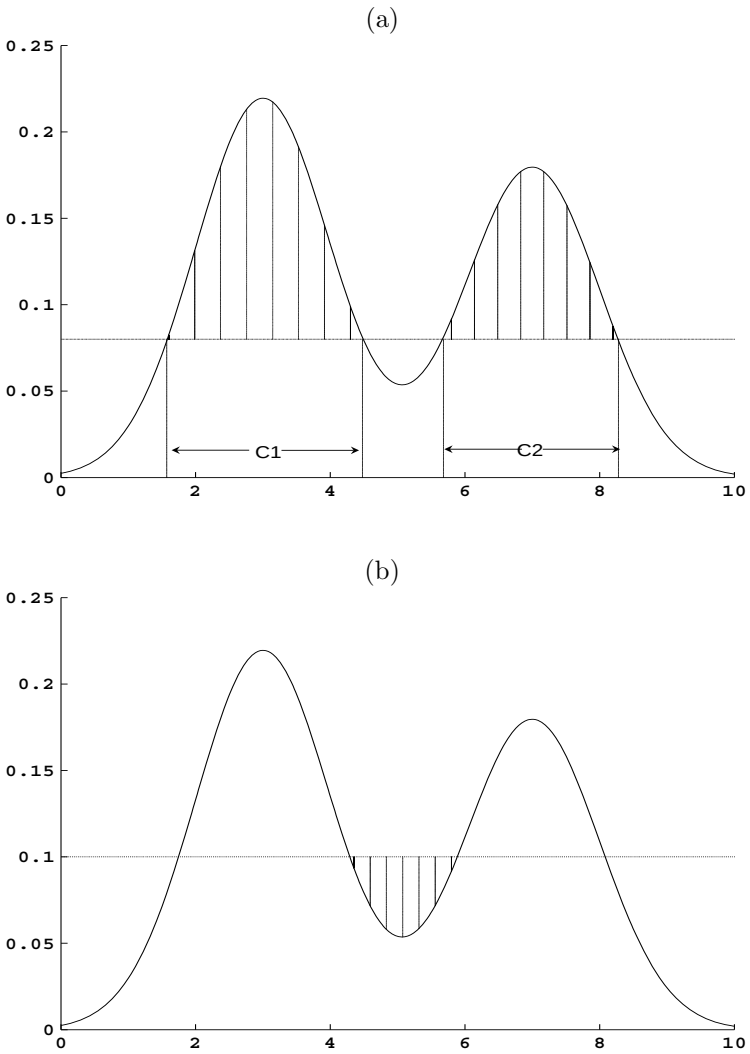

Figure 2: Probability density function for a (constructed) bimodal distribution to illustrate the concepts behind the excess mass test. (a) The shaded area represents the value of $\mathrm{EM}_2(L)$ for the choice of L. (b) The excess mass statistic Δ is the minimal amount of mass that has to be moved (from the 2nd hump above vertical bar to the shaded area) to turn two modes into one

The basic idea of the test is that two distinct modes will produce disjoint intervals with a significant density relative to the interval length, leading to a much larger value of $\mathrm{EM}_2(L)$ than when the density is unimodal. The excess mass statistic, Δ, is the maximum value of $\mathrm{EM}_2(L) - \mathrm{EM}_1(L)$ over all $L > 0$. Larger values of Δ provide stronger evidence against unimodality. Alternatively, Δ is the minimal amount of mass that has to be moved to turn a bimodal distribution

into a unimodal distribution (Müller and Sawitzki, 1991); see Figure 2(b).

The p-value of the excess mass test is evaluated through a bootstrap method developed by Cheng and Hall (1998). Under the null hypothesis that $f(t)$ is unimodal, Cheng and Hall (1998) show that the distribution of Δ depends on a single parameter that is a function of the density at the unique mode and the curvature of the density at the unique mode. Kernel density methods are used to estimate this parameter and then bootstrap samples are generated from a family of distributions indexed by this parameter. The p-value is the proportion of bootstrap samples where $\Delta^* > \Delta$.

It is important to mention that the excess mass test addresses whether the underlying distribution has one mode or two. In comparison, Silverman's test and Hall and York's test are ideally suited to assess whether the underlying distribution is unimodal or not.

The dip test (Hartigan and Hartigan, 1985) is available from the Comprehensive R Archive Network (CRAN) in the `diptest` package, written by Martin Maechler and Dario Ringach. The p-value for this test is computed differently than suggested by Cheng and Hall (1998) and is very conservative; we are unaware of an `R` package that implements Cheng and Hall's correction.

3. Mixture Models

Mixture models provide a flexible parametric approach to density estimation that is different from non-parametric kernel methods (McLachlan and Peel, 2000). The basic idea is to model $f(t)$ as a weighted sum of k normal densities

$$f(t) = \sum_{i=1}^{k} w_i \ \phi(t; \mu_i, \sigma_i),$$

where the weights w_i sum to one and $\phi(t; \mu_i, \sigma_i)$ is a normal density with mean μ_i and variance σ_i^2. The normal component $\mathcal{N}(\mu_i, \sigma_i)$ could describe the body size distribution for a sub-group of a population, while w_i represents the size of the sub-group or component. Within this framework, the i^{th} subgroup has a single mode at the mean μ_i of the i^{th} component. However, the number of modes for the population distribution given by the mixture model may have fewer than k modes (and can not have more than k modes) if the components are not well separated, or the weights of certain subgroups are small.

Mixture models have been used both by frequentists (McLachlan and Peel, 2000) and Bayesians (Roeder and Wasserman, 1997; Richardson and Green, 1997) in a variety of settings. Unlike standard frequentist inference which treats the number of components k as fixed, even when the data are used to determine the best choice for k, the Bayesian approach allows uncertainty in k to be handled

easily and naturally. Specifically, in the Bayesian analyis of the mixture model, one specifies a joint prior distribution for k, $\mu = (\mu_1, \cdots, \mu_k)$, $\sigma = (\sigma_1, \cdots, \sigma_k)$ and $w = (w_1, \cdots, w_k)$. Given the data, the prior distribution is updated using Bayes theorem to provide the posterior distribution for the parameters, which is used for inferences. Prior information is often unavailable, so it is common practice to use non-informative (sometimes called vague) priors or reference priors. Roeder and Wasserman (1997) show that the posterior distribution is improper when a standard reference prior is used with the normal mixture model. Improper posteriors can not be interpreted probabilistically, leading to a breakdown in the Bayesian paradigm. These same authors propose a partially proper prior distribution that produces a proper posterior distribution given a finite number of observations and show how to compute the posterior distribution using a Gibbs' sampling procedure. Their approach is most appropriate when the number of components k is assumed to be known, but they provide an approximate method that allows k to be unknown.

An alternative is to fit the mixture model using a reversible jump Markov chain Monte Carlo (MCMC) procedure (Richardson and Green, 1997). Traditional MCMC methods, such as Roeder and Wasserman's algorithm, sample parameters within one given model with a fixed number of components whereas the reversible jump MCMC samples parameters within the a "current" model of fixed size k but also "jumps" between models of different sizes (e.g., $k-1$ or $k+1$) with different sets of parameters, allowing for several competing mixture models with different numbers of components to be fit simultaneously. This yields an overall model that encompasses all competing models and includes k as a parameter to be estimated along with the parameters for each sub-model.

Richardson and Green's (1997) approach for the normal mixture model with an unknown but finite number of components k uses a vague, data-driven prior distribution that produces a proper posterior and works well in practice. The prior on k is a Poisson distribution with rate λ, truncated to not be larger than $K_{\max}$; we use $K_{\max} = 20$ in this paper. For each fixed value of k, the priors for μ, σ and w are assumed to be independent. The weights $(w_1, \cdots, w_k)$ have a Dirichlet$(1, 1, \cdots, 1)$ distribution, which is a multivariate generalization of a uniform distribution. The means μ_i have independent $\mathcal{N}(\xi, \kappa)$ distributions, subject to the identifiability constraint $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$. The prior mean ξ is selected to be the average of the minimum and maximum data value. The standard deviation κ is set to be the sample range R. This centers the prior in the middle of the data and keeps the prior flat over an interval of variation of the data. Decreasing κ shrinks the means towards ξ. The component precisions $\tau_i = 1/\sigma_i^2; i = 1, 2, \cdots, k$ are independent with identical gamma $\Gamma(\alpha, \beta)$ distributions where β has a $\Gamma(g, h)$ distribution and $\alpha = 2$, $g = 0.2$, and $h = 10/R^2$. This

reflects the prior belief that the precisions should be similar, but their absolute size should be left arbitrary.

Recently, Baştürk, Hoogerheide, de Knijff and van Dijk (2012) developed a Bayesian approach to detecting multimodality based on finite mixtures. Xu (2005) also examined such tests based on Dirichlet process mixtures (Escobar and West, 1995) and finite mixtures of normals (Roeder and Wasserman, 1997).

4. Comparison of Tests

For the Silverman, Hall and York, and excess mass tests, test statistics were computed for each data set, followed by 1,000 bootstrap replications to estimate their p-values. Silverman's and Hall and York's tests can be implemented in R using a modified version of Davison and Hinkley's code (1997, p. 189), which is based on built-in functions for bootstrapping and kernel density estimation. For our analysis, the minimum bandwidth tests were programmed in $C++$. A FORTRAN program kindly provided by Professor M.-Y. Cheng was used to implement the excess mass test.

4.1 Hall and York vs. Excess Mass

Hall and York's test and the excess mass test are ideally suited for testing whether a distribution is unimodal or bimodal. Hall and York (2001) and Cheng and Hall (1998) developed generalizations of these tests that apply to the setting where the null and alternative models have m and $m + 1$ modes, respectively, but aspects of implementing these tests have not been explored. For all practical purposes, the two approaches are currently restricted to checking for a single mode and do not provide a clear-cut approach for identifying the number of modes. We note that minimum bandwidth test should have power to detect multimodal alternatives, but the excess mass test is not expected to detect distributions with three or more modes, unless at least two modes are strongly identifiable.

We designed a simulation study to compare Hall and York's test and the excess mass test. Earlier separate studies (Cheng and Hall, 1996; Hall and York, 2001) compared these tests to Silverman's test, but not to each other. We will include Silverman's test in our study for completeness.

Figure 3 plots the 10 distributions used in our study. The study includes a range of distributions, from symmetric and unimodal to skewed distributions with two or three modes, some of which are not pronounced. For each distribution, we generated 1000 samples of size $n = 50$ and 200. In each sample, the three statistics were computed, followed by 1000 bootstrap replications to estimate their p-values. The proportion of the p-values below 0.10, 0.05, and 0.01 was evaluated.

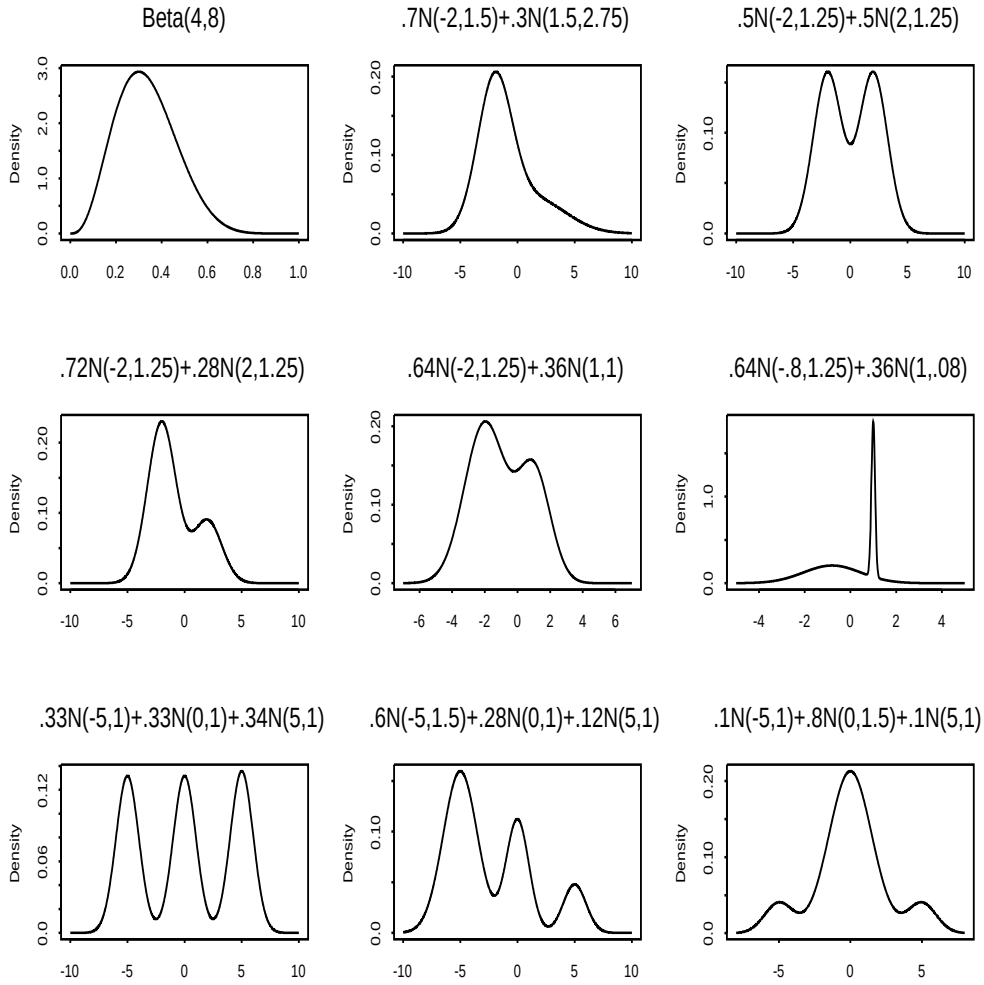

Figure 3: Distributions used in the simulation study

Table 1 gives the estimated size or power for tests with nominal levels of 0.01, 0.05, and 0.1. The excess mass test is slightly conservative under the null hypothesis, but has higher power than Hall and York's test when the distribution has prominent modes. Hall and York's test is the best for identifying bimodal or multimodal distributions where modes are small. In some cases, for example the unimodal normal mixture considered in Figure 3's top-middle panel, the size of Hall and York's test is much higher than nominal levels even for a sample sizes of 200. For this distribution, 500 observations are needed before the actual size falls to 0.05. Therefore, more advanced calibration of Silverman's test is probably needed here. Overall, neither the excess mass nor Hall and York's test is uniformly better. As expected, Silverman's test is too conservative to recommend.

Table 1: Estimated size (first 6 rows) and power of tests based on 1000 samples

			$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.1$			Bayes	
Distribution	Modes	n	Silv.	H. & Y.	E.M.	Silv.	H. & Y.	E.M.	Silv.	H. & Y.	E.M.	BF>100	BF>10
$\mathcal{N}(0,1)$	1	50	0.00	0.01	0.01	0.01	0.07	0.04	0.03	0.12	0.08	0.00	0.00
		200	0.00	0.01	0.00	0.01	0.05	0.04	0.03	0.10	0.08	0.00	0.00
Beta$(4,8)$	1	50	0.00	0.01	0.01	0.01	0.05	0.05	0.02	0.11	0.10	0.00	0.00
		200	0.00	0.01	0.01	0.01	0.05	0.04	0.02	0.10	0.08	0.00	0.00
$0.7\mathcal{N}(-2,1.5)+0.3\mathcal{N}(1.5,2.75)$	1	50	0.01	0.09	0.00	0.03	0.15	0.03	0.10	0.24	0.05	0.02	0.52
		200	0.01	0.10	0.00	0.01	0.09	0.03	0.11	0.35	0.04	0.00	0.18
$0.5\mathcal{N}(-2,1.25)+0.5\mathcal{N}(2,1.25)$	2	50	0.05	0.23	0.27	0.21	0.41	0.45	0.34	0.48	0.58	0.21	0.73
		200	0.06	0.30	0.78	0.60	0.84	0.88	0.65	0.89	0.94	0.94	1.00
$0.72\mathcal{N}(-2,1.25)+0.28\mathcal{N}(2,1.25)$	2	50	0.01	0.05	0.04	0.02	0.17	0.11	0.08	0.24	0.17	0.12	0.64
		200	0.19	0.43	0.05	0.19	0.43	0.18	0.29	0.49	0.25	0.45	0.92
$0.64\mathcal{N}(-2,1.25)+0.36\mathcal{N}(1,1)$	2	50	0.02	0.11	0.07	0.07	0.24	0.17	0.14	0.34	0.26	0.03	0.19
		200	0.03	0.28	0.16	0.12	0.47	0.29	0.23	0.56	0.40	0.27	0.66
$0.64\mathcal{N}(-.8,1.25)+0.36\mathcal{N}(1,.08)$	2	50	0.03	0.17	0.29	0.13	0.33	0.47	0.24	0.43	0.56	0.91	0.96
		200	0.16	0.63	0.56	0.47	0.78	0.75	0.62	0.84	0.82	1.00	1.00
$0.33\mathcal{N}(-5,1)+0.33\mathcal{N}(0,1)$	3	50	0.04	0.18	0.83	0.16	0.41	0.97	0.28	0.56	0.99	0.99	1.00
$+0.34\mathcal{N}(5,1)$		200	0.17	0.64	1.00	0.54	0.70	1.00	0.66	0.73	1.00	0.99	1.00
$0.6\mathcal{N}(-5,1.5)+0.28\mathcal{N}(0,1)$	3	50	0.03	0.13	0.22	0.08	0.32	0.40	0.18	0.45	0.54	0.57	1.00
$+0.12\mathcal{N}(5,1)$		200	0.17	0.59	0.74	0.56	0.67	0.90	0.62	0.72	0.95	1.00	1.00
$0.1\mathcal{N}(-5,1)+0.8\mathcal{N}(0,1.5)$	3	50	0.00	0.08	0.00	0.04	0.24	0.02	0.10	0.34	0.06	0.05	0.51
$+0.1\mathcal{N}(5,1)$		200	0.07	0.35	0.00	0.24	0.49	0.02	0.36	0.56	0.04	0.39	0.92

4.2 The Bayesian Mixture Model

We also fit the model described in Section 3 to the data sets in the previous section; posterior inference for each data set is based on 10,000 iterates thinned to 1,000 after a burn-in of 2,000. The default prior gives a 0.95 prior probability of one mode and 0.05 of two or more modes. Thus, our prior places approximately 19 times the weight on one mode as it places on two or more modes. R code to implement this model is provided in the Appendix.

An important issue with our Bayesian analysis is the extent to which the prior influences our inferences on the number of modes. To address this concern, we computed the ratio of the posterior to prior odds that the density has two or modes. This is the Bayes factor in support of the hypothesis that the density has two or more modes. Bayes factors are commonly used as an objective measure of the support in the data for an hypothesis. A standard benchmark (Jeffreys, 1961; Kass and Raftery, 1995) is that Bayes factors exceeding 10 provide "strong" evidence for a hypothesis, and exceeding 100 provides "decisive" evidence for a hypothesis. Both of these cutoffs are represented in Table 1. The cutoff $BF > 10$ seems to be a bit liberal in terms of Type II error, and so we focus only on $BF > 100$ and compare it to the excess mass and Hall and York tests for $\alpha = 0.05$.

The Bayesian test does as well as or better than excess mass for all densities when $n = 200$. For $n = 50$, the excess mass approach fares better on the top-right and middle-middle densities in Figure 3; both densities have two modes with relatively small valleys between them. In general, the Bayesian approach has about the same as or a bit worse power than excess mass for $n = 50$, but clearly outperforms the excess mass test on all densities when $n = 200$. This same finding roughly holds for the Bayesian test versus Hall and York, except that the latter test does quite a bit better than the Bayes test for the bottom-right density at $n = 50$, and a bit better at $n = 200$. However, the Bayesian test has Type II errors of only 0.02 and 0.00 for the top-middle density at $n = 50$ and $n = 200$, whereas the Hall and York test has Type II errors of 0.15 and 0.09 for nominal rate $\alpha = 0.05$.

A referee pointed out that the number of normal components necessary to adequately fit skewed distributions may be unnecessarily large, compared to models that use skewed and/or heavy-tailed components. We investigate this issue by considering a particular class of unimodal skewed t distributions given by

$$f_{\nu,\gamma}(x) = \frac{2}{\gamma + 1/\gamma} \left[f_\nu(\gamma x) I\{x < 0\} + f_\nu(x/\gamma) I\{x \geq 0\} \right],$$

where $f_\nu(x)$ is the density of a student t with degrees of freedom ν. We investigate the impact that unimodal heavy-tailed ($\nu = 3$) versus lighter-tailed ($\nu = 15$) densities have on modal estimation for three levels of skewness, $\gamma = 1$ (no skew),

$\gamma = 2$ (moderate skew), and $\gamma = 3$ (heavy skew). Table 2 gives the results from simulating 500 Monte Carlo data sets at each setting of ν and γ, with two sample sizes, $n = 50$ and $n = 200$. Again, we used the default prior described in Section 3 for the reversible jump mixture model. Overall, both heavy-tails and skew ($df = 3$, $\gamma \geq 2$) render both the Bayesian test and the Hall and York test unreliable, although the Bayesian test has much smaller type II error. Symmetric heavy-tails ($\nu = 3, \gamma = 1$) poses no problem for the Bayesian test, but destroys Hall and York's test. For lighter tails ($df = 15$), the Bayesian test performs much better than Hall and York's for every setting, but still provides Type II error of 0.15 for $n = 200$ and heavy skew. In all skewed cases, there are many data sets that produce estimates of k and the number of modes greater than one, i.e., heavy tails and/or skew artificially increases the numbers of components needed for the normal mixture model and also increases the estimated posterior number of modes. In such circumstances, finite mixtures of skewed t distributions could vastly simplify the estimates and provide more accurate estimates of posterior modes; see, e.g., Lin, Lee and Hsieh (2007) and Ho, Pyne and Lin (2012).

Table 2: Heavy-tail and/or skew simulation. First four rows are Type II errors; next four rows are the median and 95% interval of the posterior modes for the number of modes and the number of components k. Last two rows are Type II errors for Hall and York's test

		$df = 3$			$df = 15$		
		$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$
$P(\text{BF} > 10)$	$n = 50$	0.33	0.79	0.96	0.01	0.25	0.62
	$n = 200$	0.25	0.79	0.99	0.01	0.24	0.76
$P(\text{BF} > 100)$	$n = 50$	0.04	0.21	0.32	0.00	0.02	0.05
	$n = 200$	0.02	0.12	0.41	0.00	0.03	0.15
$\widehat{\text{mode}}$	$n = 50$	1 [1,2]	2 [1,3]	2 [1,3]	1 [1,1]	1 [1,2]	1 [1,2]
	$n = 200$	1 [1,2]	1 [1,2]	2 [1,3]	1 [1,1]	1 [1,2]	1 [1,2]
$\widehat{k}$	$n = 50$	2 [1,3]	2 [1,3]	3 [2,4]	1 [1,2]	2 [1,2]	2 [1,3]
	$n = 200$	3 [2,3]	3 [2,4]	4 [3,5]	1 [1,2]	2 [2,3]	3 [2,4]
Hall & York	$n = 50$	0.42	0.45	0.44	0.11	0.18	0.17
$\alpha = 0.05$	$n = 200$	0.65	0.60	0.59	0.13	0.15	0.18

5. A Case Study: Body Size Distributions in Animal Assemblages

The performance of the methods outlined above was evaluated on Holling's (1992) data for North American boreal forest birds ($n = 101$; BFB) and mammals ($n = 36$; BFM), and North American prairie birds ($n = 108$; BPB) and

mammals ($n = 53$; BPM). The four data sets represent a complete inventory of the species recorded in these biomes, each species characterized by its mass mean value obtained from the literature (Holling, 1992). These data sets were also investigated by Manly (1996) and Siemann and Brown (1999). Throughout this paper we use the $\log_{10}$-transformed values of body mass.

5.1 The Silverman, Hall and York and Excess Mass Tests

The three tests differed in their ability to distinguish whether the distributions are unimodal or not (Table 3). According to Silverman's and the excess mass tests, only the BFP and the BPM datasets, respectively, provide support for two or more modes using a benchmark of $p < 0.050$ for statistical significance. For three of the four examples, the excess mass test, which is calibrated to have the correct size in large samples, is more conservative (larger p-values) than Silverman's test, which is known to be conservative. On the other hand, Hall and York's test suggests that all four data sets have at least two modes.

Table 3: Bootstrap p-values for tests of one mode using Holling's data. The minimum bandwidth for Silverman's tests is given in parentheses

		Test		
Data Set	n	Silverman	Hall and York	Excess Mass
BPB	106	0.062 (0.862)	0.005	0.107
BFB	101	0.037 (0.393)	0.003	0.072
BPM	53	0.077 (0.392)	0.013	0.004
BFM	36	0.106 (0.910)	0.024	0.195

5.2 Mixture Models

We computed the mixture density and the corresponding number of components and modes associated with each sample from the posterior. A Bayesian density estimate was obtained by averaging the estimated mixture densities across posterior samples.

The Bayesian density estimates mimic the shape of the histograms for the data sets (Figure 4). The posterior probabilities of the number of components and the number of modes in $f(t)$ presented in Table 4 suggest that each distribution has at least two components and at least two modes. The boreal forest birds and prairie mammals have Bayes factors exceeding 100 (Table 4), indicating "decisive" evidence toward two or more modes. Prairie birds and forest mammals have "strong" evidence toward two or more modes. Overall, on the actual data analyses, the Bayesian approach agrees with the excess mass test.

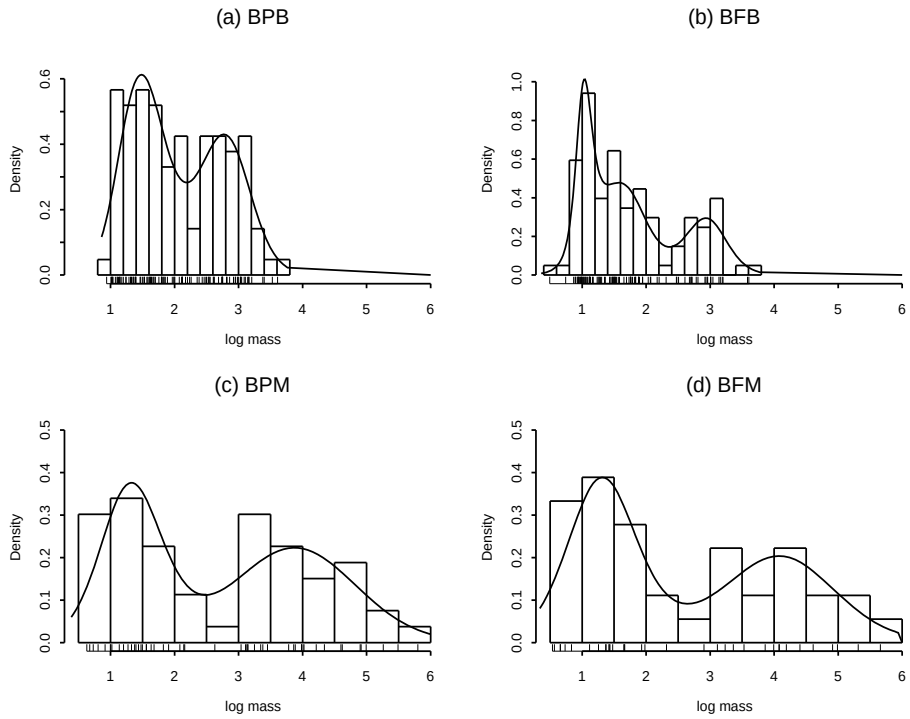

Figure 4: Holling's body mass distributions with Bayesian density estimates

Table 4: Posterior probabilities for number of components and modes in Holling's data

			Posterior Probability of				
Data	n		1	2	3	4+	Bayes Factor
BPB	106	Components	0.17	0.80	0.03	0.00	∼70
		Modes	0.20	0.80	0.00	0.00	
BFB	101	Components	0.000	0.95	0.05	0.00	∼590
		Modes	0.03	0.95	0.02	0.00	
BPM	53	Components	0.09	0.81	0.09	0.01	∼130
		Modes	0.12	0.87	0.01	0.00	
BFM	36	Components	0.13	0.73	0.12	0.02	∼70
		Modes	0.22	0.77	0.01	0.00	

6. Discussion

The four methods evaluated in this paper on a well-known dataset demonstrated the inherent richness of animal body mass distributions but also the dif-

ficulties for characterizing it. As expected, Silverman's test and the excess mass tests were less likely than Hall and York's test to detect more than one mode. Although the Bayesian approach and excess mass give similar conclusions, the Bayesian mixture model is potentially more informative, as it is able to provide quantification on the number of modes.

Our analyses using Silverman's test yielded similar results as those reported by Manly (1996). In both cases there was weak evidence for multimodality despite the fact that a visual interpretation of the histograms indicated the opposite. For example, the histogram of the boreal forest mammal data (Figure 1) appears bimodal but the p-value for Silverman's test is 0.105. Of course, the sample size is small, and the visual assessment of the histogram is influenced by the choice of bins, but similar inconsistencies are found with larger data sets in Holling's series. The primary issue here is that Silverman's test is conservative, thus this test has low power to detect deviations from a unimodal distribution. On the other hand, Hall and York's modification of Silverman's test showed evidence for two or more modes in all four datasets, whereas the excess mass test in only one (BPM, Table 3). This together with the simulations of Section 4 indicates that the performance of the methods differ depending on the characteristics of the data. Studies of the small sample properties of the three non-parametric tests considered here (Cheng and Hall, 1996; Hall and York, 2001; Xu, 2005) suggest the following four general conclusions: (1) Hall and York's modification should be used with Silverman's test to eliminate the conservativeness of the latter. (2) The excess mass test is slightly conservative under the null hypothesis of one mode, but has higher power than Hall and York's test when the distribution has two or more prominent modes. This may explain why the excess mass test appears more sensitive than Hall and York's test with the boreal prairie mammal (BPM) data, which has two prominent and somewhat separated modal regions. (3) Hall and York's test is more powerful than the excess mass test for identifying bimodal or multimodal distributions where the modes have small excess mass. (4) Neither the excess mass test nor Hall and York's test is uniformly better.

The Bayesian approach compares well to both excess mass and Hall and York. The Bayesian approach seems to have a steeper increase of power with sample size than excess mass, while maintaining acceptable Type II error in all of the densities considered in Table 1. The introduction of heavy tails and/or skew is problematic for both the Hall and York test and the Bayesian test, but much more so for the Hall and York test. Overall, we would recommend the Bayesian test for the assessment of modality.

There are a variety of other non-parametric tools that have been developed for examining modality. Hall and York (2001) and Cheng and Hall (1998) extend Silverman's test and the excess mass test, respectively, to the setting where the

null and alternative models have m and $m+1$ modes. Fraiman and Meloche (1999) estimate the number and location of modes using a kernel density estimate. The theory for these methods is complex and practical aspects of implementing these tools has not been fully explored.

We believe that non-parametric methods are valuable for exploratory analyses but that parametric mixture models have greater potential to shed light on biologically meaningful properties of multimodal distributions, such as the number and location of modes. Although frequentist methods are widely used for mixture models, the Bayesian approach presented here is more natural when the number of components is unknown. Besides quantifying the number and location of modes, the Bayesian approach could be extended to compare distributions. For example, a natural question to consider is whether the BFB and BPB distributions are "similar". Considering each as a mixture of normal distributions, the Bayesian approach would allow us to assess whether the distributions have the same number of components, and if so, whether the locations, spreads or weights might be identical. If the means and standard deviations of the normal components were similar for BFB and BPB but the mixture probabilities were very different then the overall distributions might appear markedly different when they only differ with respect to the weight given to the sub-groups of birds that comprise the various components. Such an analysis was carried out by Xu *et al.* (2010) for these data with interesting results.

Appendix: R Code to Implement the Bayesian Test for Multimodality

The R package `mixAK` (Komárek, 2009) contains a compiled R function to fit Richardson and Green's (1997) reversible jump model, `NMixMCMC`. Below, we provide a wrapper function to call `NMixMCMC` and extract the posterior number of modes and posterior number of components, as well as compute the BF for > 1 versus 1 mode. The wrapper takes every tenth MCMC iterate to reduce posterior autocorrelation, and otherwise uses the default prior specification with $K_{\max} = 20$ and the truncated Poisson with $\lambda = 1$ as described in Section 3. Note that `NMixMCMC` can incorporate other priors on k, as well as censored data, so the Bayesian approach can be generalized further.

```
 library(mixAK) # y is the data vector
# keep is the number of thinned MCMC iterates kept after burnin
modes=functio\mathcal{N}(y,keep){
 a=mi\mathcal{N}(y); b=max(y); ra=b-a; a=a-ra/4; b=b+ra/4
 gp=100; x=seq(a,b,(b-a)/(gp-1)); d=x
 d=rep(0,gp); modes=rep(0,keep)
 r=NMixMCMC(y,scale=list(shift=0,scale=1),
  prior=list(priorK="tpoisson",Kmax=20,lambda=1,xi=0.5*(a+b),ce=1/ra^2),
  nMCMC=c(burn=200,keep=keep,thin=10,info=100),keep.chains=TRUE)
 i=0
```

```
 for(j in 1:keep){
  kt=r$K[j]; w=r$w[(i+1):(i+kt)]; mu=r$mu[(i+1):(i+kt)]; si=sqrt(r$Sigma[(i+1):(i+kt)])
  for(k in 1:gp){d[k]=sum(w*dnorm(x[k],mu,si))}
  for(k in 3:gp){if(d[k-2]<d[k-1]){if(d[k-1]>d[k]){modes[j]=modes[j]+1}}}
  i=i+kt; if(modes[j]==0){modes[j]=1}
}
 ktot=max(r$K); mc=matrix(0,ncol=2,nrow=ktot)
 for(k in 1:ktot){mc[k,1]=sum(r$K==k)/keep; mc[k,2]=sum(modes==k)/keep}
 rownames(mc)=1:max(r$K); colnames(mc)=c("P(comps)","P(modes)")
 cat("BF for >1 mode=",(19/(sum(modes==1)/(keep-sum(modes==1)))),"\n")
 print("Posterior components/modes..."); mc
}
```

Here is output from the boreal prairie birds data:

```
> modes(bpb,2000)

Chain number 1
==============
MCMC sampling started on Sat Jul 06 14:13:02 2013.
Burn-in iteration 200
Iteration 2200
MCMC sampling finished on Sat Jul 06 14:13:08 2013.
BF for >1 mode= 73.00969
[1] ‘‘Posterior components/modes..."
  P(comps) P(modes)
1   0.1830   0.2065
2   0.7905   0.7925
3   0.0265   0.0010
```

Acknowledgements

We would like to thank Dr. M.-Y. Cheng for providing code to implement some of the analyses. Funding for this project was provided through NSF-HRD CREST # 0206200.

References

Allen, C. R., Garmenstani, A. S., Havlicek, T. D., Marquet, P. A., Peterson, G. D., Restrepo, C., Stow, C. A. and Weeks, B. E. (2006). Patterns in body mass distributions: sifting among alternative hypotheses. *Ecology Letters* **9**, 630-643.

Allen, C. R. and Holling, C. S. (2008). *Discontinuities in Ecosystems and Other Complex Systems*. Columbia University Press, New York.

Baştürk, N., Hoogerheide, L. F., de Knijff, P. and van Dij, H. K. (2012). A Bayesian test for multimodality with applications to DNA and economic data. Technical Report, Erasmus School of Economics, Rotterdam.

Brown, J. H., Marquet, P. A. and Taper, M. L. (1993). Evolution of body size: consequences of an energetic definition of fitness. *American Naturalist* **142**, 573-584.

Cheng, M. Y. and Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society, Series B* **60**, 579-589.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, New York.

Dazard, J. E. and Rao, J. S. (2010). Local sparse bump hunting. *Journal of Computational and Graphical Statistics* **19**, 900-929.

D'Onghia, G., Basanisi, M. and Tursi, A. (2000). Population structure, age, and growth of macrourid fish from the upper slope of the Eastern-Central Mediterranean. *Journal of Fish Biology* **56**, 1217-1238.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 57-588.

Fraiman, R. and Meloche, J. (1999). Counting bumps. *Annals of the Institute of Statistical Mathematics* **51**, 541-569.

Givens, G. H. and Hoeting, J. A. (2005). *Computational Statistics.* Wiley, New York.

Grant, P. R. (1986). *Ecology and Evolution of Darwin's Finches.* Princeton University Press, Princeton, New Jersey.

Griffiths, D. (1986). Size-abundance relations in communities. *American Naturalist* **127**, 140-166.

Hall, P. and York, M. (2001). On the calibration of Silverman's test for multimodality. *Statistica Sinica* **11**, 515-536.

Hall, P., Minnotte M. C. and Zhang, C. M. (2004). Bump hunting with non-Gaussian kernels. *Annals of Statistics* **32**, 2124-2141.

Hartigan, J. A. and Hartigan, P. M. (1985). The DIP test of unimodality. *Annals of Statistics* **13**, 70-84.

Henderson, D. J., Parmeter, C. F. and Russell, R. R. (2008). Modes, weighted modes, and calibrated modes: evidence of clustering using modality tests. *Journal of Applied Econometrics* **23**, 607-638.

Ho, H. J., Pyne, S. and Lin, T. I. (2012). Maximum likelihood inference for mixtures of skew Student-t-normal distributions through practical EM-type algorithms. *Statistics and Computing* **22**, 287-299.

Holling, C. S. (1992). Cross-scale morphology, geometry, and dynamics of ecosystems. *Ecological Monographs* **62**, 447-502.

Hutchinson, G. E. and MacArthur, R. H. (1959). A theoretical ecological model of size distributions among species of animals. *American Naturalist* **93**, 117-125.

Ipiña, S. L. and Durand, A. I. (2000). A measure of sexual dimorphism in populations which are univariate normal mixtures. *Bulletin of Mathematical Biology* **62**, 925-941.

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

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

Komárek, A. (2009). A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Computational Statistics and Data Analysis* **53**, 3932-3947.

Lin, T. I., Lee, J. C. and Hsieh, W. J. (2007). Robust mixture modeling using the skew t distribution. *Statistics and Computing* **17**, 81-92.

Manly, B. F. J. (1996). Are there clumps in body-size distributions? *Ecology* **77**, 81-86.

May, R. M. (1986). The search for patterns in the balance of nature: advances and retreats. *Ecology* **67**, 1115-1126.

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

Müller, D. W. and Sawitski, G. (1991). Excess mass estimates and tests for modality. *Journal of the American Statistical Association* **86**, 738-746.

Oksanen, L., Fretwell, S. D. and Järvinen, O. (1979). Interspecific aggression and the limiting similarity of close competitors: the problem of size gaps in some community arrays. *American Naturalist* **114**, 117-129.

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.

Rüppell, O. and Heinze, J. (1999). Alternative reproductive tactics in females: the case of size polymorphism in winged ant queens. *Insectes Sociaux* **46**, 6-17.

Scheffer, M. and van Nes, E. H. (2006). Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 6230-6235.

Siemann, E. and Brown, J. H. (1999). Gaps in mammalian body size distributions reexamined. *Ecology* **80**, 2788-2792.

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B* **43**, 97-99.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stanley, S. M. (1973). An explanation for Cope's rule. *Evolution* **27**, 1-25.

Vollmer, S., Holzmann, H. and Schwaiger, F. (2013). Peaks vs components. *Review of Development Economics* **17**, 352-364.

Wilson, E. O. (1953). The origin and evolution of polymorphism in ants. *Quarterly Review of Biology* **28**, 136-156.

Wright, S. (1968). *Evolution and the Genetics of Populations*: *Genetic and Biometric Foundations.* University of Chicago Press, Chicago.

Xu, L. (2005). *Statistical Methods for Comparing Multimodal Distributions.* Ph.D. Thesis, University of New Mexico, Albuquerque, New Mexico.

Xu, L., Hanson, T., Bedrick, E. J. and Restrepo, C. (2010). Hypothesis tests on mixture model components with applications in ecology and agriculture. *Journal of Agricultural, Biological, and Environmental Statistics* **15**, 308-326.

Received February 22, 2013; accepted July 12, 2013.

Ling Xu
Department of Mathematics and Statistics
James Madison University
Harrisonburg, VA 22807, USA
xulx@jmu.edu

Edward J. Bedrick
Department of Mathematics and Statistics
University of New Mexico
Albuquerque, NM 87131, USA
EBedrick@salud.unm.edu

Timothy Hanson
Department of Statistics
University of South Carolina
Columbia, SC 29208, USA
hansont@stat.sc.edu

Carla Restrepo
Department of Biology
University of Puerto Rico
San Juan, PR 00931, USA
crestre@hpcf.upr.edu