
Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **64**(2), 2022, 313–337 doi: 10.1111/anzs.12370

Bayesian hierarchical mixture models for detecting non-normal clusters applied to noisy genomic and environmental datasets

Huizi Zhang[1,2], Ben Swallow[2] and Mayetri Gupta[2,*]

The University of Edinburgh, and the University of Glasgow

Summary

Clustering to find subgroups with common features is often a necessary first step in the statistical modelling and analysis of large and complex datasets. Although follow-up analyses often make use of complex statistical models that are appropriate for the specific application, most popular clustering approaches are either nonparametric, or based on Gaussian mixture models and their variants, often for reasons of computational efficiency. Certain characteristics in the data, such as the presence of outliers, or non-ellipsoidal cluster shapes, that are common in modern scientific datasets, often lead these methods to fail to detect the cluster components accurately. In this article, we present two efficient and robust Bayesian clustering approaches that seek to overcome these limitations—a model-based 'tight' clustering approach to cluster points in the presence of outliers, and a hierarchical Laplace mixture-based approach to cluster heavy-tailed and otherwise non-normal cluster components—and illustrate their power and accuracy in detecting meaningful clusters in datasets from genomics, imaging and the environmental sciences.

Key words: data augmentation; Gibbs sampling; latent variable models; Markov Chain Monte Carlo; non-Gaussian clusters; SNP genotyping.

1. Introduction

A primary goal of cluster analysis is to find homogeneous groups within a dataset such that observations in the same cluster are similar to each other while objects in distinct groups tend to be different (Everitt 1974). With the advent of large and complex datasets in modern scientific research, cluster analysis has become a necessary statistical tool for exploratory data analysis before further formal investigation. Mixture models are a commonly used parametric framework for model-based clustering, and a vast literature on these and their applications are available (e.g. see McLachlan & Basford 1988). A common assumption is that each cluster comes from the same type of distribution but with different parameters, and

*Author to whom correspondence should be addressed.
[1]School of Mathematics, University of Edinburgh, Peter Guthrie Tait Road Edinburgh, EH9 3FD, UK.
[2]School of Mathematics and Statistics, University of Glasgow, University Place Glasgow, G12 8SQ, UK.
e-mail: mayetri.gupta@glasgow.ac.uk

Acknowledgements. We thank the anonymous reviewers for providing us with insightful comments and suggestions. Opinions and attitudes expressed in this document, which are not explicitly designated as Journal policy, are those of the author and are *not* necessarily endorsed by the Journal, its editorial board, its publisher Wiley or by the Australian Statistical Publishing Association Inc.

the overall population probability density is a weighted sum of the individual component densities. Gaussian mixture models (GMMs) are a practical and attractive choice due to their relative tractability, in both a classical and Bayesian framework (Diebolt & Robert 1994; McLachlan & Peel 2000). However, with the advent of more computing power, distributions beyond GMMs have been increasingly applied due to their ability to model asymmetric distributions as well as those with heavy tails, as well as a variety of non-regular features (Lee & McLachlan 2013a,b; Forbes *et al.* 2019).

The formulation of a multivariate t distribution as a multivariate Gaussian scale-mixture model was utilised in extending a Normal mixture to a t-mixture that can allow for heavier-tailed distributions in clustering (Peel & McLachlan 2000; Lin, Lee & Ni 2004; Andrews & McNicholas 2012; Lee & McLachlan 2019). Choy & Chan (2008) explored a wide class of distributions, the generalised t family of distributions, which includes the normal, t, and exponential power distributions, and illustrated a unified scale-mixture representation for this class that allowed Bayesian computational methods to be implemented easily for statistical inference. A normal scale-mixture representation can encompass a wide variety of distributional features and provide possible candidates for mixture modelling of components with lighter or heavier tails than the normal, as well as varying degrees of skewness (Johnson, Kotz & Balakrishnan 1994; Eltoft, Kim & Lee 2006a,b). For instance, the density functions of Pearson Type VII distributions (of which the t- and Cauchy distributions are special cases) can be represented as normal scale mixtures with the mixing density being a gamma (Johnson, Kotz & Balakrishnan 1994); the class of variance gamma distributions (Madan & Seneta 1990)—of which the Laplace distribution is a special case—can be represented as a Normal scale mixture with an inverse-gamma mixing density (Choy & Chan 2008); and generalised hyperbolic distributions may be represented using a generalised inverse Gaussian mixing distribution (Browne & McNicholas 2015). Skewed extensions to the elliptical distributions, such as the multivariate skewed Normal (MSN) and skew-t (MST) distributions, have also been proposed (Azzalini 2005; Lee & McLachlan 2016); and used successfully for model-based clustering in a classical (Lee & McLachlan 2013a,b) as well as Bayesian (Fruhwirth-Schnatter & Pyne 2010) framework. Lee & McLachlan (2013b) provided characterisations of several closely related parametric families of skew distributions that can be classified into four forms and illustrate their uses in clustering applications.

Distributions more widely divergent in properties from normality have also been derived starting from a Gaussian framework—an example being the geometric skew normal (GSN) distribution (Kundu 2017), based on an infinite convolution of Normal and Geometric densities. The GSN distribution can allow for skewness, heavy-tails and multimodality, and a latent variable-based Bayesian formulation was used efficiently for model-based clustering (Redivo, Nguyen & Gupta 2020). The Laplace distribution, with unique, and widely differing features to normality, has been used as a prior in hierarchical Bayesian models, in settings that favour sparsity, such as variable selection in regression model selection (Park & Casella 2008). It also can be represented as a normal scale mixture, and has the property of stability with respect to geometric summation (Kotz, Kozubowski & Podgorski 2001), analogous to the infinite divisibility property of the normal distribution under ordinary summation. In this article, we propose a new multivariate Laplace mixture-based model for clustering, which derives its motivation from the Bayesian LASSO (Park & Casella 2008), with a focus on clustering instead of variable selection. Particular forms of mix-

tures based on Laplace distributions have been previously used for clustering asymmetric distributions (Franczak, Browne & McNicholas 2014); our hierarchical approach is in some ways simpler, can be applied via minor modifications to the MCMC procedure for fitting Gaussian mixtures, and is simultaneously robust to deviations from normality in the mixture components.

More flexibility of modelling can be achieved without the restriction that subpopulations must originate from the same parametric family (e.g. Jones & McLachlan 1990). A practical problem in clustering data from biological experiments, is that of experimental noise and artefacts, which cause the cluster components to not conform exactly to symmetric, normal patterns (Ester *et al.* 1996)—patterns of noise and outliers are not always clearly known, and may not be detectable using ad-hoc approaches for data filtering and cleaning. With high volumes of complex data, clustering algorithms characterised by an intention to cluster all observations often result either in large clusters with amorphous patterns or a massive number of small clusters, neither of which provide useful information about the inherent structure of the data or resolve scientific questions. In gene expression microarray experiments, many genes are expected to be irrelevant to the biological process under investigation, prompting Tseng & Wong (2005) to propose 'tight clustering', a method producing tight and stable clusters through sequentially applying resampling procedures to clustering outcomes. Bensmail & Meulman (2003) proposed a Gaussian clustering approach, following Banfield & Raftery (1993), that allowed for random noise as a separate cluster, along with various specifications of the form of the Gaussian covariance matrix. Joo, Casella & Hobert (2010) developed a Bayesian approach for time course gene expression data. However, their definition of tight clusters, and the resulting approach, were substantially different. In this article, we also propose a Bayesian model-based tight clustering approach for datasets of large volume with a general model specification, specifying a prior structure that avoids non-identifiability issues that may be caused by empty clusters. Our approach differs from Bensmail & Meulman (2003) in avoiding the usage of different forms of the spectral decomposition of covariance matrices, in the hierarchical prior structure, and the model selection criteria used to select the number of clusters, in place of the Bayes Factor approximations based on integrated likelihoods.

In the following sections, we investigate two parallel Bayesian approaches towards clustering data displaying non-normality in their component densities—the first, a Bayesian model-based tight clustering approach that augments the normal mixture by separate noise distributions; and the second, a scale mixtures of normal distributions approach that increases robustness of the model fit in the presence of noise and outliers. In Section 2, we describe the model setup and methodology for the two proposed approaches, drawing connections to a unified Bayesian framework that encompasses several alternative mixture models, and that can be used for model-based clustering. Section 3 illustrates the applications of these methods, along with a comparison to the performance of other clustering methods on real datasets from single nucleotide polymorphism (SNP) genotype assays, North Sea fisheries catch data and 3D stereoscopic audio–visual recording data. Section 4 further explores the use of these methods in a variety of simulation studies, comparing them to a range of model-based clustering approaches, to investigate their relative performance and also the characteristics of identified clusters. Section 5 summarises our findings and discusses limitations and extensions of the proposed models and methods. Appendices to the main text are provided in the Supplementary Material.

2. Model framework and methodology

In a mixture model framework, independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, each of dimension p, within the population of interest, can be assumed to come from one of K different groups, with probabilities $\pi_1, \ldots, \pi_K$, ($\pi_k > 0$, for $k = 1, \ldots, K; \sum_{k=1}^{K} \pi_k = 1$). The pdf or pmf of $\boldsymbol{X}_i$ ($i = 1, \ldots, n$), evaluated at a realised value $\boldsymbol{x}_i$, is given by

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k), \tag{1}$$

where $f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ is the component-specific probability density or mass function for the kth mixture component, with the set of unknown component-specific parameter vectors being $\boldsymbol{\theta}^\top = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_K^\top)$. The f_ks are usually taken as densities from the same family with different component-specific parameters, but could also refer to different parametric families. The observed data likelihood arising from (1) is intractable, and numerical methods such as the EM algorithm (Dempster, Laird & Rubin 1977; McLachlan & Peel 2000) must be used for parameter estimation. In the Bayesian framework, a straightforward procedure for obtaining the posterior distributions for parameters may be achieved by means of data augmentation. Throughout this paper, we define a set of latent indicator variables Z_{ik} ($i = 1, \ldots, n; k = 1, \ldots, K$), where Z_{ik} takes the value 1 if observation i belongs to component k, and is zero otherwise. Then, with $\boldsymbol{Z}_i$ denoting the vector $(Z_{i1}, \ldots, Z_{iK})$, it can be seen that $\boldsymbol{Z}_i \sim \text{Multinomial}(1, \boldsymbol{\pi})$, where $\boldsymbol{\pi}^\top = (\pi_1, \ldots, \pi_K)$. The complete data likelihood can then be written in a simplified form, as

$$f(\boldsymbol{x}|\boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k)\right]^{Z_{ik}}. \tag{2}$$

The form of (2), taken along with appropriate (conjugate) priors for the parameters in $\boldsymbol{\theta}$, allows formulating an MCMC procedure to iteratively sample the component parameters and latent variables, through a Gibbs, or hybrid Metropolis–Gibbs sampler (Marin & Robert 2007). A Dirichlet prior, Dir($\boldsymbol{\alpha}$), where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, is typically used for $\boldsymbol{\pi}$.

2.1. Bayesian tight clustering approach for noisy data

Tight clustering (Tseng & Wong 2005) was a heuristic, resampling-based approach originally developed for finding compact data clusters in microarray experiments, without making any probabilistic model assumptions. It consisted of a sequential procedure with an inner loop (an algorithm such as k-means) that searched for tight cluster 'candidates' for a given K (number of clusters); and an outer loop that identified tight clusters sequentially based on successively increasing K. Augmenting the mixture model in (1) with a non-normal noise distribution (Dasgupta & Raftery 1998) can be considered an analogous approach in the Bayesian framework. For a set of p-dimensional observations $\boldsymbol{x}^\top = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, each generated from a mixture of K distributions, assuming an uniform distribution representing noise—the (incomplete) data likelihood takes the form

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{\pi_0}{V} + \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i|\boldsymbol{\theta}_k) \right), \tag{3}$$

where $\sum_{k=0}^{K} \pi_k = 1$, and now $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_K)$. V is the hypervolume of the data domain, where the noise is assumed to be randomly distributed. In practice, the hypervolume V is defined as

$$V = \prod_{j=1}^{p} \left(\max_{i \in \{1,\ldots,n\}} \{x_{ij}\} - \min_{i \in \{1,\ldots,n\}} \{x_{ij}\} \right),$$

and can be considered the volume of the data region (Dasgupta & Raftery 1998). Assuming a Dir$(\alpha_0, \alpha_1, \ldots, \alpha_K)$ prior for $\boldsymbol{\pi}$, its posterior full conditional distribution is Dir$(n_0 + \alpha_0, n_1 + \alpha_1, \ldots, n_K + \alpha_K)$, where $n_j = \sum_{i=1}^{n} Z_{ij}$ $(j = 0, 1, \ldots, K)$, which would constitute one step of the full Gibbs sampler. Most commonly, the component densities are assumed to be p-dimensional Gaussian, $\mathrm{N}_p(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ $(k = 1, \ldots, K)$, where we denote $\boldsymbol{\mu}^\top = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_K^\top)$, and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$. A hierarchical formulation of the prior for component-specific means can be used for mathematical tractability and faster MCMC convergence, with

$$\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k \sim \mathrm{N}_p(\boldsymbol{m}_0, \boldsymbol{\Sigma}_k / c_k); \; \boldsymbol{\Sigma}_k \sim \text{Inv-Wishart}(v_0, \boldsymbol{S}_0),$$

where the prior hyperparameters are chosen to be minimally informative while ensuring propriety of the distributions. In our examples, $\boldsymbol{m}_0$ was chosen to be a zero vector, $c_k = 10^{-3}$, $v_0 = p + 1$, and $\boldsymbol{S}_0$ a $p \times p$ identity matrix, and sensitivity analyses indicated that small variations from these settings led to no appreciable differences in posterior inference.

In the Bayesian framework, widely varying choices of f_k can be used, while still giving straightforward model fitting procedures through Gibbs sampling (occasionally including a Metropolis step), when usual conjugate prior specifications are used. These may range from discrete data models such as the Binomial and Poisson, to skewed continuous distributions such as the gamma (Wiper, Insua & Ruggeri 2001).

2.2. Clustering with Gaussian scale mixtures

Many symmetric continuous distributions can be represented as a ratio of a standard normal distribution Z and a positive random variable τ^2, independent of Z (Andrews & Mallows 1974). The probability density function of X, which is a continuous random variable with mean μ and scale parameter τ^2, has a Gaussian scale-mixture (GSM) representation if it can be written as

$$f(x|\mu, \sigma) = \int_0^\infty \phi(x|\mu, \sigma^2 g(\tau^2)) \pi(\tau^2) d\tau^2,$$

where $\phi(\cdot)$ is the Gaussian density function, $g(\cdot)$ is a positive function and $\pi(\cdot)$ is a density function defined on $\mathbb{R}^+$. With $g(\tau^2) = \tau^2$, and $\pi(\tau^2)$ set as an inverse-gamma density with parameters $(v/2, \delta/2)$, the pdf of X takes the form of a Pearson Type VII density (Johnson, Kotz & Balakrishnan 1994) with parameters (μ, σ, v, δ). When $v = \delta$, the pdf reduces to a Student t distribution with location μ, scale σ and degrees of freedom v. When $\pi(\tau^2)$ is, instead, a gamma $(\frac{v}{2}, \frac{v}{2})$ density, the resulting pdf of X is a symmetric variance-gamma distribution (Madan & Seneta 1990), which reduces to a Laplace density when $v = 2$. Logistic and generalised hyperbolic distributions can also be derived as normal scale mixtures through appropriate choices of the mixing distribution $\pi(\cdot)$ (Choy & Chan 2008).

In a multivariate setting, a number of forms of the t-distribution exist (Kotz & Nadarajah 2004), but the most commonly used one can be represented as a scale mixture of a multivariate

normal and an inverse-gamma distribution (Forbes & Wraith 2014). Lin, Lee & Ni (2004) presented an extensive exploration of multivariate t-mixtures for Bayesian model-based clustering, showing that the scale-mixture form is highly amenable to model fitting through a hierarchical Bayesian framework. For a single observation $\boldsymbol{X}_i$ ($i = 1, \ldots, n$), their model, conditioned on its cluster membership vector $\boldsymbol{Z}_{ik}$ (defined at the beginning of Section 2), for $k = 1, \ldots, K$, can be written hierarchically as

$$\begin{aligned} \boldsymbol{X}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\tau}, Z_{ik} = 1 &\sim \mathrm{N}_p(\boldsymbol{\mu}_k, \tau_i^2 \boldsymbol{\Sigma}_k), \\ \tau_i^2 &\sim \mathrm{IG}\left(\frac{v_k}{2}, \frac{v_k}{2}\right), \end{aligned} \tag{4}$$

where $\boldsymbol{\tau}^\top = (\tau_1, \ldots, \tau_n)$ is a vector of auxiliary scale variables; $\boldsymbol{\mu}^\top = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_K^\top)$, a set of K p-dimensional component mean vectors; $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, a set of K $p \times p$ covariance matrices, and $\boldsymbol{v}^\top = (v_1, \ldots, v_K)$, a K-dimensional vector of the degrees of freedom for the K components. Using appropriate conjugate priors for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, this model can be fit easily using a Gibbs–Metropolis procedure. Introducing multidimensional scale variables in a setup similar to (4) led to a new family of heavy-tailed distributions (Forbes & Wraith 2014), allowing for robust clustering on complex real-life datasets.

The Laplace distribution has strikingly different features from a Gaussian or t, notably in its sharp peak at the mode and heavy tails, motivating its use as a prior in hierarchical Bayesian models, in settings that favour sparsity. The Bayesian LASSO (Park & Casella 2008) used a model with a conditional Laplace prior specification for the regression coefficients with a non-informative scale-invariant prior on the error variance, which allowed for a Gibbs sampling-based model fitting approach. Kyung *et al.* (2010) presented a hierarchical group LASSO approach, in which the conditional prior on the regression coefficients appeared as a multivariate version of the Bayesian LASSO prior, which can be expressed as a gamma mixture of normals. As both Park & Casella (2008) and Kyung *et al.* (2010) showed, the conditioning on the variance parameter is necessary to attain a unimodal full posterior and efficient convergence of the Gibbs sampler. Our hierarchical M-Laplace approach, based on the ideas of Park & Casella (2008), gives a straightforward procedure to fit a robust mixture model for heavy-tailed data, through minor modifications to the standard MCMC procedures for fitting Gaussian mixtures. Our model specification differs significantly from the shifted asymmetric version of Laplace mixtures (Franczak, Browne & McNicholas 2014), and is described in the next section.

The scale-mixture representation is highly amenable to a Bayesian inferential approach, as the hierarchical structure, coupled with the existence of conjugate priors for most parameters, allows for the straightforward computation of the posterior full conditional densities (corresponding to well-known distributions) that are required for a Gibbs sampler. Careful choice of the prior structure also allows for minimising correlation in the joint posterior distributions of parameters, speeding up MCMC convergence. In mixtures in particular, there is much evidence to show that MCMC approaches are often more successful compared to maximum likelihood approaches (such as EM) in avoiding local maxima (Rydén 2008). The Bayesian approach also allows us to approximate the full joint posterior distribution of the parameters, even in complex multimodal likelihoods, where the choice of appropriate (informative) priors can regularise the posterior distributions when the likelihood is non-identifiable; or direct the sampler towards the regions of highest posterior density, permitting valid inference.

2.2.1. The M-Laplace mixture model

We now develop a model and methodology for a Bayesian Laplace mixture model-based clustering approach. If $X_i|Z_{ik} = 1$ has a Laplace distribution with parameters μ_k and σ_k/λ, then this model may be written as a Bayesian hierarchical model (Park & Casella 2008):

$$X_i|Z_{ik} = 1, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \tau_i^2 \sim \mathrm{N}\left(\mu_k, \sigma_k^2\tau_i^2\right) \quad (i = 1, \ldots, n)$$
$$\tau_i^2 \sim \mathrm{Ga}\left(1, \frac{\lambda^2}{2}\right) \equiv \mathrm{Expo}(\lambda^2/2).$$

A multivariate version of this model may also be developed hierarchically, with $\boldsymbol{X}_i$ being a p-variate normally distributed vector, and a scale factor of an (exponentially distributed) τ_i^2 for its covariance matrix (Eltoft, Kim & Lee 2006a,b).

Here, adapting ideas from the hierarchical group LASSO approach, we introduce an alternate, simplified version of a multivariate Laplace distribution by using a different, gamma-distributed scale factor instead of an exponential one. Unlike the Bayesian group LASSO model, however, which considers groups of varying numbers of regression coefficients, we are interested in a multivariate vector of observations that could arise from one out of K possible distributions that are a simplified version of a multivariate Laplace (or M-Laplace) model. We assume that we have a set of p-variate observations $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$, and the hierarchical model setup is as follows:

$$\begin{aligned} \boldsymbol{X}_i|Z_{ik} = 1, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau_i^2 &\sim \mathrm{N}_p\left(\boldsymbol{\mu}_k, \tau_i^2\boldsymbol{\Sigma}_k\right), \\ \tau_i^2|Z_{ik} = 1, \lambda_k^2 &\sim \mathrm{Ga}\left(\frac{p+1}{2}, \frac{\lambda_k^2}{2}\right), \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}_k$ is a p-dimensional mean vector, and $\boldsymbol{\Sigma}_k$ denotes a non-singular covariance matrix of order p ($k = 1, \ldots, K$). An observation $\boldsymbol{X}_i$ is assumed to be generated from the mixture component k with probability π_k ($k = 1, \ldots, K$). We use the notation $\boldsymbol{\tau}^2$ to denote the set of variables $(\tau_1^2, \ldots, \tau_n^2)$, $\boldsymbol{\lambda}$ to denote $(\lambda_1, \ldots, \lambda_K)$, $\boldsymbol{\mu}$ to denote $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, and $\boldsymbol{\Sigma}$ to denote $(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$. This leads to Proposition 1.

Proposition 1. *Under the hierarchical model setting* (5)*, the distribution of* $\boldsymbol{X}_i|Z_{ik} = 1, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ *is M-Laplace, with parameters* $\boldsymbol{\mu}_k$ *and* $\lambda_k^{-2}\boldsymbol{\Sigma}_k$*, and a pdf given by*

$$p(\boldsymbol{X}_i|Z_{ik} = 1, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) \propto \lambda_k^p|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left[-\lambda_k\left\{(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}^{1/2}\right]. \tag{6}$$

The proof of Proposition 1 is straightforward, following similar lines as other Gaussian scale mixtures, and is given in Appendix I for completeness. The form of the M-Laplace distribution in (6) is somewhat different to multivariate versions of the Laplace distribution as presented elsewhere, for example, in Eltoft, Kim & Lee (2006b). This is due to the usage of the gamma mixing distribution for the scale parameter τ_i^2, instead of an exponential distribution, which leads to the more commonly used (but relatively complex) form. When $\boldsymbol{X}_i$s can be assumed to have a diagonal covariance matrix, that is, $\boldsymbol{\Sigma}_k = \sigma_k^2\mathbf{I}_p$ (where $\mathbf{I}_p$ is an identity matrix of order p), the pdf in (6) simplifies to

$$p(\boldsymbol{X}_i|Z_{ik} = 1, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \propto \left(\frac{\lambda_k}{\sigma_k}\right)^p \exp\left[-\frac{\lambda_k}{\sigma_k}\left\{(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\}^{1/2}\right].$$

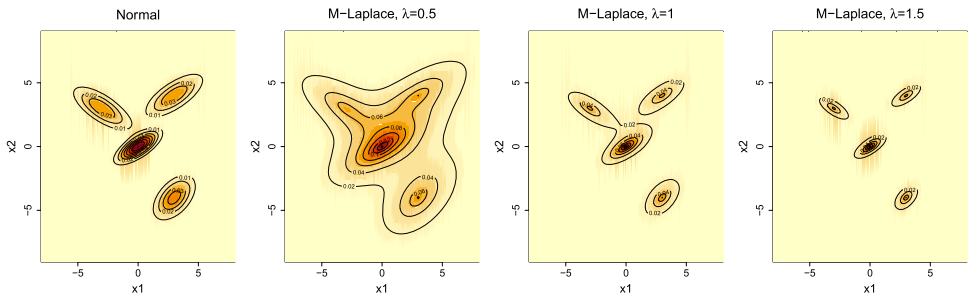

Figure 1. Mixture of 4 bivariate M-Laplace distributions with choices of λ as $0.5, 1, 1.5$, respectively, compared to a 4-component bivariate normal mixture. The means of the four components are: $(0,0)^\top, (-3,3)^\top, (3,4)^\top$ and $(3,-4)^\top$. Other parameters are specified in the text.

2.2.2. Features of the M-Laplace model

In order to motivate the use of the M-Laplace model for clustering, we briefly demonstrate some of its features compared to alternatives. In the univariate setting, the M-Laplace and multivariate Laplace both reduce to an identical form, with tails of the Laplace distribution being heavier than the Normal for every choice of scale parameter (Figure S8 in Appendix II). Comparing features of the bivariate forms of the Laplace distribution with the bivariate normal (Figure S9 in Appendix II), we see that all forms of the Laplace densities have heavier tails, with the M-Laplace with $\lambda = 1$ having the heaviest tails with the largest spread of sample values among them. This feature is replicated in the marginal histograms and densities of each variable (Figure S10 in Appendix II).

The properties of the univariate distributions are replicated in higher dimensions and in mixtures. A four-component bivariate mixture with unequal covariance matrices

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \text{ and } \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

is shown in Figure 1. With a smaller λ in the M-Laplace mixture, the component distributions have heavy tails—the component modes even appear to merge in the case of $\lambda = 0.5$—suggesting that this may be a more appropriate model to fit for data with more spread out distributions or outliers. This also suggests that the λ_ks can act as 'tuning' parameters, giving a hierarchy of clusterings of the same dataset that can be utilised through informative prior settings in a Bayesian framework.

2.2.3. Priors

Choosing conjugate model priors, if available, allows for efficient, closed-form Gibbs sampler steps for fitting the M-Laplace model. Following Park & Casella (2008), we choose, for $k = 1, \ldots K$, $\lambda_k^2 \sim \text{Ga}(r, \delta)$, $\boldsymbol{\mu}_k \sim \text{N}(\boldsymbol{m}_0, g\boldsymbol{\Sigma}_k)$, and $\boldsymbol{\Sigma}_k \sim \text{Inv-Wishart}(v_0, \boldsymbol{S}_0)$. As previously, we assume a Dirichlet–Multinomial model for $\boldsymbol{\pi}$ and $\boldsymbol{Z}|\boldsymbol{\pi}$. The priors may be made weakly informative by appropriate choices of the fixed hyperparameters $r, \delta, \boldsymbol{m}_0, g, v_0, \boldsymbol{S}_0$ and $\boldsymbol{\alpha}$ (while ensuring that $v_0 > p+1$ and $\boldsymbol{S}_0$ is positive definite). The fully conjugate prior specification for $\boldsymbol{\mu}_k$ leads to a more efficient Gibbs sampler with faster convergence. By using weakly informative priors for clustering, improper posterior distributions can be avoided

(for instance, allowing for empty clusters), while still allowing posterior inference to be data driven rather than strongly influenced by priors.

2.2.4. Bayesian clustering with M-Laplace mixtures

The hierarchical model formulation for the M-Laplace mixture, as discussed at the start of Section 2.2.1, allows for the posterior full conditional distributions for all but one of the parameters to be derived in closed form, which can then be sampled through efficient Gibbs sampling steps. In addition, the sampling steps are straightforward to implement as extensions to a standard normal mixture model-based Gibbs sampler. The forms of the posterior full conditional distributions are given in Proposition 2 and derived in Appendix I.

Proposition 2. *The posterior full conditional distributions of the parameters for the model described by* (5)*, with the prior settings as given in Section* 2.2.3*, are as follows*:

$$
\begin{aligned}
&\boldsymbol{\mu}_k|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\tau}^2,\boldsymbol{\Sigma}_k \sim \mathrm{N}_p\left(\boldsymbol{E}_k,\boldsymbol{V}_k\right), \quad \textit{where}\\
&\qquad \boldsymbol{E}_k = \left(\sum_{i=1}^{n}\frac{z_{ik}\boldsymbol{x}_i}{\tau_i^2}+\frac{\boldsymbol{m}_0}{g}\right)\Bigg/\left(\sum_{i=1}^{n}\frac{z_{ik}}{\tau_i^2}+\frac{1}{g}\right); \quad \boldsymbol{V}_k = \left(\sum_{i=1}^{n}\frac{z_{ik}}{\tau_i^2}+\frac{1}{g}\right)^{-1}\boldsymbol{\Sigma}_k;\\
&\boldsymbol{\Sigma}_k|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\tau}^2,\boldsymbol{\mu}_k \sim \text{ Inv-Wishart}\left(v_0+n_k,\boldsymbol{S}_k\right), \quad \textit{where}\\
&\qquad \boldsymbol{S}_k = \boldsymbol{S}_0+\sum_{i=1}^{n}\frac{z_{ik}}{\tau_i^2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top+\frac{1}{g}(\boldsymbol{\mu}_k-\boldsymbol{m}_0)(\boldsymbol{\mu}_k-\boldsymbol{m}_0)^\top,\\
&\qquad \lambda_k^2|\boldsymbol{Z},\boldsymbol{\tau}^2 \sim \text{ Ga}\left(\frac{(p+1)n_k}{2}+r,\frac{\sum_{i=1}^{n}\tau_i^2 z_{ik}}{2}+\delta\right), \textit{ and } \boldsymbol{\pi}|\boldsymbol{Z} \sim \text{ Dir}(\boldsymbol{n}+\boldsymbol{\alpha}),
\end{aligned}
\tag{7}
$$

where $n_k = \sum_{i=1}^{n} z_{ik}$ *and* $\boldsymbol{n} = (n_1,\ldots,n_K)$.

In addition, for the remaining latent variables, we have, for $i = 1,\ldots,n$:

$$
p(\tau_i^2|\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\sigma}^2) \propto (\tau_i)^{-1}\exp\left[-\frac{1}{2\tau_i^2}\sum_{k=1}^{K} z_{ik}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)-\frac{\tau_i^2}{2}\sum_{k=1}^{K}\lambda_k^2 z_{ik}\right]. \tag{8}
$$

One could, in principle, sample from the distribution in (8) using a Metropolis–Hastings-type step, but in practice, this appeared to significantly slow down convergence, so we used an adaptive Metropolis–Hastings (ARMS) step instead (Gilks, Best & Tan 1995). Finally, the latent cluster indicator variables $\boldsymbol{Z}$ may be sampled through evaluating the posterior cluster membership probabilities, using Bayes' Theorem; here we marginalised the densities over the latent variable τ_i, to increase the efficiency of the Gibbs sampler and improve convergence (Liu, Wong & Kong 1994).

2.3. Model selection and assessment in model-based clustering

To assess the performance of clustering methods in simulation studies and real datasets where the cluster labels were known (or identifiable), we used the adjusted rand index (ARI) (Hubert & Arabie 1985), that evaluates the accuracy of clustering results based on the prior knowledge of clusters. High agreement between two clusterings is indicated by ARI values close to 1, while values approaching 0 indicate poor concordance. Additionally, given the true labels, the correct classification rate (CCR)—defined as the proportion of correctly

allocated cluster points among all objects in clusters—was used to compare the performance of different methods when the true number of clusters was known.

The other aspect of assessing clustering performance was through model comparison, especially important in the context of real-life datasets with no established cluster membership. One criterion used was the Bayesian information criterion (BIC) (Schwarz 1978), that maximises the data likelihood while penalising model complexity. In Bayesian models, the BIC can be evaluated at the estimated posterior mode rather than the MLE, while another option is based on estimating the integrated likelihood from posterior simulation (Raftery *et al.* 2007). Bensmail & Meulman (2003) approximated the integrated Bayes Factor based on a Laplace–Metropolis approximation to decide on the best model after excluding points classified as noise; our selected approach for model selection was based on an alternative criterion, the WAIC (Watanabe 2010). The WAIC evaluates the predictive accuracy of the fitted model to the data using log pointwise predictive densities based on the posterior distribution of the parameters: a description is provided in Appendix I.

3. Applications

The Bayesian uniform-normal mixture (Section 2.1) and M-Laplace mixture model (Section 2.2.1) were fitted on a number of datasets, and compared with other commonly used mixture model-based clustering methods in the same scenarios. We present results from three applications: (i) a GWAS of bone density (Estrada *et al.* 2012), (ii) data from bottom-trawl fishing in Scotland and Northern Europe (ICES 2020) and (iii) audio–visual recording data from the CAVA database (Arnaud *et al.* 2008). In the first two cases, the cluster identities of samples were either fully known or had been inferred through manual identification techniques.

Along with the two proposed methods, we also applied the following—(i) tight clustering (TC)—implemented through the tightClust R package (Tseng & Wong 2018), (ii) normal mixture models and normal mixtures with uniform noise fitted with the EM algorithm (N-EM and UN-EM)—both implemented using versions of the R package mclust (Scrucca *et al.* 2016), (iii) Normal mixture models (N-GS) fitted via Gibbs sampling using the R package mixAK (Komarek 2009), (iv) the EM algorithm for mixtures of t-distributions implemented in the R package EMMIXskew (Wang, Ng & McLachlan 2009) and (iv) mixtures of shifted asymmetric Laplace distributions or MSAL implemented in the R package MixSAL (Franczak *et al.* 2018). Model selection was done using BIC for the non-Bayesian models, and using both BIC and WAIC for the Bayesian ones. All implementation of our proposed models was done in the R statistical software environment (R Core Team 2021).

3.1. Genotype identification

Single nucleotide polymorphisms (SNPs) are base-pair variations (A, C, G or T) at a locus in an individual's DNA sequence, which provide important markers in genetic association or linkage studies to locate genes that may be responsible for various diseases (Auton *et al.* 2015). In genome-wide association studies (GWAS), tens of thousands of SNPs are interrogated simultaneously in order to detect those that may be associated with a phenotype or disease. Most SNPs are biallelic, with only two possible nucleotides out of the 4 occurring, say A and C, giving three possible genotypes at that position as AA, AC or CC. The data from a genotyping assay are bivariate, representing the quantitative levels

of fluorescent intensities of two probes designed to capture each of the two alleles—the stronger the intensity, the stronger is the signal for that allele. Clustering of the data from multiple individuals is used to partition samples into the three possible genotypes, either with a proprietary clustering algorithm built into the genotyping platforms, such as Illumina (Zhao *et al.* 2018), or a (usually Gaussian) mixture model (Erickson & Callaway 2016). The typical shapes of the cluster components, volume of the data (both SNPs and numbers of individuals), along with experimental noise and instrumental limitations, complicate this process, leading to numerous errors, that then have to be corrected through a labour-intensive procedure of manual curation, magnified by the sheer order of repetition for the hundreds of thousands of SNPs. It is therefore important to be able to automatically and accurately cluster SNPs into the correct genotype categories using powerful probabilistic clustering methods.

3.1.1. Tight genotype clustering

We applied the Bayesian tight clustering method, as well as several other approaches on four SNP datasets, each from 5094 individuals in a GWAS of bone density (Estrada *et al.* 2012). The data are in fairly dense clusters, with a number of outliers, and some points in-between the main cluster centres (first column of Figure 2). For all SNP datasets, we obtained manually curated sets of genotype calls which could be used as a benchmark for the accuracy of predicted calls.

Comparing the classification of points to genotype groups, for the SNP tagged as `rs6665426` in dbSNP (Sherry *et al.* 2001), UN-GS was the only method that was successful in

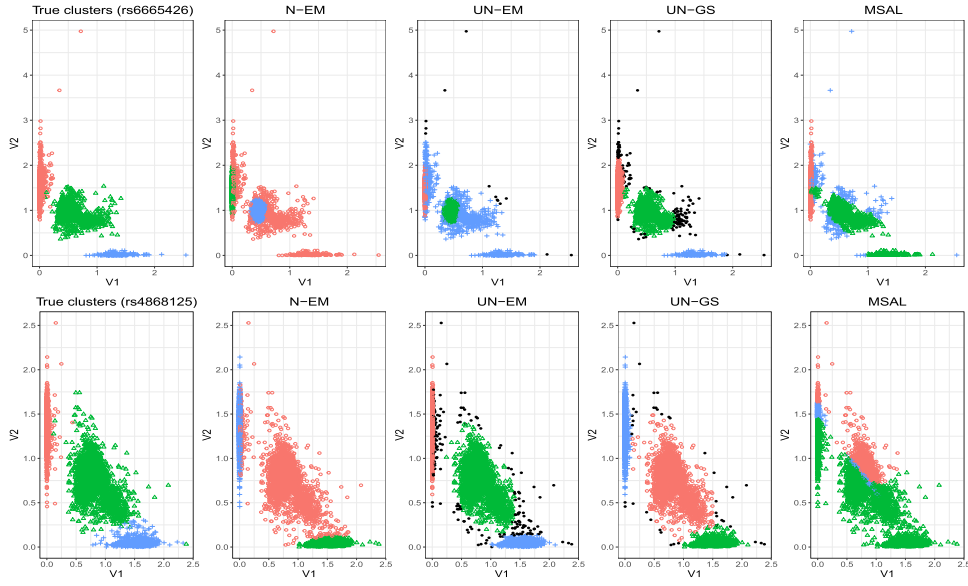

Figure 2. Comparisons of Single nucleotide polymorphism (SNP) genotype prediction for three methods: N-EM (column 2), UN-EM (column 3), UN-GS (column 4) and MSAL (column 5) compared to the gold standard of manual curation (column 1) for two SNPs: `rs6665426` and `rs4868125` (rows). The black points correspond to predicted 'noise' by UN-EM and UN-GS, that are not allocated to any of the three clusters.

Table 1. Adjusted rand index (ARI), correct classification rate (CCR), Bayesian information criterion (BIC) and predicted percentage of noise for the genotyping data for four SNPs, fitted using N-EM, UN-EM, UN-GS and MSAL.

SNP	Measure	N-EM	UN-EM	UN-GS	MSAL
rs2993122	ARI	0.5674	0.5605	0.7735	0.9070
	CCR	0.8386	0.8129	0.9163	0.9499
	Noise%	–	4.21	1.59	–
rs4868125	ARI	0.8187	0.9247	0.9715	0.0161
	CCR	0.9432	0.9495	0.9866	0.4402
	Noise%	–	4.91	0.83	–
rs6665426	ARI	0.6602	0.6310	0.9447	0.7234
	CCR	0.8229	0.8073	0.9699	0.8579
	Noise%	–	0.26	2.90	–
rs6683715	ARI	0.4305	0.6781	0.9355	0.8317
	CCR	0.6713	0.8600	0.9317	0.9107
	Noise%	–	4.57	2.22	–

determining the groups fairly accurately (Figure 2), although the number of points classified as noise was over-estimated. UN-EM classified fewer points as noise, but severely misclassified the groups, one severely non-normal group (denoted by + symbols) containing observations from all three genotypes. For the SNP rs4868125, UN-GS again more accurately classified SNPs into the correct groups, and classified fewer points as noise, compared to UN-EM. This behaviour was also observed in the CCR and ARI (Table 1)—for three of four SNPs, UN-GS produced the highest values among all methods, sometimes by as high a margin as 0.3 or 0.4. For the fourth SNP, rs2993122, MSAL appeared to have a slightly higher ARI and CCR, but on closer observation, it appeared that it still failed to detect the cluster separation accurately, and the lower values for UN-EM and UN-GS were due to a larger proportion of points in one cluster being labelled as noise. On average, UN-GS classified fewer points as noise compared to UN-EM. The BIC tended to overestimate the number of groups (typically by 1) for these four datasets. Running 1000 iterations of UN-GS on each SNP dataset took an average of 10.34 minutes on a 2 GHz Intel Core i5 processor.

3.1.2. SNP genotype clustering using M-Laplace mixture models

We next investigated the performance of the M-Laplace mixture model on SNP datasets (Estrada *et al.* 2012) which did not have clear outliers, but more spread out clusters, for which the component distributions were likely to have heavier tails than a Normal. We present here results for a particular SNP, rs1926463, that showed visual separation in its genotype distributions, yet could not be clustered accurately using the Illumina software as well as simple and widely used clustering methods such as k-means. We fitted the M-Laplace mixture model on this dataset, with K fixed at 3, and initially with weak priors set for $\boldsymbol{\mu}_k$, ($\boldsymbol{m}_0 = \mathbf{0}$ and $g = 10^6$), $\boldsymbol{\Sigma}_k$ (with $\nu_0 = 5$ and $\boldsymbol{S}_0$ a diagonal matrix of dimension 2) and $\boldsymbol{\pi}$ ($\alpha_j = 1$, for $j = 1, \ldots, K$). It was found that the posterior sampling step for λ_k^2 caused slow MCMC convergence of the algorithm and we therefore experimented with fixing the values of λ_k^2 and checking the sensitivity of the results to a range of values over 0.1 to 100 (discussed later). The MCMC algorithm, after 10,000 iterations, did not appear to show signs

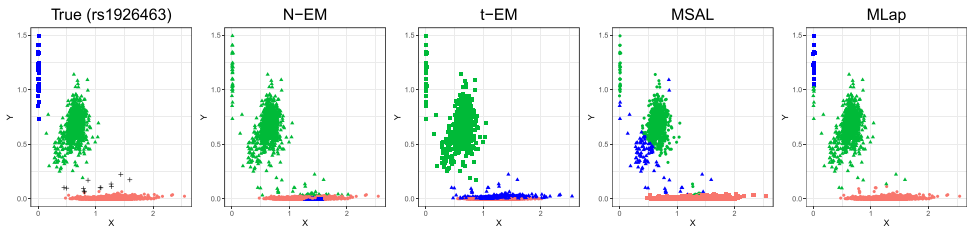

Figure 3. Comparisons of genotype cluster prediction for four methods: Normal mixtures (Panel 2), t-mixtures (Panel 3), shifted asymmetric Laplace mixtures (Panel 4) and M-Laplace mixture (Panel 5) compared to manual curation (Panel 1). The plot symbols indicate the cluster labels according to manual curation, while the point colours indicate the cluster labels predicted by each method.

of non-convergence, and the results here, are from a run when $\lambda_k^2 = 2$ with the first 20% of the chain taken as burn-in. Using the M-Laplace mixture model, 5052 out of the 5094 individual genotypes were correctly called, with two of the three categories being classified perfectly (CCR: 0.9918; ARI: 0.9887). Keeping other parts of the model unchanged, but a more informative prior on $\boldsymbol{\mu}_k$, based on clustering patterns in a number of other SNP datasets ($\boldsymbol{\mu}_1 = (1.5, 0)^\top$, $\boldsymbol{\mu}_2 = (0.7, 0.7)^\top$, $\boldsymbol{\mu}_3 = (0, 1.5)^\top$), the results were slightly improved further, with 5072 of the genotypes correctly classified (CCR: 0.9980; ARI: 0.9924). We also fitted a normal mixture (N-EM), a t-mixture (t-EM), and MSAL on the same dataset. These methods had lower accuracy, with N-EM having a CCR of 0.715 (ARI: 0.4007), t-EM a CCR of 0.6048 (ARI: 0.3534) and MSAL a CCR of 0.9748 (ARI:0.9681). Figure 3 shows the SNP data labelled by the predicted genotype IDs from the manual curation method and each of the fitted models. It is appears that while the M-Laplace picked up the separation of the clusters quite clearly, the other three algorithms did not. For N-EM and t-EM one cluster cannot be detected separately, while two other predicted clusters actually correspond to a single cluster. For MSAL, the middle cluster is split into two, while the third cluster is merged with one of these components. Posterior summaries for the cluster-specific parameters and convergence diagnostics (for the $\lambda_k^2 = 2$ setting) are shown in Table 2. We conducted a sensitivity analysis for λ_k^2 for the M-Laplace mixture; posterior credible intervals of $\boldsymbol{\mu}$ from different settings of λ_k^2 (Supplementary Figure S15 in Appendix II) showed very little variation, and high degrees of overlap, suggesting that the precise choice of λ_k^2 did not strongly matter. Credible intervals for the other parameters also remained similar with variations in λ_k^2, except when λ_k^2 was set at the highest value (100), in which case the clustering results of the M-Laplace method appeared to become similar to mclust (splitting the cluster along the horizontal axis into two clusters and merging the other two clusters into one), indicating a form of limiting behaviour.

The typical pattern and shapes of the three genotype clusters, along with the imbalance of points across clusters (the category with two rare alleles usually has a very small frequency), appeared to make it difficult for Gaussian or t-mixture based clustering methods to detect clusters accurately, even when there was clear visual separation. Our analysis was replicated on several other SNPs with similar genotype patterns, indicating that the M-Laplace mixture may have the potential to provide a robust and accurate method for automatic SNP genotype-calling in large-scale GWAS. MSAL allows for heavy tails and multimodality in the data, however, this approach often ran into numerical issues, when certain runs of the algorithm failed, and had to be re-started from a different initial value. An additional feature that was

Table 2. Credible intervals and effective sample size (ESS) for model parameters of M-Laplace mixture model. 2.5th and 97.5th percentiles of the posterior distributions of each parameter are shown.

Parameter	(2.5th, 97.5th)			ESS (mean)
$\boldsymbol{\mu}_1$	(1.300, 1.375)	(0.004, 0.005)		10304.7
$\boldsymbol{\mu}_2$	(0.588, 0.675)	(0.619, 0.711)		9933.3
$\boldsymbol{\mu}_3$	(−0.129, 0.299)	(0.850, 1.684)		5302.2
$\boldsymbol{\Sigma}_1$	(1.809, 2.113)	(0.005, 0.008)	(0.0003, 0.0004)	9445.9
$\boldsymbol{\Sigma}_2$	(0.348, 0.522)	(0.332, 0.503)	(0.392, 0.596)	9445.9
$\boldsymbol{\Sigma}_3$	(0.0297, 0.920)	(-0.381, 0.318)	(0.241, 3.028)	9445.0
π_1	(0.8353, 0.8553)			9100.8
π_2	(0.1399, 0.1612)			6853.5
π_3	(0.0005, 0.0072)			3050.4

observed even in successful runs of MSAL with relatively high CCR was that the correct 'patterns' or directions of clusters failed to be determined, for instance, clusters being split along incorrect axes or subdivided into two or more parts with wide spreads; see, for example, Figures 2 and 3. The results of different runs sometimes varied, suggesting that the EM algorithm could get trapped at suboptimal modes of the likelihood.

3.2. Fisheries data

Data from the North Sea International Bottom Trawl Survey (NS-IBTS) (ICES 2020), consisting of distributions of size structure of the caught fish, are important for understanding the impact of both natural and human pressures on fish populations (Weerarathne, Monk & Barrett 2021). We selected two species of fish commonly caught in the survey as our second case study: these are the European sprat *Sprattus sprattus* and Atlantic mackerel *Scomber scombrus*. The data, collected over the period 1990–1999, consist of two measurements—cost per unit effort (CPUE) per hour and length class (measured in 5 cm bands, with the measurement taken the lower bound of the band). While length class is a discrete measurement, it represents a continuous scale and our treatment of it as a continuous measurement seemed reasonable.

Each set of data, corresponding to a species, had a highly non-normal structure, and there was not a complete visual separation between groups, suggesting that clustering accurately could be difficult (Figure 4). Observations that were zero on both axes (no fish of that species and/or length class being observed) were non-informative in the context of clustering and were removed prior to fitting the algorithms, but measurements that were zero on a single axis were retained. It was evident that there were large numbers of outliers in the data, with a large spread in the upper tail. It is common to consider the CPUE on a log-scale to make distributions more spherical. When some observations have a measurement of zero on one of the axes, a standard approach is to add a small constant to all data points before taking logarithms. Bellégo, Benatia & Pape (2021) discuss possible pitfalls of this approach, and show that bias in log-linear model parameter estimates is minimised for a constant approximately equal to 0.7. Small values were likely to dominate the zero values, whereas larger values could potentially change the variance of the clusters and risk distortion of the results.

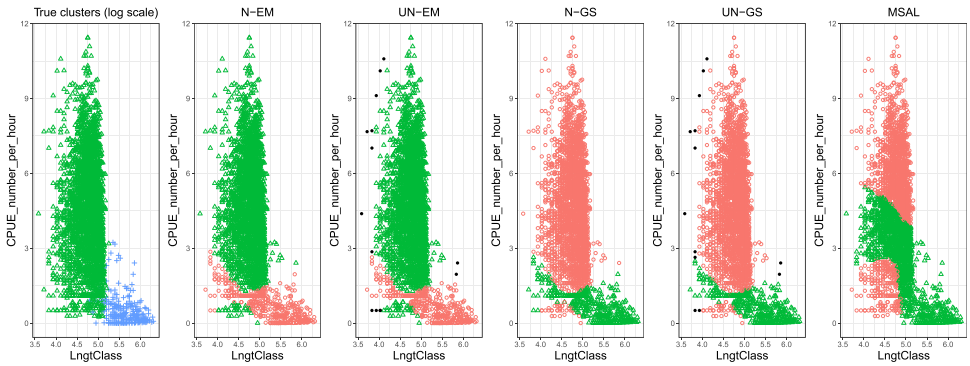

Figure 4. Comparisons of cluster prediction results for five methods applied on the Fisheries data on the log-scale, compared to actual clusters (leftmost panel). The black points correspond to predicted 'noise' by UN-EM and UN-GS, that are not allocated to any of the two clusters.

3.2.1. Tight clustering and comparison with other methods

The Fisheries data were filtered, by removing observations with zero in both measurements, incrementing all values by 1 and taking logarithms: the first column in Figure 4 shows the two clusters: *Sprattus sprattus* ($n = 2904$) and *Scomber scombrus* ($n = 256$). We implemented the N-EM, UN-EM, N-GS, UN-GS and MSAL methods on the log-transformed data. For UN-GS, we ran a chain of length 55000, after which a burn-in of 5000 and a thinning of 5 were applied. Generally, the results for all methods except for MSAL were quite similar (Figure 4), where the larger cluster appeared to be segmented into two by the smaller cluster. The noise points identified by UN-EM and UN-GS were also similar. From Table 3, it can be seen that UN-GS had the highest CCR (0.9386) and ARI (0.6348) among all the methods, followed by N-GS. The low ARI and CCR for MSAL are evident in Figure 4, where a large number of observations in the large (circles) group were wrongly allocated to the other. The BIC values of the corresponding models were calculated as 15220.51 (UN-GS), and 18178 (MSAL). Running 1000 iterations of UN-GS took approximately 5.38 minutes on a quad core 10th generation Intel Core i5 CPU processor with a speed of 2 GHz (Turbo Boost up to 3.8GHz).

Posterior credible intervals of the parameters from N-GS and UN-GS, showed considerable similarity to confidence intervals from N-EM and UN-EM, calculated using non-parametric bootstrap (Table S6 in Appendix II). To investigate the influence of the constant value added to the data before taking logarithms, we also applied the above five methods when a constant of 0.5 was used. The results were essentially unchanged (Figure 4). UN-GS retained the highest CCR and ARI, followed by N-GS. Such similar outcomes can be attributed to the cluster shape not changing substantially between a constant of 1 or 0.5.

Table 3. Adjusted rand index (ARI) and correct classification rate (CCR) for the Fisheries data, using the methods N-EM, UN-EM, N-GS, UN-GS and MSAL.

	Measure	N-EM	UN-EM	N-GS	UN-GS	MSAL
Logarithmic	ARI	0.6088	0.6177	0.6285	0.6348	0.0529
Scale	CCR	0.9355	0.9345	0.9383	0.9386	0.6196

However, with a constant of 0.01, the shape of the smaller (blue) group in the leftmost panel (Figure 4) was altered considerably, leading to a worse performance for all methods– this matched observations in Bellégo, Benatia & Pape (2021). Therefore, a choice of 1, or values around 1, that preserved the shape of the clusters, appeared suitable for this dataset.

Finally, as N-EM and UN-EM are both designed to cluster in the presence of noise, we did an additional test by adding uniform noise points in the data space before clustering. In this case, UN-GS gave the most accurate results (CCR: 0.9259, ARI: 0.6584), followed by UN-EM (CCR: 0.9241, ARI: 0.6537). N-GS performed slightly worse (CCR: 0.9127, ARI: 0.6264). In all cases, the large cluster was still split by the smaller one (Figure S16 in Appendix II). For N-EM (CCR: 0.7472, ARI: 0.2760), many more observations in the large cluster were incorrectly classified and more noise points were allocated into the clusters. The Fisheries dataset was difficult to cluster because of zero-inflated data; a more sophisticated mechanism for dealing with the zeroes may succeed in improving the clustering results. In this dataset, when the total number of clusters was allowed to vary from 2, the BIC overfitted every type of model, by favouring more clusters in the data.

3.3. 3-D data from stereoscopic camera pair

As our final application, we compared the performance of the previously discussed methods on a location-tracking dataset from the CAVA database (Arnaud *et al.* 2008), previously analysed by Forbes & Wraith (2014). The three-dimensional data are audio–visual recordings of three moving and speaking people, using binocular and binaural camera/microphone pairs. After removing instrumental artefacts, the data appear as three elongated clusters, each corresponding to a person—the goal is to distinguish the locations of the three through clustering methods.

Figure 5 shows that a number of differences are visible in the results of applying the seven methods on the data: (i) the normal mixture fitted with EM can only differentiate two clusters; the Bayesian approach also is unable to separate the two closest clusters; (ii) when the uniform-normal mixture is used, both the EM and Bayesian versions do much better,

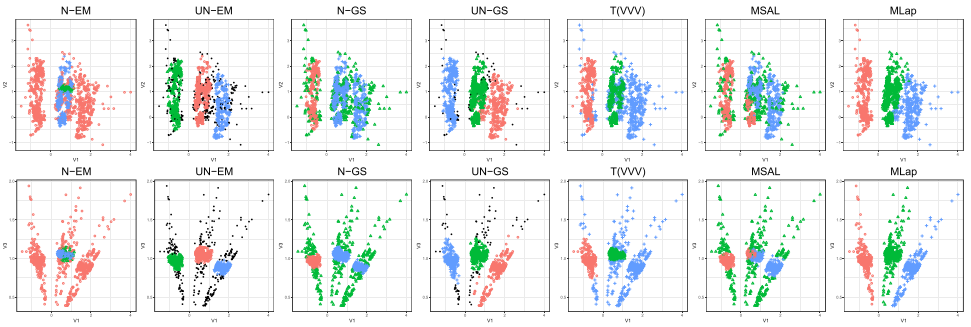

Figure 5. Comparisons of cluster prediction results for methods applied on the stereoscopic data on variables V1 and V2 (top row), and V1 and V3 (bottom row) with K set as 3. The black points correspond to predicted 'noise' by UN-EM and UN-GS, that are not allocated to any of the three clusters. Tight clustering does not perform well (as expected for heavy tailed component distributions), with three very compact clusters found, and the vast majority of points classified as outliers, and is not included in the figure.

however a large number of points are predicted as noise (17.08% for EM; 5.05% for Gibbs) and left unclassified; (iii) the *t*-mixture appears to separate the clusters well on one set of variables (V1 and V2) but misclassifies severely on another pair (V1 and V3) where all points in the tails are absorbed into a single cluster; (iv) MSAL is unable to separate the clusters well on either pair of axes, in spite of allowing for heavier tails of the component distributions; and (v) the M-Laplace mixture distinguishes the clusters accurately along both sets of axes—the heavy tails allowing outlying points to be allocated to the appropriate clusters.

4. Simulation studies

To get a better idea of the extent of observed differences in the performance of various methods in the real datasets, we next designed simulation studies involving the two proposed Bayesian clustering methods, as well as the previously used methods, in a variety of simulation settings where the 'true' clustering was known. We compared their performance under three broad headings: (i) impact of within-cluster variance, levels of overlap across clusters, and the total number of cluster components, (ii) impact of model misspecification and (iii) performance in application-inspired datasets. Results of the simulation studies presented are based on ten independent replicates under each setting.

4.1. Impact of cluster variance, overlap and total number

The simulation settings for all datasets used in the following sections are given in Table S5 in Appendix II. In case study 1, we simulated multiple datasets using a five-component Gaussian mixture, with two overlapping clusters, and 'noise' points that constituted about 20% of the dataset. For N-EM and UN-EM, we used results both from the best model chosen by BIC (this turned out to be the 'VII' model for all 10 datasets in N-EM and 'VII' (9) and 'VEI'(1) for UN-EM) as well as the unrestricted (variable volume, shape and orientation: VVV) model, which was the most comparable to the general model fitted by Gibbs sampling. Priors for the Bayesian models were chosen to be proper to ensure identifiability of the mixtures (if empty clusters were obtained) but minimally informative to avoid biasing the parameter estimates in any direction. The boxplots of the ARI and CCR for all the fitted models (Fig. 6), and the MSE for parameters of the Gaussian mixture models show that UN-EM and UN-GS had similarly high ARI and CCR, and the lowest average MSE (Table 4), while tight clustering and MSAL had the lowest CCR and ARI.

Figure S11 in Appendix II shows the results from a single replicate dataset, which illustrates the general pattern of performance across the methods. Tight clustering split one of the clusters and lost an entire cluster by classifying it as 'noise', leading to a low ARI and CCR. The noise points are mostly found as a separate cluster by the normal mixture-based models, while the general *t* and MSAL approaches split up some clusters as well as the noise points. UN-GS and UN-EM have the lowest misclassification rates, being able to distinguish most outlying points, while the *t*-mixture and normal mixtures had similar CCRs.

Next, we investigated the performance of the model selection criteria for a subset of case study 1, where the data were generated from a three-component mixture, but the number of clusters (excluding the noise) was allowed to vary between 2 and 5, for the normal and normal-uniform mixtures. For N-EM, the BIC appeared to select the correct number of

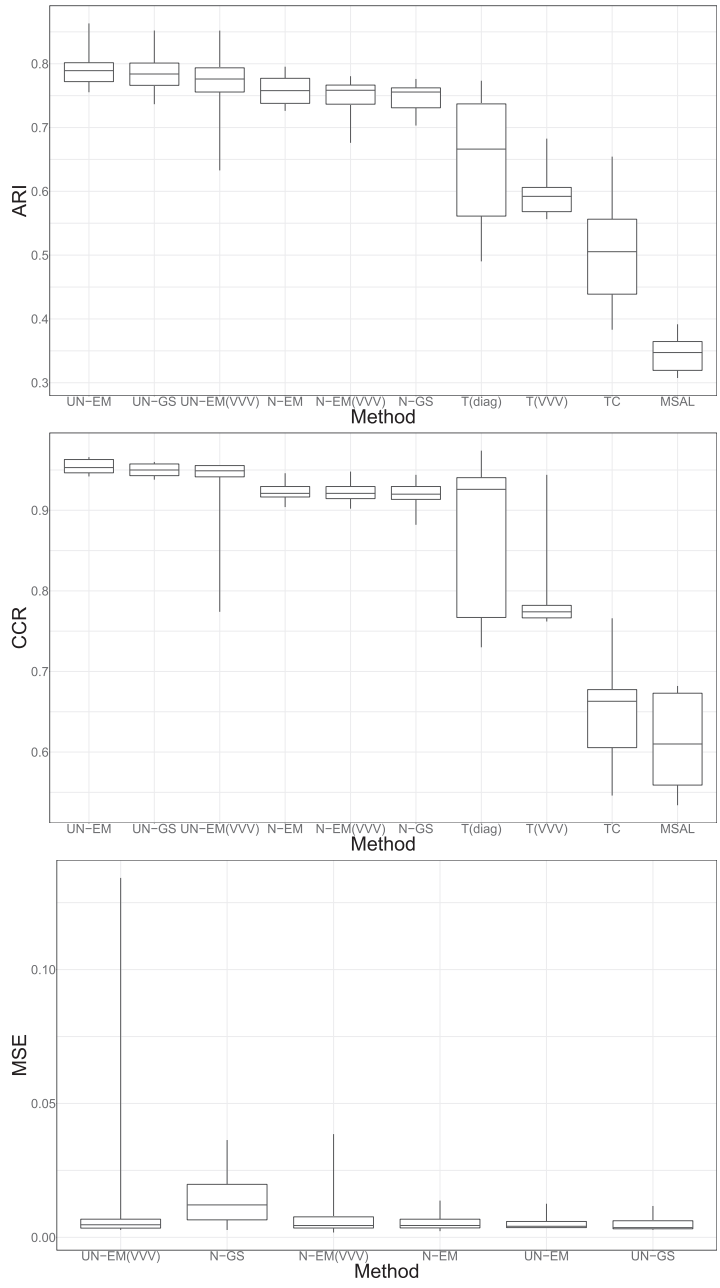

Figure 6. Adjusted rand index (ARI), correct classification rate (CCR) and mean MSE for clustering methods over 10 replicated datasets in Case study 1. Models used are: normal mixtures fitted with (i) EM [N-EM], (ii) EM with the unrestricted model [N-EM(VVV)] (iii) Gibbs sampling [N-GS], normal mixtures with uniform noise component fitted with (iv) EM [UN-EM] and (v) Gibbs [UN-GS], t-mixture with diagonal variance (vi) [T(diag)] and (vii) general variance [T(VVV)], (viii) tight clustering [TC] and (ix) mixtures of shifted asymmetric Laplace distributions [MSAL].

Table 4. Average Adjusted rand index (ARI) and correct classification rate (CCR) in Case studies 1, 2 and 3, using the methods: (i) tight clustering (TC), (ii) normal mixture-model based clustering (N-EM), (iii) normal mixture model-based clustering with uniformly distributed noise (UN-EM), (iv) Bayesian normal mixture model (N-GS), (v) Bayesian normal mixture model with uniformly distributed noise (UN-GS), (vi) t-mixture models (T), (vii) shifted asymmetric Laplace mixtures (MSAL) and (viii) M-Laplace mixture models (MLap). N-EM, and UN-EM results are shown for both the highest Bayesian information criterion (BIC) model and the unrestricted covariance model (VVV) while the t-mixture results are shown for the unrestricted (VVV) and diagonal (diag) covariance models. MLap was not used in Case study 1 as it is not appropriate for data with very compact, overlapping clusters that it would tend to merge into a single cluster. The number of clusters K was set at 5, 3, 3 for UN-EM and UN-GS, while the other methods were allowed an extra cluster.

	Case study	1	2	3
TC	ARI	0.508	0.519	0.163
	CCR	0.649	0.646	0.156
N-EM(VVV)	ARI	0.747	0.780	0.779
	CCR	0.922	0.907	0.858
N-EM	ARI	0.758	0.788	0.828
	CCR	0.923	0.911	0.902
UN-EM(VVV)	ARI	0.769	0.804	0.648
	CCR	0.932	0.926	0.773
UN-EM	ARI	0.791	0.801	0.708
	CCR	0.954	0.924	0.817
N-GS	ARI	0.746	0.784	0.919
	CCR	0.919	0.912	0.960
UN-GS	ARI	0.785	0.809	0.923
	CCR	0.950	0.930	0.995
T(VVV)	ARI	0.595	0.800	0.861
	CCR	0.800	0.941	0.918
T(diag)	ARI	0.648	0.802	0.893
	CCR	0.868	0.942	0.941
MSAL	ARI	0.345	0.567	0.643
	CCR	0.614	0.833	0.772
MLap	ARI	–	0.652	0.922
	CCR	–	0.978	0.996

clusters, while for UN-EM, the number of clusters was over-estimated by 1 on average. For the Bayesian models, the WAIC was able to choose the correct number of clusters in every case, whereas the BIC occasionally underestimated the number of clusters by 1. With the correct K, the average ARI computed for UN-GS (0.747) was slightly higher than that for N-EM (0.7464) and N-GS (0.7132).

4.2. Impact of model misspecification

The next study was designed to examine how well these methods performed in the case of model misspecification. The datasets in case study 2 were simulated from a three-component mixture of t-distributions, each component with 3 df, and with other parameter and noise settings given in Table S5 in the Appendix II. The ARI still appeared to be the highest for UN-GS (Figure S12 in Appendix II) but the CCR from the t-mixture model was in general slightly higher (as expected). UN-GS had the next highest average CCR, followed by UN-EM. Tight clustering continued to give overly compact clusters, identifying many cluster points as noise; and increasing the total number of clusters to 4 only led to further cluster splitting. Figure S13 in Appendix II shows an example of the performance of the various methods on t-mixture data, with the Normal mixtures with noise components doing slightly better than the non-noise versions, and the t-mixture doing slightly better overall. The asymmetric Laplace merged two clusters and split another, leading to a higher misclassification rate.

4.3. Performance in application-inspired datasets

Our final study was motivated by the goal of replicating the observed performance of the methods in datasets that were simulated to have similar characteristics to the SNP genotyping data in Section 1. Ten replicated datasets, for three 'genotype' clusters each, were simulated from a mixture of three truncated normal distributions in the proportions $1/2:1/3:1/6$, restricted to the positive quadrant, and noise points that constituted about 3% of the dataset, were added (parameter settings in Table S5 in Appendix II).

Applying the N-GS model on this dataset, in most cases, exhibited label-switching, which had to be manually corrected by visualising the posterior density plots and re-ordering observations. This could potentially be a problem with using N-GS in clustering large numbers of genotyping datasets in GWAS. The difference between the actual ARI and CCR was not significantly impacted by the relabelling, in this example, but is a possibility. UN-GS did not exhibit symptoms of label-switching in any of the simulated runs or real datasets. UN-GS and the Bayesian M-Laplace mixture had the highest average ARI and CCR, and were also most consistent over the multiple replicated datasets (Fig. 7). There were slight differences in the cluster components detected by UN-GS and M-Laplace, the M-Laplace tending to have heavier component tails, thus including more distant points into the clusters, whereas UN-GS classified more distant points as noise; however, both methods successfully classified the majority of the data points into their correct clusters (an example is shown in Figure S14 in Appendix II). This study demonstrated that even when the data significantly mismatched model assumptions (reflecting characteristics of SNP genotyping), the M-Laplace mixture and UN-GS appeared relatively robust and accurate in clustering.

4.4. Summary

Overall, it was found that (i) tight clustering tended to find small and compact clusters with low within-cluster variance, with a tendency to split up high variance clusters, thus performing relatively worse compared to model-based methods, with heavy-tailed data; (ii) in the presence of noise, Uniform-Normal mixtures showed better performance in terms

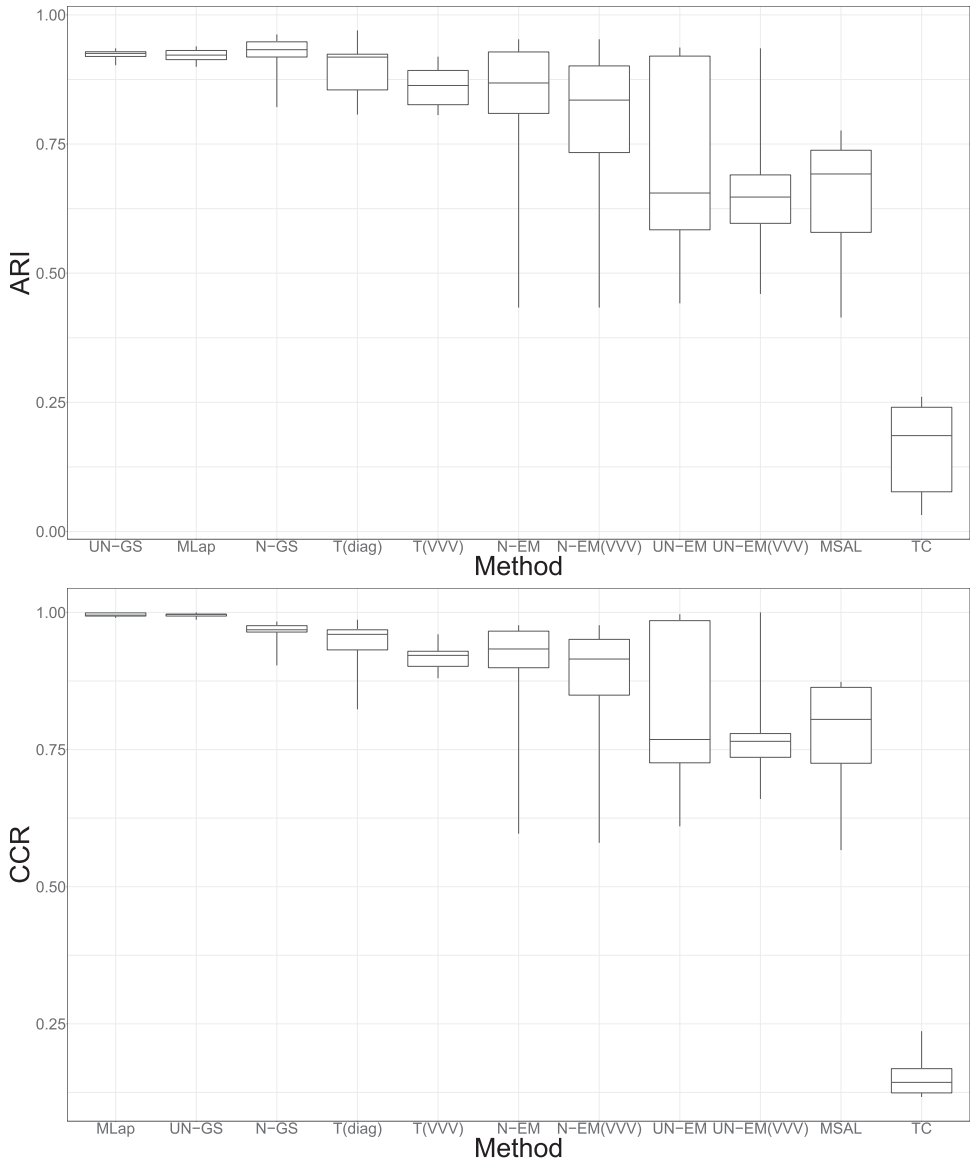

Figure 7. Adjusted rand index (ARI) and correct classification rate (CCR) summarized over 10 replications for all clustering methods in Case study 3.

of ARI, CCR and BIC in most scenarios, and Gibbs sampling was more robust to the presence of noise than EM-based methods; (iii) in a scenario where data were simulated from a t-mixture with additional noise, UN-GS still performed comparably to the generative model, which could be ascribed to the EM algorithm having a tendency to get trapped at local optima in the presence of noise, while Gibbs sampling is successful in exploring the posterior distributions more broadly; and (iv) the M-Laplace mixture was

highly successful in detecting clusters in data inspired by real genotyping studies, that had significant departures from normality in the form of bounded data with elongated, non-ellipsoidal groups.

5. Discussion

In this article, we have explored methods of Bayesian model-based clustering in the presence of noise and outliers and features of non-normality. We have proposed an extension to the method of Bayesian Gaussian mixture modelling through incorporating outliers, as well as an M-Laplace mixture model-based approach for heavy-tailed and non-normal data clusters. Both approaches were implemented through efficient data augmentation algorithms, and appeared robust to departures from normality, and less prone to problematic aspects of Bayesian mixture modelling, such as label-switching, in the datasets under investigation. Our approaches gave promising results in three real-life applications—a genotyping experiment, an ecological study and image classification—however, they could generally be applied to a variety of clustering problems in other areas of science.

In some of the real data examples, the clustering models fitted will certainly not contain the true cluster model due to the variables being bounded and/or discrete. Our approach aims to study the use of the new approaches to challenging data types where standard approaches do not perform well. Examples of the use of real-valued models for non-real-valued data can be widely found in the clustering literature, for example, the success of unrestricted skew-t mixtures in clustering bounded flow cytometry data (e.g. Lee & McLachlan 2013a,b), and the widespread use of real-valued mixture distributions for non-negative or discrete data (e.g. Andrews & McNicholas 2012; Lee & McLachlan 2013a,b; Franczak, Browne & McNicholas 2014)—although we acknowledge that other specific distributions such as restricted skew distributions (Lee & McLachlan 2013a,b) or tree models (Poon, Liu & Zhang 2018) could be sometimes more appropriate. Such model misspecification may not be ideal, and could lead to slight biases in parameter estimation, but our numerical studies (including a simulation study under a truncated data scenario) show that as long as there are sufficient data to build the cluster structure, the accuracy of cluster grouping is not significantly affected, along with substantial gains in computational efficiency and stability over some of the more complex models.

In practical terms, these methods have much future scope for improvement. The Bayesian tight clustering algorithm assumes the noise structure is uniform, which may not always be realistic—the model could be further extended to account for more complex and informed noise patterns, based on the specific application. For model selection, the BIC appeared to give an overestimate of the total number of clusters, and although the performance of the WAIC appeared more promising for the Bayesian models, more investigation into this would be desirable. Alternative approaches, such as variational approximations (Forbes *et al.* 2019) may also be explored. In our applications here, it is also important to observe that we have looked at comparatively low-dimensional datasets. In many other applications in genomics where clustering is needed, for example, in single-cell RNA sequencing experiments (Kiselev, Andrews & Hemberg 2019), both the dimensionality and size of datasets is very large, running into tens of thousands or higher. In such situations, the interpretation of clusters and clustering results presents an extra challenge, beyond even the computational challenges that may ensue from a scaling-up of the methods presented here.

Supporting information

Additional supporting information may be found in the online version of this article at http://wileyonlinelibrary.com/journal/anzs.

References

ANDREWS, D.F. & MALLOWS, C.L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36**, 99–102.

ANDREWS, J.L. & MCNICHOLAS, P.D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing* **22**, 1021–1029.

ARNAUD, E., CHRISTENSEN, H., LU, Y.C. *et al.* (2008). The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements. In ICMI '08.

AUTON, A., BROOKS, L.D., DURBIN, R.M. *et al.* (2015). A global reference for human genetic variation. The 1000 Genomes Project Consortium. *Nature* **526**, 68–74.

AZZALINI, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics* **32**, 159–188.

BANFIELD, J.D. & RAFTERY, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

BELLÉGO, C., BENATIA, D. & PAPE, L.D. (2021). Dealing with the log of zero in regression models. Working papers, Center for Research in Economics and Statistics. Available from URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3444996.

BENSMAIL, H. & MEULMAN, J.J. (2003). Model-based clustering with noise: Bayesian inference and estimation. *Journal of Classification* **20**, 49–76.

BROWNE, R. & MCNICHOLAS, P. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* **43**, 176–198.

CHOY, S.B. & CHAN, J.S. (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zealand Journal of Statistics* **50**, 135–146.

DASGUPTA, A. & RAFTERY, A.E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* **93**, 294–302.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–38.

DIEBOLT, J. & ROBERT, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)* **56**, 363–375.

ELTOFT, T., KIM, T. & LEE, T.W. (2006a). Multivariate scale mixture of Gaussians modeling. In *Independent Component Analysis and Blind Signal Separation*, eds. J. Rosca & D. Erdogmus. pp. 799–806. Springer-Verlag, Berlin.

ELTOFT, T., KIM, T. & LEE, T.W. (2006b). On the multivariate Laplace distribution. *IEEE Signal Processing Letters* **13**, 300–303.

ERICKSON, S. & CALLAWAY, J. (2016). SNPMClust: Bivariate Gaussian genotype clustering and calling for Illumina microarrays. *Journal of Statistical Software* **71**, 1–9.

ESTER, M., KRIEGEL, H.P., SANDER, J. & XU, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96*, eds. E. Simoudis, J. Han, U. Fayyad & M. Usama, pp. 226–231. AAAI Press, Portland, Oregon.

ESTRADA, K., RIVADENEIRA, F. & EVANGELOU E. *et al.* (2012). Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature Genetics* **44**, 491–501.

EVERITT, B. (1974). *Cluster Analysis*. London: Heinemann Educational Publishers.

FORBES, F., ARNAUD, A., LEMASSON, B. & BARBIER, E. (2019). Component elimination strategies to fit mixtures of multiple scale distributions. In *RSSDS 2019, Communications in Computer and Information Science*, vol. **1150**, pp. 81–95. Melbourne, Australia: Springer.

FORBES, F. & WRAITH, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing* **24**, 971–984.

FRANCZAK, B.C., BROWNE, R.P. & MCNICHOLAS, P.D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis & Machine Intelligence* **36**, 1149–1157.

FRANCZAK, B.C., BROWNE, R.P., MCNICHOLAS, P.D. & BURAK, K.L. (2018). MixSAL: Mixtures of Multivariate Shifted Asymmetric Laplace (SAL) Distributions. R package version 1.0.

FRUHWIRTH-SCHNATTER, S. & PYNE, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11**, 317–336.

GILKS, W.R., BEST, N.G. & TAN, K.K.C. (1995). Adaptive rejection Metropolis sampling. *Applied Statistics* **44**, 455–472.

HUBERT, L. & ARABIE, P. (1985). Comparing partitions. *Journal of Classification* **2**, 193–218.

ICES (2020). North Sea International Bottom Trawl Survey (1972-2020). Consulted on 2020-11-30. Available from URL: http://datras.ices.dk.

JOHNSON, N.L., KOTZ, S. & BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions*. 2nd edn. vol. **2**. New York: John Wiley & Sons.

JONES, P.N. & MCLACHLAN, G.J. (1990). Laplace-normal mixtures fitted to wind shear data. *Journal of Applied Statistics* **17**, 271–276.

JOO, Y., CASELLA, G. & HOBERT, J. (2010). Bayesian model-based tight clustering for time course data. *Computational Statistics* **25**, 17–38.

KISELEV, V.Y., ANDREWS, T.S. & HEMBERG, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics* **20**, 273–282.

KOMAREK, 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 & Data Analysis* **53**, 3932–3947.

KOTZ, S., KOZUBOWSKI, T. & PODGORSKI, K. (2001). The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance. Birkhäuser, Boston, MA.

KOTZ, S. & NADARAJAH, S. (2004). *Multivariate T-Distributions and Their Applications*. Cambridge University Press, Cambridge, UK.

KUNDU, D. (2017). Multivariate geometric skew-normal distribution. *Statistics* **51**, 1377–1397.

KYUNG, M., GILL, J., GHOSH, M. & CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* **5**, 369–411.

LEE, S. & MCLACHLAN, G. (2013a). Model-based clustering and classification with non-normal mixture distributions. *Statistical Methods & Applications* **22**, 427–454.

LEE, S.X. & MCLACHLAN, G.J. (2013b). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification* **7**, 241–266.

LEE, S.X. & MCLACHLAN, G.J. (2016). Finite mixtures of canonical fundamental skew t-distributions. *Statistics and Computing* **26**, 573–589.

LEE, S.X. & MCLACHLAN, G.J. (2019). On mean and/or variance mixtures of normal distributions. In *Statistical Learning and Modeling in Data Analysis*, eds. S. Balzano, G.C. Porzio, R. Salvatore, D. Vistocco & M. Vichi. pp. 117–127. Springer, Cham.

LIN, T., LEE, J. & NI, H. (2004). Bayesian analysis of mixture modelling using the multivariate t distribution. *Statistics and Computing* **14**, 119–130.

LIU, J.S., WONG, W.H. & KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

MADAN, D. & SENETA, E. (1990). The variance gamma (Vg) model for share market returns. *The Journal of Business* **63**, 511–524.

MARIN, J.M. & ROBERT, C.P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics (Springer Texts in Statistics)*. Springer-Verlag, Berlin.

MCLACHLAN, G. & BASFORD, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Statistics, textbooks and monographs. vol. **84**, Marcel Dekker, New York.

MCLACHLAN, G. & PEEL, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Hoboken.

PARK, T. & CASELLA, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681–686.

PEEL, D. & MCLACHLAN, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* **10**, 339–348.

POON, L.K., LIU, A.H. & ZHANG, N.L. (2018). Uc-ltm: unidimensional clustering using latent tree models for discrete data. *International Journal of Approximate Reasoning* **92**, 392–409.

R CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, Available from URL: https://www.R-project.org/.

RAFTERY, A., NEWTON, M., SATAGOPAN, J. & KRIVITSKY, P. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics* **8**, 1–45.

REDIVO, E., NGUYEN, H.D. & GUPTA, M. (2020). Bayesian clustering of skewed and multimodal data using geometric skewed normal distributions. *Computational Statistics & Data Analysis* **152**, 107040.

RYDÉN, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: a computational perspective. *Bayesian Analysis* **3**, 659–688.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

SCRUCCA, L., FOP, M., MURPHY, T.B. & RAFTERY, A.E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* **8**, 289–317.

SHERRY, S.T., WARD, M.H., KHOLODOV, M., *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308–311.

TSENG, G. & WONG, W. (2005). Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**, 10–16.

TSENG, G.C. & WONG, W.H. (2018). *tightClust: Tight Clustering*. R package version 1.1.

WANG, K., NG, S. & MCLACHLAN, G. (2009). Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data. In *Conference of Digital Image Computing: Techniques and Applications, Melbourne*, eds. H. Shi, Y. Zhang, M. Bottema, B. Lovell & A.M. Los Alamitos, pp. 526–531. California: IEEE Computer Society.

WATANABE, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571–3594.

WEERARATHNE, I.A., MONK, J. & BARRETT, N. (2021). Sample-size requirements for accurate length-frequency distributions of mesophotic reef fishes from baited remote underwater stereo video. *Ecological Indicators* **122**, 107262.

WIPER, M., INSUA, D.R. & RUGGERI, F. (2001). Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics* **10**, 440–454.

ZHAO, S., JING, W., SAMUELS, D.C., SHENG, Q., SHYR, Y. & GUO, Y. (2018). Strategies for processing and quality control of Illumina genotyping arrays. *Briefings in Bioinformatics* **19**, 765–775.