Statistical Modelling 2022; **22(1–2)**: 33–45

Reflections on Murray Aitkin's contributions to nonparametric mixture models and Bayes factors

Alan Agresti[1], Francesco Bartolucci[2] and Antonietta Mira[3,4]
[1]University of Florida, Gainesville, Florida, USA
[2]University of Perugia, Perugia, Italy
[3]Università della Svizzera italiana, Lugano, Switzerland
[4]University of Insubria, Como, Italy

Abstract: We describe two interesting and innovative strands of Murray Aitkin's research publications, dealing with mixture models and with Bayesian inference. Of his considerable publications on mixture models, we focus on a nonparametric random effects approach in generalized linear mixed modelling, which has proven useful in a wide variety of applications. As an early proponent of ways of implementing the Bayesian paradigm, Aitkin proposed an alternative Bayes factor based on a posterior mean likelihood. We discuss these innovative approaches and some research lines motivated by them and also suggest future related methodological implementations.

Key words: Bayes factor, Bayesian inference, generalized linear mixed model, mean likelihood, mixture models, nonparametric random effects

1 Introduction

We are delighted to be invited to contribute to this issue in honour of Professor Murray Aitkin. In a journal devoted to issues of statistical modelling, relating to a statistical society in which he was instrumental in the early days by helping to make statisticians more fully aware of the capabilities of generalized linear models (e.g., through workshops and the influential text by Aitkin et al., 1989), the Editors have an excellent idea to devote an issue to his creative contributions as well as those of others whom he has influenced.

Murray Aitkin's research interests have always been quite broad, with specialties including Bayesian and likelihood theory, generalized linear models and some particular cases such as item response models, mixture models including latent class models and random effects models, statistical computing and neural network models. Anyone who has ever attended a statistics conference at which Professor Aitkin has

Address for correspondence: Alan Agresti, Department of Statistics, University of Florida, Gainesville, Florida 32611, USA.
E-mail: agresti@ufl.edu

10.1177/1471082X20981312

been present can attest to the insightful comments he invariably makes following a presentation that deals with any aspect of statistical science.

In our contribution, we focus on some of Aitkin's many research publications that deal with mixture models and with Bayesian inference. Some of his contributions relate to our own interests and have motivated our own research work. A considerable number of his publications over the years, dating back to about 1980 and continuing to the present, have focused on mixture models of various types. In particular, his published output includes some of the first articles dealing with random effects in generalized linear models (e.g., Aitkin et al., 1981a; Bock and Aitkin, 1981). In Section 2 of this article, we discuss his proposal of a nonparametric random effects approach in such models, illustrating with an example. Aitkin also has been a frequent contributor to the literature on Bayesian inference, starting with an influential discussion paper for the Royal Statistical Society (Aitkin, 1991). This important early contribution dealt with Bayes factors (BFs), and he proposed an alternative formulation based on the mean of the likelihood function with respect to the posterior rather than the prior distribution. In Section 3, we discuss this contribution as well as some literature that has dealt with it. In each section, we also suggest possible future research work and methodological implementations that are motivated by this discussion.

2 Contributions on nonparametric mixture modelling

Many of Murray Aitkin's research publications have dealt with mixture models of a wide variety of types. Such models include latent class models and other finite mixture models and the generalized linear model that includes random effects, that is, the generalized linear mixed model (GLMM). Here we focus on an innovative idea of his for using nonparametric structure instead of assuming normality for the random effects in GLMMs.

2.1 Nonparametric random effects in generalized linear mixed models

For clustered data such as with repeated measures or in a longitudinal study, let y_{it} denote observation t in cluster i and let $\boldsymbol{x}_{it}$ denote a column vector of explanatory variables associated with this observation. Here we consider a simple model with univariate random effect u_i, traditionally assumed to have a $N(0, \sigma^2)$ distribution. For $\mu_{it} = E(Y_{it} \mid u_i)$, a GLMM of simple random-intercept form is

$$g(\mu_{it}) = u_i + \boldsymbol{x}_{it}^T \boldsymbol{\beta}, \tag{2.1}$$

for link function $g(\cdot)$ and fixed-effect model parameters $\boldsymbol{\beta}$. The Aitkin et al. (1981a) discussion paper for the Royal Statistical Society and Bock and Aitkin (1981) are highly cited and among the first articles about GLMMs and particular special cases such as clustered data in educational research, latent class models and item response models. Their influence was enhanced by paying practical attention to implementable computations using an EM algorithm. Also very heavily cited from these early days is Anderson and Aitkin (1985), which dealt with interviewer variability and the use of unbalanced ANOVA methods with binary data when the number of interviews conducted varies by interviewer.

Since $\{u_i\}$ are unobserved in model (2.1), the normality assumption is not easily checked. A natural concern is the impact on the bias and efficiency of estimating $\boldsymbol{\beta}$ due to violating this assumption when the true random effects distribution is very different from normality. Aitkin (1996, 1999a) promoted an interesting nonparametric approach, using an unspecified finite mixture distribution for u_i. For this model, maximum likelihood fitting is computationally straightforward, not requiring numerical integration. Estimation of the unspecified distribution usually has relatively few mass points, even for very large samples, but this is not problematic when the mixing distribution is a nuisance parameter rather than of direct interest. Aitkin used this approach for a variety of applications, including dealing with measurement error in explanatory variables in generalized linear models (Aitkin and Rocci, 2002) and modelling longitudinal binary and count responses (Aitkin and Alfó, 1998; Alfó and Aitkin, 2006).

We refer to Table 1, from the 2018 General Social Survey (GSS) in the USA, for a simple example in which such an approach may be natural. Subjects were asked whether they supported legalized abortion in each of three situations. (We ignore here seeming contradictions in people's responses, such as some subjects supporting legalization for any reason but being opposed in a particular situation.) A cluster is a set of the three observations for a particular subject, with subjects classified by gender. With such a controversial issue, the population might be polarized, with some people likely to support legalization regardless of the context, and some likely to oppose it regardless of the context. A third group of subjects may have response dependent on the context. With y_{it} the response for subject i in situation t, consider the model

$$\text{logit}[P(Y_{it} = 1 \mid u_i)] = u_i + \beta_t + \gamma x_i, \tag{2.2}$$

where $x_i = 1$ for females and 0 for males, the situation effects $\{\beta_t\}$ satisfy a constraint such as $\beta_1 = 0$, and we could add additional categorical and/or quantitative explanatory variables as well as interaction terms. Because of the likely polarization, it would seem implausible to assume a $N(\mu, \sigma^2)$ distribution for the random intercept in this model, even conditional on values of other potential explanatory variables.

One of us (Agresti) used Aitkin's nonparametric approach with GLMMs for a variety of scenarios. For example, for a logit model for a vector of binary responses observed under multiple conditions, Agresti (1997) showed that a nonparametric treatment of the random effects vector implies marginally a multivariate loglinear model having quasi-symmetric structure for the cross-classification of responses at the various conditions. For binary responses in multi-centre data comparing two treatments, such as clinical trials, Agresti and Hartzel (2000) used the nonparametric GLMM approach with the logit and other link functions to describe the mean and variability of centre-specific effects such as log odds ratios and risk ratios. Hartzel et

Table 1 Support (1 = yes, 0 = no) for legalized abortion in three situations: (a) if the family has a very low income and cannot afford any more children, (b) when the woman is not married and does not want to marry the man, and (c) when the woman wants it for any reason

	Sequence of responses							
Gender	(1,1,1)	(1,1,0)	(0,1,1)	(0,1,0)	(1,0,1)	(1,0,0)	(0,0,1)	(0,0,0)
Male	283	27	12	7	14	26	29	262
Female	325	15	13	9	22	24	51	362

al. (2001b) employed this approach with ordinal responses, describing centre-specific log odds ratios that result from models that apply the logit link to cumulative response probabilities or to pairs of adjacent response probabilities. Hartzel et al. (2001a) used it in a more general context for GLMMs with multinomial (nominal and ordinal) responses and generalized Aitkin's EM algorithm. In all such cases, simulations suggested that the method performed well for estimating fixed effects even when the standard model with a normal random effect truly holds. However, the nonparametric approach did not estimate the mixture distribution or its variance well. This is not surprising, because the maximum likelihood estimate of the mixture distribution (with an unspecified number of mass points) is typically highly discrete, with relatively few mass points, even though it often facilitates a good approximation of the marginal likelihood.

Agresti et al. (2004) investigated the impact of misspecification of the random effects distribution, using models with various actual random effects structures. For instance, consider estimation of the mean log odds ratio for comparison of two groups in several 2×2 contingency tables. Conditional on a random effect u_i in table i, suppose (y_{i1}, y_{i2}) are independent binomials with log odds ratio $\beta + u_i$. That is, for all i, conditional on u_i, y_{ij} is bin(n_{ij}, μ_{ij}) where

$$\text{logit}(\mu_{i1}) = \alpha + (\beta + u_i)/2, \quad \text{logit}(\mu_{i2}) = \alpha - (\beta + u_i)/2, \tag{2.3}$$

and where $E(u_i) = 0$ and $\text{var}(u_i) = \sigma^2$. (In practice, α could also vary in i, but this analysis focused on estimation of β.) One possibility for a severely non-normal true random effects distribution is a two-point mixture, such as when we expect the population to be polarized on a controversial issue. In this case, the nonparametric approach was much more efficient in estimating β than the standard model assuming normality for the random effects. The improvement was more substantial as σ increases. At the same time, the nonparametric approach lost little efficiency when the normality assumption truly held. Both the normal and nonparametric approaches were adequate in terms of bias in estimating β. Similar results about efficiency improvement held for other models. For example, for a frailty model for survival that assumes a gamma distribution for a random effect, the nonparametric approach was much more efficient when the actual distribution was binary, again more so as the random-effects variability increases.

Table 2 Fitting of GLMM in R to Table 1, with nonparametric and normal treatment of random effect term

```
Abortion <- read.table("agresti_bartolucci_mira.dat",header=TRUE)
> head(Abortion, 3)
  person gender situation response
1      1      0         1        1
2      1      0         2        1
3      1      0         3        1
> library(npmlreg)

> fit.npml <- allvc(response ~ gender + factor(situation), random=~1|person,
+       random.distribution="np", family=binomial(link=logit), data=Abortion, k=3)
> summary(fit.npml)         # nonparametric fitting of GLMM with 3 mass points
                     Estimate Std. Error     t value
gender            -0.62866964  0.1296554  -4.8487716
factor(situation)2 -0.56856051  0.1603709  -3.5452841
factor(situation)3  0.15467540  0.1543484   1.0021189
MASS1             -4.33740718  0.2793745 -15.5254219
MASS2              0.05946193  0.1320789   0.4501999
MASS3              5.27859211  0.2837316  18.6041766
Mixture proportions:    MASS1      MASS2      MASS3
                   0.3852812  0.2072261  0.4074927
Random effect distribution - standard deviation:  4.284974
-2 log L:          3906.6     Convergence at iteration  140

> fit.normal <- allvc(response ~ gender + factor(situation), random=~1|person,
+      random.distribution="gq", family=binomial(link=logit), data=Abortion, k=100)
> summary(fit.normal)       # GLMM with normal random effect and 100 quadrature points
                    Estimate Std. Error   t value
(Intercept)        0.4230881  0.1321061  3.202639
gender            -0.7362493  0.1308990 -5.624561
factor(situation)2 -0.5549031  0.1582741 -3.505961
factor(situation)3  0.1591908  0.1565966  1.016566
z                  6.9636465  0.2565615 27.142214       # standard deviation estimate
-2 log L:          3916.5      Convergence at iteration  23
```

For years, one could fit many GLMMs with nonparametric random effects using a GLIM macro, and recently Einbeck et al. (2018) provided the R package `npmlreg` for fitting of such models. Table 2 shows edited output for nonparametric fitting of the simple GLMM (2.2) to Table 1, specifying three mass points to represent polarization with a middle group having opinion depending on the situation. At the estimated locations for the polarized mass points, which are quite extreme and about equally likely, for each gender the probabilities of supporting legalization are close to 0 for each situation or close to 1 for each. In this output, we regard standard errors and log-likelihood values informally, as the GSS uses sampling more complex than a simple random sample. It is informative to note, however, that when we instead assume a normal random effect, Gaussian quadrature with 100 quadrature points has $-2(\log L)$ value 9.9 higher and a large estimated standard deviation (7.0) for the random intercept.

Although simple, the nonparametric approach has disadvantages. These include often poor estimates of the variance component, standard asymptotic theory for model comparison not being appropriate (when we regard the number of mixture mass points as unknown), identifiability issues, and less adaptability to multivariate random effects modelling and multilevel modelling than using a multivariate normal random effect.

2.2 Further research and implementation of GLMMs with non-normal random effects

Aitkin (1999b) used the nonparametric random effects approach for meta-analysis of multi-centre trials. He used the logit model, focusing on variability in log odds ratios such as described above. In practice, it would often be desirable to use a log or an identity link function, in which case summary effects relate to ratios or differences of proportions, which are simpler to interpret and sometimes more relevant. Agresti and Hartzel (2000) used these links for multiple 2×2 tables but noted the structural problem in using a continuous random effects distribution in modelling a probability or its log. Estimates exist that deal with heterogeneity in meta-analyses outside the context of a logit model. For instance, DerSimonian and Laird (1986) weighted sample estimates inversely proportional to estimated variances. However, Wald-type methods for categorical data that use estimated variances often behave poorly, especially for applications in which the probability of the outcome of interest is close to zero for each group, in which case sometimes all members of a group make the same response. How successful might a nonparametric random effects approach be for obtaining estimates for those non-standard link models?

For GLMMs, some authors have suggested replacing a normal random effects distribution by a finite mixture of normals. For instance, Agresti et al. (2004) used the $\rho N(\mu_1, \sigma^2) + (1 - \rho)N(\mu_2, \sigma^2)$ distribution for a mixture parameter ρ in a GLMM with logit link, and Caffo et al. (2007) did this with the probit link. Komàrek and Lesaffre (2008) considered this more generally, estimating the parameters with a Bayesian approach and using a penalized approach to estimate the weights of the mixture components. Pan et al. (2020) used a penalized EM algorithm to fit the model and proposed a type of likelihood-ratio test to determine the number of components in the mixture. The normal mixture approach can accommodate a wide variety of shapes and includes the binary mixture distribution as the special case with $\sigma = 0$. However, simulations for a model such as the logistic for binary multi-centre clinical trials, suggested that this approach had results much like those of a single normal random effect. Future research could consider a variety of models to evaluate whether such a random effects distribution might often provide a useful compromise between the ordinary normal random effects distribution and the nonparametric approach while enabling greater flexibility, less potential for serious misspecification, and better characterization and estimation of random-effects variability. It would also be useful to have an `R` function that can employ a mixture normal random effect for any GLMM. The package `glmmAK` has some capability in this direction, following the Komàrek and Lesaffre (2008) approach.

A particularly important case to consider is when the variance of the random effects depends strongly on values of covariates, as this can result in substantial bias in estimating fixed effects in ordinary GLMMs (Heagerty and Zeger, 2000). Also, for a GLMM, how can one estimate an unspecified but continuous mixture distribution with more precision than the nonparametric random effects approach does with its relatively few mass points, and does this make any difference for the resulting inference? Can the nonparametric approach be used effectively with multiple random effects and possibly multilevel structure, especially when good estimates are needed of variance components are highly heterogeneous? A referee pointed out that the nonparametric random effects approach would be useful for classification purposes, because it implicitly produces posterior probabilities of class memberships.

The research results quoted above dealt with modest numbers of parameters, such as multi-centre trials for about 30 centres. In these days of big data, it would be of interest to study the effect of misspecification of the random effects distribution for a sparse asymptotic framework in which the number of parameters grows with the sample size or even exceeds it. For instance, under what conditions does one obtain consistency of estimation of an average treatment effect and of variance components? Can regularization methods, such as the lasso, apply directly to GLMMs with nonparametric random effect? Finally, for all types of applications, further attention could be paid to the disadvantages mentioned at the end of Section 2.1.

3 Contributions on Bayes factors

Murray Aitkin has a long record of developing ways to implement Bayesian inference. Early work focused on an alternative to the BF introduced in a highly cited discussion paper for the Royal Statistical Society (Aitkin, 1991). A more recent work focuses on fundamental issues such as ways of assessing the credibility of confidence intervals and prediction intervals (Aitkin and Liu, 2018). His publications include a book (Aitkin, 2010) that presents a unified Bayesian treatment of inference and model comparisons using simple diffuse prior specifications and that re-visits long-term interests of his, such as finite mixture models and variance component models. Here we focus on his innovative idea of an alternative type of BF and some related research results.

3.1 Bayes factor using posterior mean likelihood

One of Professor Aitkin's most relevant contributions related to Bayesian inference is the proposal of the *posterior* BF, initially formulated in Aitkin (1991). Suppose that, with reference to a set of data represented by the vector $\boldsymbol{y}$, we need to compare two models, say M_1 and M_2, characterized by the vectors of parameters denoted by $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. Let us denote the likelihood functions under these models by $L_j(\boldsymbol{\theta}_j)$ and the prior distributions by $\pi_j(\boldsymbol{\theta}_j)$, with $j = 1, 2$. The posterior BF for M_1 versus M_2 is defined as the ratio between the posterior means $\bar{L}_1^A$ and $\bar{L}_2^A$, which are defined as

$$\bar{L}_j^A = \int L_j(\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}_j, \quad j = 1, 2, \tag{3.1}$$

where $\pi_j(\boldsymbol{\theta}_j|\boldsymbol{y})$ denotes the posterior distribution of the parameters of model M_j. Clearly, the main difference between this definition of the BF and the conventional one (see, for instance, Kass and Raftery, 1995) is that in the latter the marginal likelihood $\bar{L}_j^B = \int L_j(\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j$ is used in place of $\bar{L}_j^A$. A critical review of the use of these marginal likelihoods based on the prior distribution and their ratio is provided in Aitkin (2010); see in particular Section 2.8. It is objected that the prior mean of the likelihood says nothing about the variability of the likelihood and that it does not correspond to the distribution of the observed data, because it refers to a two-stage sampling process in which parameter values and observable variables are sequentially drawn. Other difficulties arise with the use of improper priors and in connection with specific model choice and hypothesis testing problems. Among others, it is shown that with nested hypotheses of a certain type, the BF based on the prior marginal likelihoods corresponds to the posterior mean of the likelihood ratio, and as a consequence of this equivalent formulation it is clear that data are used twice, violating in this way the principle of 'temporal coherence'.

The posterior BF proposed by Aitkin (1991) is used in the usual way, so that model M_1 is preferred to M_2 when its value is larger than 1 and, in general, the larger is this value, the stronger is the evidence in favour of M_1. Examples are given of inferential tasks in which the posterior BF has certain advantages compared to the conventional BF and, in particular, it represents a reasonable solution to the Lindley paradox when diffuse priors are used to compare two hypotheses (Lindley, 1957). Aitkin (1991) illustrated the application of the posterior BF to more sophisticated problems, as when, in a normal regression model, the best set of covariates has to be selected. In this case the comparison is among an arbitrary number k of statistical models, denoted by $M_1, \ldots, M_k$, each one characterized by a different set of covariates and which are not necessarily nested. Another interesting application illustrated in Aitkin (1991) concerns the binomial sample size problem, where the aim is to draw conclusions on the size of a binomial distribution on the basis of a sample drawn from this distribution, considering the probability of success as a nuisance parameter.

Aitkin has also discussed the idea of relying on the posterior mean of the likelihood function for selecting a statistical model in other papers, among which it is worth recalling Aitkin (1997), Aitkin et al. (2014) and Aitkin et al. (2015). The latter, in particular, is focused on two very interesting applications, namely the use of the posterior BF for selecting the number of components in a finite mixture model of normal distributions and the number of classes in a latent class model. Aitkin et al. (2014) focuses on models for using networks to investigate social group structure. In it, groups are identified via latent classes, the number of which is selected using Bayesian methodology. These applications inspired methodological extensions, and this is a characteristic of Prof. Aitkin's research attitude. He proposed to rely on a more sophisticated criterion than merely taking the posterior mean of the likelihood under each model. In particular, for each model M_j, the likelihood function is transformed into the deviance $D_j(\boldsymbol{\theta}_j) = -2 \log L_j(\boldsymbol{\theta}_j)$. Then, M_1 is preferred to M_2 if the posterior distribution of the deviance under the first model is stochastically smaller than under the second model. Details on how to compute the deviance from the MCMC output are given, and Aitkin et al. (2014) observes that the deviance is unaffected by label-switching (an issue occurring in mixture as well as in latent class models), since it is a symmetric function of the class labels, and also invariant under their permutation. As a result, the comparison of the class models is shown to be equivalent to the comparison of the distributions of the class deviances. What is interesting in this article is that the Bayesian approach through comparison of posterior deviance distributions leaves us uncertain whether there are two or three classes. However additional information helps to identify the model with two classes as the best one, since the third class is unstable and effectively empty. The article concludes by observing that the proposed approach is computationally intensive for large networks and suggesting that, in this setting, variational Bayesian methods that approximate the full posterior distribution with simpler structure are more computationally effective, though the degree of agreement with the full analysis has not been clearly established, a statement which is still largely true.

In Aitkin et al. (2015) a certain advantage emerges of this selection criterion in comparison to popular alternatives, such as the Deviance Information Criterion (Spiegelhalter et al., 2002). This result relates to an observation, formulated by Aitkin (2001), about Bayesian inference for finite mixture models. In particular, evidence is shown of the sensitivity of the selected number of components to the assumed prior structure. This criticism also occurs in an interesting overview of different Bayesian analyses of the popular Galaxy data (Roeder, 1990), showing that very different conclusions have been reached about the number of mixture components suitable to properly model these data.

3.2 Further research and implementation of Bayesian inference

A standard technique to perform Bayesian inference is based on the use of Monte Carlo algorithms, in particular in the Markov chain version (MCMC); see Robert and Casella (2013). As is well known, for a certain model M_j, an algorithm of this type produces a sequence of realizations, $\boldsymbol{\theta}_j^{[s]}$, $s = 1, \ldots, S$, from the posterior distribution of the parameters given the observed data $\boldsymbol{y}$, where S denotes the number of simulation steps performed in the algorithm. It is then natural to obtain a simulated posterior mean of the likelihood, when this is analytically available in a closed-form expression, from the corresponding realizations $L_j^{[s]} = L_j(\boldsymbol{\theta}_j^{[s]})$, $s = 1, \ldots, S$, while the posterior mean defined in (3.1) may be reliably estimated as $\hat{\bar{L}}_j^A = \frac{1}{S} \sum_{s=1}^S L_j^{[s]}$.

The possibility of obtaining the posterior mean of the likelihood function by a simple average along the path of the Markov chain is an interesting aspect that, in our opinion, has not been sufficiently stressed and exploited. In particular, once an MCMC algorithm is available, obtaining the posterior BF requires one merely to write and run a small amount of extra code; compare this with the ordinary BF based on the ratio of marginal likelihoods, in which case more sophisticated methods are required (see, for instance, Chib, 1995). Moreover, a simple parallelization is possible, when computing the posterior BF, in the setting in which the same model is estimated on the basis of parallel chains run on the entire dataset.

As outlined in Section 3.1, Aitkin has also proposed a more sophisticated criterion than that simply based on computing the posterior mean of the likelihood, which relies on stochastic ordering. More precisely, on the basis of parallel MCMC chains run separately for each model, Aitkin et al. (2015) proposed a consensus criterion to compare the models based on computing

$$W_j^+ = \sum_{s=1}^{S} W_j^{[s]}, \quad j = 1, \ldots, k, \tag{3.2}$$

where $W_j^{[s]}$ is a indicator variable equal to 1 if $D_j^{[s]}$ is the smallest among the deviances $D_1^{[s]}, \ldots, D_k^{[s]}$ obtained at the *s*-step, which are defined as $D_j^{[s]} = -2 \log L_j^{[s]}$. The favourite model is the one with the largest value of W_j^+, which is the 'most often best' model across the draws. Even in this case, a parallelization may be simply implemented when different chains are used to estimate the same model on the overall set of data. The computation of $D_j^{[s]}$, $s = 1, \cdots, S$, is performed on each one of the $j = 1, \ldots, k$ parallel chains separately and does not require the chains to communicate among each other. The values $W_j^{[s]}$ and, consequently, W_j^+, are then computed all at once, at the end, upon post-processing the obtained values of $D_j^{[s]}$. Moreover, when the same model is estimated by separate chains run on different scores using non-overlapping subsets of the observed data, we propose to use an extension of this criterion. In particular, if we use *C* cores to estimate each of the *k* models—and on each one of the core a subset of the data is considered—we can obtain *C* consensus values $W_j^{+[c]}$, $c = 1, \ldots, C$, defined as in (3.2), by comparing, in a suitable way, the simulation results. Then these quantities are simply aggregated to obtain the overall consensus of a single model that directly compares with the original W_j^+, using suitable weights if the subsets of data have different size or importance.

A final point of interest concerns the computation of the posterior BF on the basis of the output of the Reversible Jump (RJ) algorithm (Green, 1995; Richardson and Green, 1997) that is used when (quoting Professor Peter Green), 'the number of things you do not know is one of the things you do not know'. In fact, while the conventional BF is obtained from this output on the basis of the number of times each single model has been visited by the Markov chain that moves between subspaces of different dimensions, the posterior BF is obtained by elaborating the subchains referred to each single model treated as if these chains were produced from separate MCMC algorithms. In other words, the computation is performed within each subspace. Moreover, Bartolucci et al. (2006) showed that the RJ output may be elaborated in a more efficient way in order to obtain the conventional BF on the basis of the acceptance probabilities between models, a quantity that is computed at each step. An open research question is then whether, also in a RJ framework, the number of visits to each model may be somehow exploited to estimate the posterior BF and, subsequently, if the technique of Bartolucci et al. (2006) may be exploited also in this case to improve this estimate. Indeed, a similar question concerns the use of a parallel MCMC algorithm to obtain the posterior BF, where techniques such as the one proposed in Meng and Wong (1996) could be profitably used to improve the precision of the estimate of the posterior BF; see also Mira and Nicholls (2004) and Bartolucci et al. (2018).

We conclude by noting, again, how versatile Aitkin's contributions are, ranging from likelihood to Bayesian approaches to statistical inference; from methodological to application driven papers, from computational to theoretical. This is not the right venue to discuss them, but his topics of application have been highly diverse, as illustrated by article titles such as 'Stillbirths among offspring of male radiation workers at Sellafield nuclear reprocessing plant' (a co-authored *Lancet* article in 1999) and 'Teaching styles and pupil progress' (Aitkin et al., 1981b). In these difficult days for our planet, we would very much value a modelling expert like Professor Aitkin wrangling 'coronavirus data' and making reliable predictions on where we are heading.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

References

Agresti A (1997) A model for repeated measurements of a multivariate binary response. *Journal of the American Statistical Association*, **92**, 315–21.

Agresti A, Caffo B and Ohman-Strickland P (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, **47**, 639–53.

Agresti A and Hartzel J (2000) Tutorial in biostatistics: Strategies for comparing treatments on a binary response with multi-center data. *Statistics in Medicine*, **19**, 1115–39.

Aitkin M (1991) Posterior Bayes factors. *Journal of the Royal Statistical Society, Series B*, **53**, 111–28.

——— (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–62.

——— (1997) The calibration of p-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, **7**, 253–61.

——— (1999a) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–28.

——— (1999b) Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, **18**, 2343–51.

——— (2001) Likelihood and Bayesian analysis of mixtures. *Statistical Modelling*, **1**, 287–304.

——— (2010) *Statistical Inference: An Integrated Bayesian/Likelihood Approach*. Boca Raton, FL: CRC Press.

Aitkin M and Alfo M (1998) Regression models for binary longitudinal responses. *Statistics and Computing*, **8**, 289–307.

Aitkin M, Anderson D, Francis BJ and Hinde J (1989). *Statistical Modelling in GLIM (Revised in 2005 for GLIM4 and in 2009 for R)*. Clarendon Press.

Aitkin M, Anderson D and Hinde J (1981a) Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society, Series A*, **144**, 419–61.

Aitkin M, Bennett SN and Hesketh J (1981b) Teaching styles and pupil progress: A re-analysis. *Journal of Educational Psychology*, **51**, 170–86.

Aitkin M and Liu C (2018) Confidence, credibility and prediction. *Metron*, **76**, 251–68.

Aitkin M and Rocci R (2002) A general maximum likelihood analysis of measurement error in generalized linear models. *Statistics and Computing*, **12**, 163–74.

Aitkin M, Vu D and Francis B (2014). Statistical modelling of the group structure of social networks. *Social Networks*, **38**, 74–87.

——— (2015) A new Bayesian approach for determining the number of components in a finite mixture. *Metron*, **73**, 155–76.

Alfo M and Aitkin M (2006). Variance component models for longitudinal count data with baseline information: Epilepsy data revisited. *Statistics and Computing*, **16**, 231–38.

Anderson DA and Aitkin M (1985) Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203–10.

Bartolucci F, Bacci S and Mira A (2018) On the role of latent variable models in the era of big data. *Statistics & Probability Letters*, **136**, 165–69.

Bartolucci F, Scaccia L and Mira A (2006) Efficient Bayes factor estimation from the reversible jump output. *Biometrika*, **93**, 41–52.

Bock RD and Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, **46**, 443–59.

Caffo B, An M-W and Rohde C (2007) Flexible random intercept models for binary outcomes using mixtures of normals. *Computational Statistics & Data Analysis*, **51**, 5220–35.

Chib S (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**, 1313–21.

DerSimonian R and Laird N (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177–88.

Einbeck J, Darnell R and Hinde J (2018). The npmlreg package. Nonparametric maximum likelihood estimation for random effect models, version 0.46-5.

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–32.

Hartzel J, Agresti A and Caffo B (2001a) Multinomial logit random effects models. *Statistical Modelling*, **1**, 81–102.

Hartzel J, Liu I-M and Agresti A (2001b) Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials. *Computational Statistics & Data Analysis*, **35**, 429–49.

Heagerty PJ and Zeger SJ (2000) Marginalized multilevel models and likelihood inference. *Statistical Science*, **15**, 1–19.

Kass RE and Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–95.

Komarek A and Lesaffre E (2008) Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution. *Computational Statistics & Data Analysis*, **52**, 3441–58.

Lindley DV (1957) A statistical paradox. *Biometrika*, **44**, 187–92.

Meng X-L and Wong WH (1996) Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, **6**, 831–60.

Mira A and Nicholls G (2004) Bridge estimation of the probability density at a point. *Statistica Sinica*, **14**, 603–12.

Pan L, Li Y, He K, Li Y and Li Y (2020) Generalized linear mixed models with Gaussian mixture random effects: Inference and application. *Journal of Multivariate Analysis*, **175**. URL https://doi.org/10.1016/j.jmva.2019.104555 (last accessed 8 January 2021).

Richardson S and Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B*, **59**, 731–92.

Robert C and Casella G (2013). *Monto Carlo Statistical Methods*. Berlin: Springer.

Roeder K (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**, 617–24.

Spiegelhalter DJ, Best NG, Carlin BP and Van Der Linde A (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.