PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Review

Cite this article: Wade S. 2023 Bayesian cluster analysis. *Phil. Trans. R. Soc. A* **381**: 20220149.
https://doi.org/10.1098/rsta.2022.0149

Received: 22 July 2022
Accepted: 3 January 2023

One contribution of 16 to a theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

Subject Areas:
artificial intelligence, pattern recognition, statistics

Keywords:
Bayesian analysis, clustering, ensembles, mixture models, model misspecification

Author for correspondence:
S. Wade
e-mail: sara.wade@ed.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6423927.

THE ROYAL SOCIETY PUBLISHING

Bayesian cluster analysis

S. Wade

School of Mathematics and Maxwell Institute for Mathematical Sciences, University of Edinburgh, James Clerk Maxwell Building, Edinburgh, UK

SW, 0000-0002-6547-5555

Bayesian cluster analysis offers substantial benefits over algorithmic approaches by providing not only point estimates but also uncertainty in the clustering structure and patterns within each cluster. An overview of Bayesian cluster analysis is provided, including both model-based and loss-based approaches, along with a discussion on the importance of the kernel or loss selected and prior specification. Advantages are demonstrated in an application to cluster cells and discover latent cell types in single-cell RNA sequencing data to study embryonic cellular development. Lastly, we focus on the ongoing debate between finite and infinite mixtures in a model-based approach and robustness to model misspecification. While much of the debate and asymptotic theory focuses on the marginal posterior of the number of clusters, we empirically show that quite a different behaviour is obtained when estimating the full clustering structure.

This article is part of the theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

1. Introduction

Clustering is one of the canonical forms of unsupervised learning, which aims to divide data points into *similar* groups, and has been used in various applications. Examples include astronomy to discover types of stars by clustering astrophysical measurements [1,2]; geosciences to detect minefields or seismic faults from spatial data [3]; natural language processing where topic models employ clustering to infer latent topics across documents for information retrieval [4]; biomedicine to discover groups of individuals or genes with similar patterns in gene expression or omics data [5,6]

and many more. In many applications, the allocations and patterns within each cluster are of direct interest, while in other settings, clustering may be used in data preprocessing or feature engineering, or it may be used, not to recover homogeneous sub-populations, but, rather, as a building block in kernel methods for flexible density estimation or regression [7].

Algorithmic approaches, such as hierarchical, partition-based or density-based clustering, are commonly used in clustering. Hierarchical clustering builds a tree of clustering solutions, either through an agglomerative (bottom-up) and divisive (top-down) strategy [8,9], and results crucially depend on the choice of dissimilarity and linkage. Partition-based algorithms, including k-means [10] and k-mediods [9], aim to divide data into subsets by minimizing a specified loss function. In contrast to hierarchical algorithms, they provide a single clustering solution which is revisited and iteratively optimized. While the k-means algorithm is by far the most popular tool for clustering [11], there are several drawbacks (e.g. only covers numerical variables, sensitive to local optimum, requires the number of clusters k to be pre-specified). Lastly, density-based algorithms, such as DBSCAN [12,13], are based on the general idea of defining a cluster as a connected dense component. They are capable of discovering clusters of arbitrary shapes but lack interpretability. Although such algorithmic approaches are widely used, they are largely heuristic and not based on formal models, prohibiting the use of statistical tools, for example, in determining the number of clusters, and they lack measures of uncertainty in the clustering solution.

An alternative approach is model-based clustering, which uses mixture models, where each (non-empty) mixture component corresponds to a cluster [14–16]. Problems of determining the number of clusters and the component probability distribution can be dealt with through statistical model selection, for example, through various information criteria. The expectation–maximization (EM) algorithm is typically used for maximum likelihood estimation (MLE) of the mixture model parameters, consisting of the prior group probabilities and the local parameters of each component probability distribution. Given the MLEs of the parameters, the posterior probability that a data point belongs to a group can be computed through the Bayes rule. The cluster assignment of the data point corresponds to the component with maximal posterior probability, with the corresponding posterior probability reported as a measure of uncertainty. Importantly, however, this measure of uncertainty ignores uncertainty in the parameter estimates. As opposed to MLE, Bayesian mixture models incorporate prior information on the parameters and allow one to assess uncertainty in the clustering structure unconditional on the parameter estimates.

In this article, we provide a review of Bayesian approaches to clustering, including both model-based and loss-based methods (§2), along with an illustrative application to highlight the advantages of the Bayesian approach in §3. In addition, §4 brings together recent research on estimating the number of clusters and robustness to model misspecification in the model-based approach. This literature highlights the fundamental trade-off of mixture models between density estimation and clustering. As a simple solution, we discuss how one can separate the task of clustering by framing it in a decision-theoretic context. Importantly, this allows the mixture model to retain optimal statistical properties for density estimation, while also providing more robust clustering estimates.

2. Bayesian cluster analysis

In the context of clustering, the observed data consists of measurements $\boldsymbol{y} = (y_1, \ldots, y_n)$ drawn from a heterogeneous population consisting of an unknown number of homogeneous sub-populations. The observed $y_i \in \mathcal{Y}$ may be continuous, discrete, mixed or more complex in nature (e.g. functional data). Each data point is associated with a discrete latent variable z_i (also called the allocation variable) indicating the group membership of the data point, i.e. $z_i = j$ if y_i belongs to the jth group, and $z_i = z_{i'}$ if y_i and $y_{i'}$ belong to the same sub-population. We are interested in obtaining estimates of clustering structure characterized by the latent $\boldsymbol{z} = (z_1, \ldots, z_n)$ as well as describing the patterns within each cluster and understanding uncertainty in clustering structure.

To achieve this, the Bayesian approach constructs a posterior distribution over clusterings, $\pi(z|y) \propto p(y \mid z)\pi(z)$, where $\pi(z)$ represents the prior over the space of clusterings and $p(y \mid z)$ can be defined through a model-based (§2a) or a loss-based (§2b) approach.

It is worth emphasizing that clustering is often referred to as an ill-posed problem, as it aims to discover unknown patterns or structures in the data. The notion of a cluster depends on the application at hand and can often be challenging to characterize formally. A unique clustering solution often does not exist [17]. Thus, one must carefully consider the model or loss employed and, importantly, also characterize uncertainty in the clustering solution. To achieve the latter, Bayesian cluster analysis provides a formal framework through both the posterior distribution over the entire space of clusterings and by creating an ensemble of clustering solutions sampled from the posterior. Moreover, this also helps to mitigate sensitivity to local optima which adversely impact all clustering algorithms due to the sheer size of the space.

As a note, in this article, we focus on clustering based on a single dataset. However, the massive growth in data acquisition and technologies has led to a number of interesting extensions. This includes combining multiple data sources through data integration [18–20], hierarchical Bayesian frameworks for partially exchangeable or nested data [21–28], hidden Markov models and other extensions for temporal data [29,30], accounting for spatially indexed data [31–33], incorporating general covariate information [34–37] and more.

(a) Model-based approach

The most popular approach to Bayesian clustering employs a model-based framework through mixture models [38,39]. In this case, the data are assumed to be conditionally i.i.d. from a convex combination of parametric components:

$$y_i|\boldsymbol{w}, \boldsymbol{\theta}, \psi \overset{\text{iid}}{\sim} \sum_{j=1}^{J} w_j f(\cdot|\theta_j, \psi) = \int f(\cdot|\theta, \psi)\, \mathrm{d}H(\theta), \tag{2.1}$$

where $f(y \mid \theta, \psi)$ is a fixed parametric density, often referred to as the kernel, with component-specific parameters contained in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ and global parameters ψ, and the mixture weights $\boldsymbol{w} = (w_1, \ldots, w_J)$ are non-negative and sum to one. In the equivalent integral representation on the right-hand side of (2.1), $H = \sum_{j=1}^{J} w_j \delta_{\theta_j}$ represents the mixing measure. Yet another equivalent representation, useful for clustering, makes use of allocation variables z:

$$y_i|z_i = j, \theta_j, \psi \overset{\text{ind}}{\sim} f(\cdot|\theta_j, \psi), \quad z_i \overset{\text{iid}}{\sim} \text{Cat}(w_1, \ldots, w_J),$$

where $\text{Cat}(\cdot)$ represents the categorical distribution with parameter $\boldsymbol{w}$. In the Bayesian setting, the model is completed with a prior on the unknown parameters $\boldsymbol{w}$, $\boldsymbol{\theta}$ and ψ (or equivalently on the unknown mixing measure H and ψ).

In order to obtain clusters of practical relevance, the kernel $f(\cdot|\theta, \psi)$ should be carefully selected to reflect the shape and properties of a cluster for the application at hand. A standard choice is the multivariate Gaussian distribution, $f(\cdot|\theta, \psi) = N(\cdot \mid \mu_j, \Sigma_j)$. In fact, the widely used k-means algorithm can be seen as a limiting case of the EM algorithm for Gaussian mixture models, where the kernel is $N(\cdot \mid \mu_j, \sigma^2 I)$ [40,41]. This highlights that k-means imposes restrictive cluster shapes, specifically, all clusters have the same spherical shape of equal size in all dimensions, with only the centres μ_j allowed to differ across clusters. More generally, Gaussian mixture models relax this assumption by allowing different ellipsoidal shapes and sizes across clusters. The cluster-specific covariance matrices can be parametrized as $\Sigma_j = \lambda_j D_j A_j D_j^T$, where λ_j, D_j and A_j control the volume, orientation and shape, respectively, of the ellipsoid and each parameter can be cluster-specific or global for general geometric cross-cluster constraints [14,42]. Other types of constraints on the covariance matrices can also be considered, such as mixtures of factor analysers [43,44] and mixtures of Gaussian graphical models within a casual framework [45,46].

However, depending on the data characteristics and aim, different kernels are more appropriate. For continuous data, skewed shapes and/or robustness to outliers can be accounted for through multivariate skew-normal or t-distributions [47,48], shifted asymmetric Laplace distributions [49] and normal-inverse Gaussian distributions [50]. For directional data on the unit sphere, examples include mixtures of Kent distributions [51], von-Mises–Fisher distributions [52] or Gaussian distributions in distinct tangent spaces [53]. For discrete data, mixtures of Bernoulli or multinomial distributions, known as latent class models, are appropriate for categorical data [4,54]; latent variable approaches using a logistic or probit transformation are employed for ordinal data [55,56]; and mixtures of Plackett–Luce models are used for rankings [57]. For count data, examples include mixtures of Poisson distributions [58,59], negative-binomial distributions [60] and rounded continuous kernels [61,62], as well as zero-inflated Poisson or negative-binomial distributions for sparse counts [63]. Mixed data of different types can be modelled by assuming either conditional independence, combining appropriate kernels through a product operation, or through a latent variable approach [64,65]. Moreover, the kernels may be themselves mixtures for increased flexibility [66,67].

In high-dimensional settings, challenges arise from both a computational [68] and a theoretical [69] perspective. In particular, Chandra *et al.* [69] show that one needs to be extremely careful in specifying both the kernel and the prior on θ in high-dimensions; otherwise, the posterior can degenerate on extreme clustering structures. One solution to overcome this employs variable selection methods within the mixture model to identify relevant variables that are informative for clustering, either using spike-and-slab priors [70–73] or shrinkage priors [74,75]. An alternative approach is to incorporate dimension reduction methods within the mixture model. This includes cluster-specific dimension reduction methods to reduce the number of parameters, such as mixtures of factor analysers [76] or parsimonious Gaussian mixtures [77], as well as approaches that conduct clustering directly on the lower dimensional space [69]. While approaches mainly focus on incorporating linear dimension reduction, extensions based on nonlinear dimension reduction can also be considered [78].

(b) Loss-based approach

Partition-based clustering algorithms aim to minimize a specific loss function and are widely adopted but lack any quantification of uncertainty in the clustering solution. To address this and bridge the gap between partition-based and model-based approaches, the recent work of Rigon *et al.* [79] employs a generalized Bayesian framework through the use of Gibbs posteriors [80]. Specifically, the generalized posterior is defined as

$$\pi(\boldsymbol{z}|\boldsymbol{y}) \propto \exp(-\lambda \ell(\boldsymbol{z}, \boldsymbol{y}))\pi(\boldsymbol{z}),$$

where the loss function has the form $\ell(\boldsymbol{z}, \boldsymbol{y}) = \sum_{j=1}^{k} \sum_{i:z_i=j} \mathcal{D}(y_i, \boldsymbol{y}_j)$ with $\boldsymbol{y}_j = \{y_i : i = z_j\}$ denoting the observations belonging to the jth cluster and $\mathcal{D}(y_i, \boldsymbol{y}_j) \geq 0$ quantifying the discrepancy of y_i from the jth cluster. A simple example is the k-means loss which sets $\mathcal{D}(y_i, \boldsymbol{y}_j) = \| y_i - \boldsymbol{y}_j \|^2$ (additional examples can be found in Rigon *et al.* [79]). Fixing the number of clusters k, the prior $\pi(\boldsymbol{z})$ is chosen to be uniform over the set of partitions with k clusters. In this case, the maximum a posterior (MAP) estimator $\widehat{\boldsymbol{z}}_{\text{MAP}} = \text{argmax}_{\boldsymbol{z}} \pi(\boldsymbol{z} \mid \boldsymbol{y})$ corresponds to minimizing the loss function, e.g. under the k-means loss, $\widehat{\boldsymbol{z}}_{\text{MAP}}$ is the k-means solution. This provides an important link to partition-based approaches but also a significant enhancement through the uncertainty quantification offered by the Bayesian framework. However, a drawback of Bayesian loss-based clustering is that assumptions defining the notion of a cluster are less explicit compared with the model-based approach.

In addition, we highlight other interesting work integrating algorithmic approaches within a Bayesian framework. This includes combining density-based methods with a Bayesian model-based approach [81]; Bayesian hierarchical clustering which builds a tree of hierarchical clustering solutions based on Bayesian nonparametric (BNP) mixture models [82–85] and Bayesian distance-based clustering based on pairwise distances between observations [86–89].

(c) Priors

(i) Number of clusters

One of the most difficult and important questions in clustering regards the choice of the number of clusters. In the model-based approach, the distinction between the number of components J and the number of clusters k requires emphasis. In fact, there may be no observations allocated to some components in the mixture, with possibly very small or even zero weight w_j for some $j \in \{1, \ldots, J\}$. Thus, the number of components provides an upper bound, i.e. $k \leq J$. In general, there are four approaches to infer the number of clusters:

(i) Model selection tools or information criteria can be used to compare the mixture model under different choices of J [90]; in this case, penalization for empty clusters is implicitly included, so that the number of components corresponds to the number of clusters.
(ii) Mixtures of finite mixtures (MFM) [91–93] extend the hierarchy of the model with a prior on the number of components.
(iii) Overfitted mixtures specify J as an upper bound on the number of clusters with a sparsity promoting prior on the weights, which implicitly defines a prior on the number of clusters [74,94–96].
(iv) BNP mixtures [97] assume $J = \infty$ and can be viewed as a limiting case of overfitted mixtures, with the Dirichlet process (DP) mixture [98] being the most widely-used example.

In the first approach, uncertainty on the number of clusters is lost with model fits based on all other choices of J disregarded. Instead, the subsequent three approaches are more natural from a Bayesian perspective, as they provide a posterior on the number of clusters, reflecting uncertainty.

(ii) Weights

To specify the prior on the weights, the standard choice for finite mixtures is the symmetric Dirichlet distribution, $(w_1, \ldots, w_J) \sim \text{Dir}(\alpha, \ldots, \alpha)$, due to its conjugacy with respect to the categorical distribution for $\boldsymbol{z}$. A small value of the parameter α promotes sparsity in the weights, and in the extreme case when $\alpha \to 0$, all prior mass is placed on the vertices of the simplex, with all weight on a single component. In overfitted mixtures, this sparsity property is essential to effectively regularize and prune extra components [95]. The parameter α has an influential role, and van Havre *et al.* [96] develop a parallel tempering algorithm to explore different values of α. While asymmetric Dirichlet priors may also be considered, symmetry with respect to relabelling of the clusters no longer holds. More generally, other distributions beyond the Dirichlet may be considered, such as the Generalized Dirichlet distribution [99], multinomial Pitman–Yor (PY) process [100], BFRY priors [101], normalized jumps of a finite point process [102] or non-informative Jeffreys priors [103].

For infinite mixture models, the prior on the weights must be constructed carefully to ensure the infinite sequence of weights $(w_1, w_2, \ldots)$ sum to one. A popular construction is stick-breaking [104,105], with a discussion on choice of hyperparameters in Giordano *et al.* [106]. Alternatively, the weights can be marginalized with a prior defined directly on the partition of data points into clusters. Exchangeable partition probability functions (EPPFs) [107] are a natural class of priors that result from the basic assumptions of exchangeability and invariance with respect to cluster labels, leading to priors that only depend on $\boldsymbol{z}$ through the cluster sizes $n_j = \sum_{i=1}^{n} \mathbf{1}(z_i = j)$. For example, the EPPF obtained from the DP [108] has the form

$$\pi(\boldsymbol{z}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^k \prod_{j=1}^{k} \Gamma(n_j),$$

where $\alpha > 0$ is a hyperparameter reflecting prior belief in the number of clusters. While this form is simple, intuitive and computationally appealing, it places most prior mass on highly imbalanced

clusters with only a single parameter α to control prior uncertainty. Thus, there has been increased interest in exploring priors in the wider class of EPPFs and beyond, such as the general class of Gibbs-type priors [109,110], which contain the EPPF of the DP, PY process [111] and MFM [91] as special cases. Other proposals [112–114] aim to mitigate the *rich-get-richer* property of the DP to prefer highly imbalanced clusters; however, exchangeability often no longer holds. Subjective priors can also be specified which further enrich the parameter space by centring around prior information on the clustering structure [115,116]. In general, a BNP prior can be placed directly on the mixing measure H, which induces a prior on both the sequence of weights and the random partition. Indeed, the DP and PY are two widely used examples because they induce nice analytic priors for both the weights and partition. See Lijoi & Prünster [117] for an overview of priors beyond the DP.

(iii) Atoms

Often overlooked, the prior on the cluster-specific atoms $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ also plays an important role, especially in high-dimensional settings. Typically, the atoms are assumed to be i.i.d. from a *base measure* H_0, i.e. $\theta_j \overset{\text{iid}}{\sim} H_0$. A popular choice for H_0 is the conjugate prior to the kernel $f(y|\theta)$, which has the main advantage of computational convenience. The hyperparameters of the base measure can either be selected subjectively based on prior knowledge of the component-specific parameters, set empirically or inferred with additional hyperpriors; alternatively, data-dependent or non-informative priors can be used for H_0 [118]. For example, consider the Gaussian scale-location mixture with kernel $N(y \mid \mu_j, \sigma_j^2)$. The conjugate prior is the normal-inverse gamma:

$$\mu_j \mid \sigma_j^2 \sim N\left(\mu_0, \frac{\sigma_j^2}{c}\right), \quad \sigma_j^2 \sim \text{IG}\left(\frac{\nu}{2}, \frac{\delta^2}{2}\right).$$

A data-dependent choice for μ_0 is the empirical mean of the data. However, the scale parameter σ_j^2 represents the within cluster variance, and thus, the empirical variance provides an upper bound, where ν and δ should be carefully chosen to have most prior mass concentrated on values smaller than the empirical variance. As ν represents the degrees of freedom in the marginal t prior on μ_j, Fraley & Raftery [119] suggest fixing ν to the smallest integer that gives finite variance, i.e. $\nu = d + 2$, and setting δ^2 to be the empirical variance divided by $\widehat{k}^2$, where $\widehat{k}$ represents the prior guess on the number of clusters. The parameter c should be less than 1 to ensure higher between variance and can either be fixed, e.g. Fraley & Raftery [119] suggest a value of $c = 0.01$, or assigned a hyperprior. Other examples of data-dependent priors can be found in Diebolt & Robert [120]; Richardson & Green [93]; and Wasserman [121]. We note that vague priors are not appropriate, as they will be highly influential on the posterior distribution, often favouring high within variance and one large cluster. In addition, while non-informative Jeffreys priors [122] for the atoms often lead to improper posteriors, non-informative priors can be combined with hierarchical hyperpriors to produce proper posteriors [123].

In order to favour components that are well separated, the independence assumption on the atoms can be relaxed through the use of repulsive priors [124–126], determinantal point processes [127] or non-local priors [128]. In particular, this form of prior regularization helps to improve interpretation and encourages more meaningful clustering structures, however, it also results in more complicated posterior computations. An alternative strategy is posterior regularization, which aims to find the variational solution with minimal Kullback–Leibler (KL) divergence to the posterior in a constrained space; this has been used to impose a max-margin constraint on DP mixtures [129,130] to ensure well-separated clusters.

3. Example: discovering cell subtypes

The rise in single-cell RNA sequencing (scRNA-seq) technology allows researchers to go beyond bulk RNA measurements and understand gene expression patterns at the single-cell level.

Cells are heterogeneous in nature, and with existing technology able to record measurements on thousands of cells across thousands of genes, clustering has become an important tool to characterize latent cell types with similar expression patterns and summarize the data [131,132].

To highlight the advantages of Bayesian cluster analysis, we consider an experimental scRNA-seq dataset [133][1] collected to shed light on the development and fates of embryonic cells and the importance of the transcription factor PAX6 in the process. More generally, a single cell develops into an estimated 30 trillion cells in humans, and there is great interest in using single-cell technology and data to understand this process. In particular, PAX6 plays an important role in early development, and to empirically study this phenomenon the experimental data was collected at day E13.5 from mouse embryos under control (HET) and mutant (HOM) conditions in which PAX6 has been deleted. In the following, we present highlights of the analysis and results found in [60], which employs Bayesian model-based clustering to identify cell types, investigate how cell-type proportions change when PAX6 is knocked out, and explore if there are unique patterns when PAX6 is not present.

Most approaches for clustering scRNA-seq data separate the workflow into the steps of global normalization, dimension reduction and clustering. In particular, the data are often simply log-transformed, after adding an offset to avoid taking the logarithm of zero, and normalized in order to apply standard statistical tools. Figure 1 displays the log-transformed counts for two genes, Id4 and Meg3, and compares to data simulated from a spherical Gaussian mixture model. While it is standard to apply heuristic algorithms, such as k-means, to the log-transformed data, the incompatibility between the transformed data and the spherical clusters implied by k-means is clearly evident in figure 1. Instead, as discussed in §2a, the kernel in a model-based approach (or loss in a loss-based approach) should be more carefully considered to reflect the notion of a cluster. Moreover, the data are typically further transformed through dimension reduction methods, making the interpretation and specification of the notion of a cluster even more challenging. Indeed, as shown in Prabhakaran *et al.* [134] and Vallejos *et al.* [135], separating the workflow into normalization, dimension reduction and clustering can adversely affect the analysis, resulting in improper clustering and characterization of cell types.

Instead, the methodology and analysis of Liu *et al.* [60] follows more recent proposals which integrate normalization and clustering in a combined model-based framework [63,134,136,137]. Not only does this allow simultaneous recovery of clusters, inference of cell types and normalization, but it also provides measures of uncertainty that are propagated through the model hierarchy and coherent Bayesian updating. Importantly, the model-based approach allows for a more explicit definition of a cluster through careful specification of the kernel. Specifically, a negative-binomial kernel is employed to directly account for the count nature and overdispersion present in scRNA-seq data; that is the kernel is assumed to factorize across genes and for each gene is $\mathrm{NB}(y \mid \beta\mu_j, \phi_j)$, where μ_j and ϕ_j represent the cluster-specific mean expression and dispersion and β is the cell-specific capture efficiency, representing the fraction of transcripts recovered. For more robust estimates in the case of sparse data or small clusters, a hierarchical prior for the atoms (μ_j, ϕ_j) is used that accounts for the mean-variance relationship [138]. Borrowing of strength and shared clustering across the mutant and control conditions is permitted through a hierarchical Bayesian framework, namely the hierarchical DP [28]. For full details, see [60].

The Bayesian approach permits us to produce a range of graphical tools and tables to visualize and summarize not only point estimates but also uncertainty in the clustering structure and all parameters. The posterior estimated latent counts (corrected by the posterior capture efficiencies) are shown in figure 2*a*. Solid yellow lines separate the cells by cluster, and within each cluster, the dashed yellow line separates cells from the HOM and HET conditions. When focusing on the latent counts and genes identified as differently expressed, the clusters are visually well separated (figure 2*b*). In addition to the clustering estimate, we can also visualize uncertainty

[1]Collected and prepared by Dr Kai Boon Tan and the research group lead by Prof. David Price and Prof. John Mason at the Centre for Brain Discovery Science, University of Edinburgh.

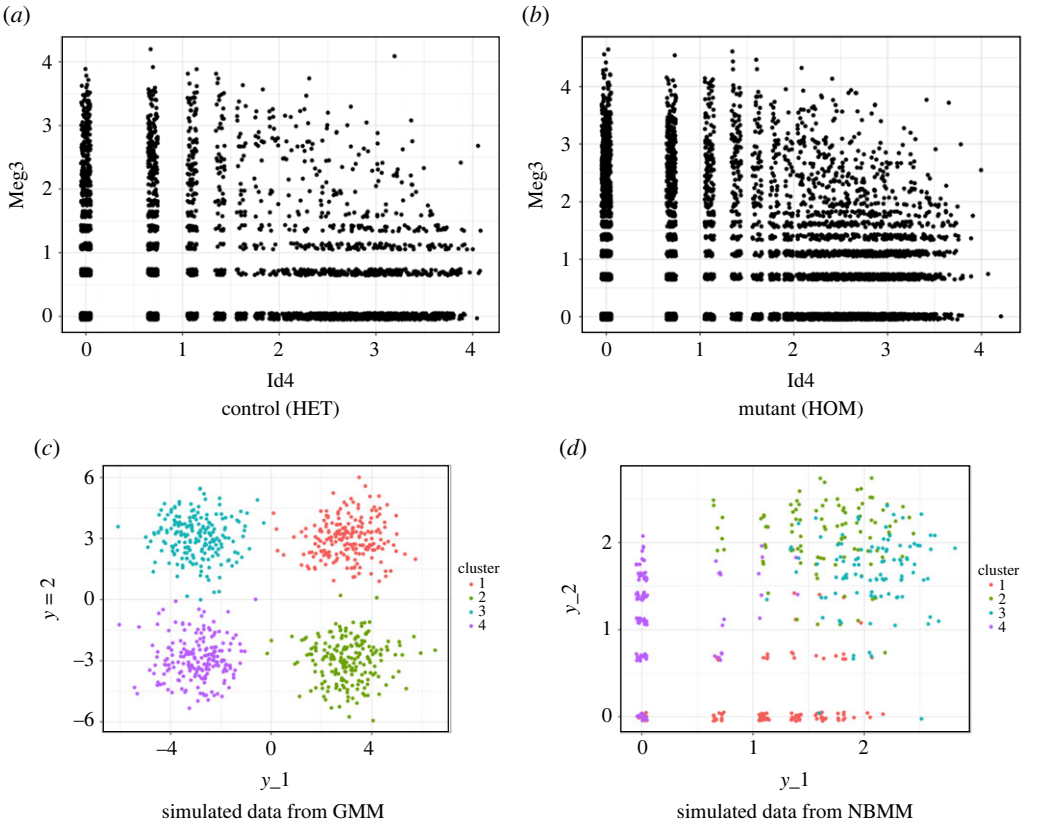

Figure 1. In order to highlight limitations of the standard workflow for scRNA-seq data, which firsts log-transforms data and then applies tools, such as *k*-means for clustering, we plot in (*a*,*b*) the log-transformed counts across all cells for two genes, Id4 and Meg3, and in (*c*) data simulated from a Gaussian mixture model (GMM); incompatibility and different characteristics are clearly observed between the real data (*a*,*b*) and simulated data (*c*). Instead, (*d*) plots log-transformed data generated from a negative-binomial mixture model (NBMM), which more closely resembles the real data. (Online version in colour.)

in the clustering structure, for example through the *posterior similarity matrix*, whose elements represent the posterior probability that two cells are clustered together. While the blocks of red highlight evident clusters of cells with posterior probability close to one, there are also some cells with more uncertainty in their allocation. Alternative tools to visualize and describe uncertainty through credible balls are provided in Wade & Ghahramani [140]. In summary, the model estimates a total of eight clusters, which are all shared in the control and mutant conditions (with some uncertainty on further splitting some clusters). Certain clusters are under or over represented in the mutant condition when PAX6 is knocked out, and further discussion on the results can be found in [60].

4. Estimating the number of clusters and model misspecification

In this last section, we bring together recent research on estimating number of clusters, with a focus on the debate between finite mixtures (MFM, overfitted) versus infinite mixtures (BNP), and robustness to model misspecification. Infinite BNP mixtures assume that the number of clusters depends on the sample size and grows unboundedly as more data are collected. With advancements in computing and general inference schemes such as Markov chain Monte Carlo (MCMC), BNP mixtures can be easily implemented. Moreover, well-established theory validates the use of BNP mixtures for asymptotically optimal density estimation [141–144]. Together these

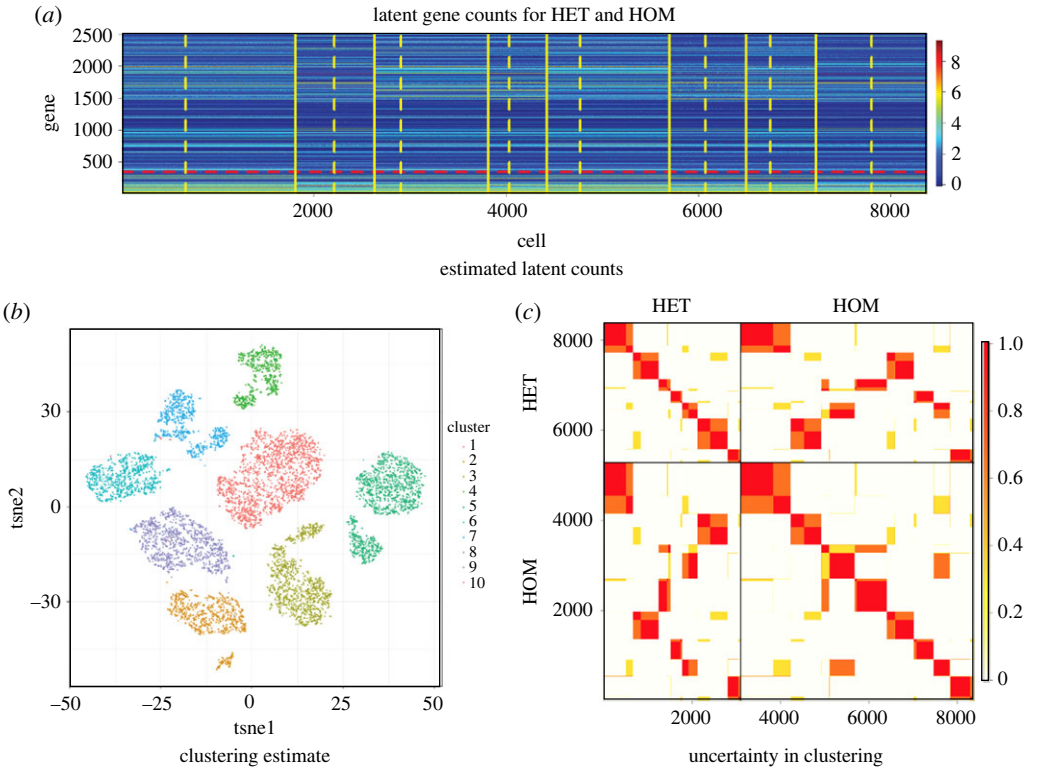

Figure 2. Highlights of the analysis of Liu *et al.* [60]. (*a*) Heat map of the posterior estimated latent RNA counts (corrected by the posterior capture efficiencies) for each cell (*x*-axis) and gene (*y*-axis). Cells from different clusters are separated by solid yellow lines, and within each cluster, the dashed yellow line separates HOM and HET. Genes above the red horizontal line are identified as differentially expressed across the clusters. (*b*) Visualization of the clustering estimate in the two-dimensional space obtained through t-distributed stochastic neighbour embedding (t-SNE [139]) of the high-dimensional data. (*c*) Uncertainty in clustering characterized by the posterior similarity matrix. (Online version in colour.)

properties and developments have led to the huge growth and adoption of BNP mixtures, especially DP mixtures, for a variety of applications in statistics and machine learning in the twenty-first century.

However, this enthusiasm was dampened by the negative results of Miller & Harrison [145,146] that provided a simple example in which the posterior on the number of non-empty components in DP mixtures is inconsistent when true number is finite. In fact, the posterior is demonstrated to be *severely* inconsistent, as the posterior probability that the number of non-empty components equals the truth asymptotically tends to zero. This is in contrast to overfitted mixtures, which asymptotically prune extra components [95,118], and MFM which yield consistent estimates for the number of components [147,148]. While overfitted mixtures can be viewed as truncated approximations to DP mixtures, this seemingly contradictory result can be explained by noting that BNP mixtures are misspecified when the true number of components is finite, and in this case, the true density lies at the boundary of the prior support [147]. Indeed, it is well-known that DP mixtures can introduce many small extra clusters. To overcome this, Guha *et al.* [147] develop a post-processing procedure to consistently estimate the true number of components by suitably truncating components with small weights and merging similar components. Instead, consistency can also be achieved by adapting the concentration parameter α of the DP to be sample-size dependent or via a suitable hyperprior, which is in fact standard in practice [149,150]. Furthermore, Frühwirth-Schnatter & Malsiner-Walli [151] illustrate that the choice of the hyperprior on the weights is far more influential on the number of clusters than whether an overfitted or DP mixture is considered.

In practice, we can expect that mixture models are misspecified in some way; either in the kernel or mixing measure, or both. Optimal asymptotic results of Bayesian mixtures for density estimation still hold (in the sense of convergence to the KL projection of the true density into the prior's support) [152,153]. However, Guha *et al.* [147] show that mild misspecification leads to very slow contraction rates of the mixing measure (with respect to its KL projection) and that the choice of the kernel is especially important; moreover, BNP mixtures are better suited to adapt to complex forms of the density in the misspecified setting. Cai *et al.* [154] also show that for MFM, even slight model misspecification leads to inconsistency for the number of clusters. As mixture models are inherently built for density estimation, it is intuitively reasonable that an overestimation of the number of clusters occurs in the misspecified case, since more components are required to accurately recover the density. This highlights the fundamental trade-off between clustering and density estimation for mixtures.

To improve model-based clustering in the misspecified setting, robust clustering methods have been developed. Examples include coarsened posteriors for mixture models [155] as well as modal-based clustering [156]. While such approaches may result in more robust clustering solutions, optimal statistical properties for density estimation may be lost. In general, both density estimation and clustering may be of interest. Thus, we consider Bayesian model-based clustering via mixtures, to retain optimal properties for density estimation, and focus on the comparison of different estimates for the number of clusters and clustering solution. In fact, we find that the number of clusters can change drastically depending on the estimator used.

The literature discussed above estimates the number of clusters via the marginal posterior on the number of non-empty components. Alternatively, the full clustering solution can be estimated, without conditioning on the number of clusters, thus also implicitly providing an estimate of this number. In fact, Rajkowski [157] demonstrates that the MAP clustering has desirable asymptotic properties in the simple example of Miller & Harrison [145], in stark contrast to the severe inconsistency of the marginal posterior on the number of clusters. To estimate the clustering solution, various *ad hoc* methods have been proposed [158–161]. Instead, we focus on a decision-theoretic approach, obtaining the optimal clustering by minimizing the posterior expectation of a specified loss function measuring the discrepancy between the true and estimating clustering. The MAP clustering is obtained under the 0–1 loss, and various search algorithms have been developed to locate the MAP solution [82,162–164]. Alternative loss functions were considered in Fritsch & Ickstadt [165]; Lau & Green [166]; Quintana & Iglesias [167] and Wade & Ghahramani [140]. Two widely used loss functions, which are considered below, are Binder's loss[2] [168] and the variation of information (VI) [169]. General algorithms to optimize the posterior expected loss can be found in [140,166], with more recent schemes in Dahl & Müller [170], Dahl *et al.* [171] and Rastelli & Friel [172] that are particularly suited to large sample sizes and parallel computations.

To illustrate the differences between the estimators, we consider two simple examples: data generated from (i) a standard normal distribution and (ii) a uniform distribution on the unit circle, and in both cases, a DP location-scale mixture of Gaussians is employed for model-based clustering. The famous example of Miller & Harrison [145] corresponds to (i) and the marginal posterior on the number of clusters was demonstrated to be severely inconsistent. Rajkowski [157] considered the second example and proved that when the within cluster variance is set too small (in a DP location mixture of Gaussians with fixed within cluster variance), the MAP clustering is not unique and partitions the ball into several, seemingly arbitrary convex sets. In all experiments, the posterior is approximated via MCMC with 10 000 iterations. For each example, 50 replicated datasets are generated and different sample sizes of $n = 100, 200$ and 500 are considered. Sensitivity to the choice of DP concentration parameter α is explored, with $\alpha = 0.5, 1$ and 2 and a sample-size dependent choice of $\alpha = 1/\log(n)$. The same search algorithm is performed for the MAP, Binder and VI estimates; specifically, first, we select the clustering which minimizes the posterior expected loss among both the MCMC draws and the set of clusterings

[2]Note that minimizing the posterior expected Binder's loss is equivalent to maximizing the posterior expected Rand Index or Hamming distance.

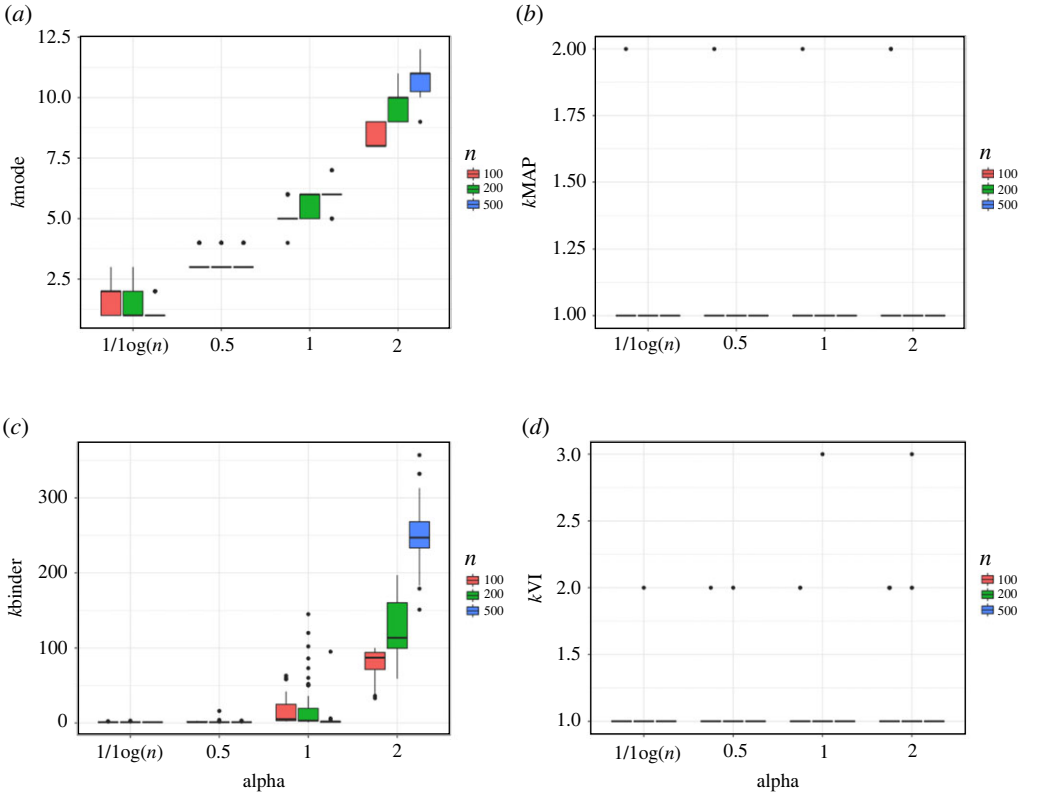

Figure 3. Comparison of different estimators for the number of clusters in the example of Miller & Harrison [145], where the true clustering contains only a single cluster. The DP mixture of Gaussians is considered for model-based clustering with different choices of the concentration parameter α. The box plots display variability in the estimates across the 50 replicated datasets, with colour corresponding to a sample size of $n = 100$, 200 or 500. (*a*) Marginal mode of k. (*b*) MAP clustering k. (*c*) Binder clustering k. (*d*) VI clustering k. (Online version in colour.)

obtained through a hierarchical clustering algorithm (with dissimilarity equal to one minus the posterior similarity), and then, we perform a greedy search [140] starting from this clustering for any possible further improvements.

Figure 3 compares the different estimators for the example of Miller & Harrison [145]. Focusing on the mode of the marginal posterior on k (figure 3*a*), we empirically observe the inconsistency shown by Miller & Harrison [145]. Results are also sensitive to the choice of α, and the sample-size dependent choice of $\alpha = 1/\log(n)$ empirically helps to mitigate this behaviour (as expected based on Ascolani *et al.* [149] and Ohn & Lin [150]). Instead, the MAP clustering (figure 3*b*) contains only a single cluster (as proved by Rajkowski [157]) and is robust to the choice of α. Depending on α, the DP mixture tends to create small extra clusters at each iteration. As discussed in Wade & Ghahramani [140], Binder's loss has a preference to split off small clusters over merging, and thus, when α is too large, the Binder clustering (figure 3*c*) extremely overestimates the number of clusters. The VI is a more symmetric metric in this regard, and therefore, the VI clustering (figure 3*d*) only contains a single cluster, in almost all replicates, that is also robust to the choice of α.

Results for the example of Rajkowski [157] are shown in figure 4. This is a misspecified example; while the true clustering under the uniform kernel contains only a single cluster, the DP mixture of Gaussians explores various arbitrary partitions of the data to approximate the uniform distribution. While Rajkowski [157] found that the MAP clustering also partitions the unit circle into several arbitrary sets when the within cluster variance is fixed and set too small,

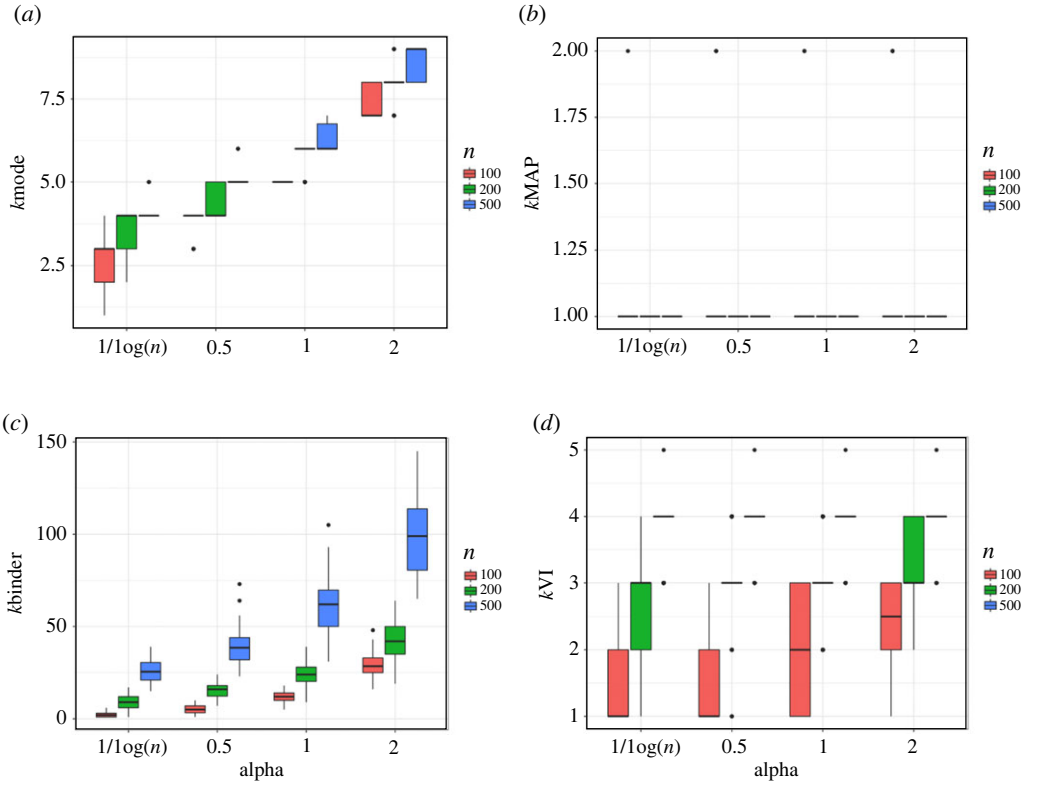

Figure 4. Comparison of different estimators for the number of clusters in the misspecified example of Rajkowski [157], where the true clustering contains only a single cluster under the uniform kernel. The DP mixture of Gaussians is considered for model-based clustering with different choices of the concentration parameter α. The box plots display variability in the estimates across the 50 replicated datasets, with colour corresponding to a sample size of $n = 100$, 200 or 500. (*a*) Marginal mode of k. (*b*) MAP clustering. (*c*) Binder clustering. (*d*) VI clustering. (Online version in colour.)

we, however, empirically observe a different behaviour when incorporating uncertainty on the within cluster variance. In fact, the MAP clustering (figure 4*b*) contains only a single cluster in almost all replicates and is robust to the choice of α. Again, the marginal posterior on k (figure 4*a*) and the Binder clustering (figure 4*c*) are quite sensitive to the value of α, with the Binder clustering extremely overestimating the number of clusters for larger n and α. The VI clustering (figure 4*d*) contains only a single clustering in some replications, and in others contains two to four clusters, particularly for larger sample sizes. The former can be explained by the fact that the VI solution is obtained by minimizing a function of the posterior similarity matrix, and as each posterior sample corresponds to an arbitrary partition of the data points into convex sets, each pair of data points may have a relatively high probability of being clustered together.

These examples highlight that the choice of estimator can greatly affect the number of clusters and clustering solution. In fact, while most asymptotic theory focuses on the behaviour of the marginal posterior on the number of clusters, quite a different behaviour is observed when estimating the full clustering solution. As practitioners are interested in the full clustering solution, this is an important aspect to consider. There are a number of interesting directions to expand this study, including further investigating the performance of the different estimators in the misspecified setting, as well as the case when the clusters are not well separated (Rajkowski [157] finds that MAP tends to underestimate the number clusters in this setting), and sensitivity to hyperparameters. Other estimators can be studied, e.g. Dahl *et al.* [171] develops generalized forms of Binder's loss and VI with unequal penalties, which provide more control over the estimated number of clusters but require specifying an additional parameter of the generalized

loss. Moreover, this study can be expanded by empirically comparing different models (MFM, sparse mixtures and infinite mixtures beyond the DP), as well as quantifying uncertainty, e.g. through empirical coverage of credible balls around the estimators [140]. Finally, while we have focused on general, commonly used estimators, it must be emphasized that in applications more problem-specific estimators should also be considered. This is achieved by defining an application-specific loss function in the decision-theoretic framework; examples include clinical trials [173,174] and earthquake studies [175].

5. Conclusion

The article contains an overview of Bayesian cluster analysis, which offers substantial benefits over algorithmic approaches by providing not only point estimates but also uncertainty in all parameters. More specifically, through the posterior over the clustering structure, an ensemble of clustering solutions is obtained. This ensemble and associated uncertainty can be visualized and described through various graphical tools and quantities, such as the posterior similarity matrix, credible balls, cluster comparison criterion [169] and stability indices [176]. The benefits are showcased in an application to cluster cells and discover latent cell types in scRNA-seq data to improve understanding of embryonic cell development [60].

We have provided a review of two approaches to Bayesian cluster analysis: model-based and loss-based. In both, careful consideration of the kernel or loss is emphasized for clustering solutions of practical relevance. The Bayesian paradigm requires specification of priors over the unknown parameters. Most often this includes the number of clusters, and a review of relevant approaches is given.

Lastly, we have focused on the ongoing debate between finite and infinite mixtures in a model-based approach and robustness to model misspecification. While much of the debate and asymptotic theory has focused on the marginal posterior of the number of clusters, we have empirically shown that quite a different behaviour is obtained when estimating the full clustering solution. As the full clustering solution is required in applications, the results highlight that more emphasis should be placed on this aspect. All models are misspecified in some way, and while careful consideration of the kernel in mixture models helps, robustness to misspecification should be acknowledged. Mixture models are inherently built for density estimation; if robust clustering methods are employed, optimal density estimation is sacrificed. Instead, our simple experiments highlight that mixtures models can still be employed, to retain optimal density estimation, with robust clustering via separation of the clustering problem in a decision-theoretic framework and careful consideration of the loss and estimators used. The MAP and VI clustering solutions are general and provide robust estimates in the examples presented, but in applications, more problem-specific estimators should also be explored [173].

Data accessibility. The data are provided in electronic supplementary material [177].

Conflict of interest declaration. We declare we have no competing interests.

Funding. Sara Wade is a Royal Society of Edinburgh (RSE) Sabbatical Research Grant Holder; this work was supported by the RSE under grant no. 69938.

Acknowledgements. I would like to thank David Dunson for his interesting lectures at the Bayesian Nonparametrics Networking Workshop held in Nicosia, Cyprus 2022, as well as the other lecturers Yanxun Xu and Aad van der Vaart, the organizers and participants; indeed some parts of this article were inspired by discussions at the workshop.

References

1. Cheeseman P, Kelly J, Self M, Stutz J, Taylor W, Freeman D. 1988 Autoclass: a Bayesian classification system. In *Machine learning proceedings 1988* (ed. J Laird), pp. 54–64. San Francisco, CA: Elsevier.
2. Kuhn MA, Feigelson ED. 2019 Applications in astronomy. In *Handbook of mixture analysis* (eds *S* Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 463–489. New York, NY: Chapman and Hall/CRC.

3. Dasgupta A, Raftery AE. 1998 Detecting features in spatial point processes with clutter via model-based clustering. *J. Am. Stat. Assoc.* **93**, 294–302. (doi:10.1080/01621459.1998.10474110)
4. Blei DM, Ng AY, Jordan MI. 2003 Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022.
5. Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. 2019 Evaluation of integrative clustering methods for the analysis of multi-omics data. *Brief. Bioinform.* **21**, 541–552. (doi:10.1093/bib/bbz015)
6. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, Achas M, Adebiyi E. 2016 Clustering algorithms: their application to gene expression data. *Bioinf. Biol. Insights* **10**, 237–253.
7. Titterington DM, Afm S, Smith AF, Makov U. 1985 *Statistical analysis of finite mixture distributions*. John Wiley & Sons Incorporated.
8. Jain AK, Dubes RC. 1988 *Algorithms for clustering data*. Prentice-Hall, Inc.
9. Kaufman L, Rousseeuw PJ. 1990 *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.
10. Hartigan J, Wong M. 1979 Algorithm AS 136: a *k*-means clustering algorithm. *J. R. Stat. Soc. C* **28**, 100–108.
11. Jain AK. 2010 Data clustering: 50 years beyond *k*-means. *Pattern Recognit. Lett.* **31**, 651–666. (doi:10.1016/j.patrec.2009.09.011)
12. Ester M, Kriegel H-P, Sander J, Xu X. 1996 A density-based algorithm for discovering clusters in large spatial databases with noise. In *Conf. on Knowledge Discovery and Data Mining, Portland, OR, 2–4 August 1996*, pp. 226–231. Washington DC: AAAI Press.
13. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. 2017 DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42**, 1–21. (doi:10.1145/3068335)
14. Fraley C, Raftery AE. 2002 Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **97**, 611–631. (doi:10.1198/016214502760047131)
15. Fruhwirth-Schnatter S, Celeux G, Robert CP. 2019 *Handbook of mixture analysis*. CRC Press.
16. McLachlan GJ, Peel D. 2004 *Finite mixture models*. John Wiley & Sons.
17. Hennig C. 2015 What are the true clusters? *Pattern Recognit. Lett.* **64**, 53–62. (doi:10.1016/j.patrec.2015.04.009)
18. Gabasova E, Reid J, Wernisch L. 2017 Clusternomics: integrative context-dependent clustering for heterogeneous datasets. *PLoS Comput. Biol.* **13**, e1005781. (doi:10.1371/journal.pcbi.1005781)
19. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. 2012 Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297. (doi:10.1093/bioinformatics/bts595)
20. Lock EF, Dunson DB. 2013 Bayesian consensus clustering. *Bioinformatics* **29**, 2610–2616. (doi:10.1093/bioinformatics/btt425)
21. Beraha M, Guglielmi A, Quintana FA. 2021 The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions. *Bayesian Anal.* **16**, 1187–1219. (doi:10.1214/21-BA1278)
22. Camerlenghi F, Dunson DB, Lijoi A, Prünster I, Rodríguez A. 2019 Latent nested nonparametric priors (with discussion). *Bayesian Anal.* **14**, 1303–1356. (doi:10.1214/19-BA1169)
23. Denti F, Camerlenghi F, Guindani M, Mira A. 2021 A common atoms model for the Bayesian nonparametric analysis of nested data. *J. Am. Stat. Assoc.* 1–12. (doi:10.1080/01621459.2021.1933499)
24. Lijoi A, Nipoti B, Prünster I. 2014 Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**, 1260–1291. (doi:10.3150/13-BEJ521)
25. Lijoi A, Prünster I, Rebaudo G. 2022 Flexible clustering via hidden hierarchical Dirichlet priors. *Scand. J. Stat.*
26. Müller P, Quintana F, Rosner G. 2004 A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. B* **66**, 735–749. (doi:10.1111/j.1467-9868.2004.05564.x)
27. Rodriguez A, Dunson D, Gelfand A. 2006 The nested Dirichlet process. *J. Am. Stat. Assoc.* **103**, 1131–1154.
28. Teh Y, Jordan M, Beal M, Blei D. 2006 Hierarchical Dirichlet process. *J. Am. Stat. Assoc.* **101**, 1566–1581. (doi:10.1198/016214506000000302)

29. Kaufmann S. 2019 Hidden Markov models in time series, with applications in economics. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 309–341. New York, NY: Chapman and Hall/CRC.
30. Maheu JM, Zamenjani AS. 2019 Applications in finance. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 407–437. New York, NY: Chapman and Hall/CRC.
31. Fernández C, Green PJ. 2002 Modelling spatially correlated data via mixtures: a Bayesian approach. *J. R. Stat. Soc. B* **64**, 805–826. (doi:10.1111/1467-9868.00362)
32. Forbes F. 2019 Mixture models for image analysis. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 385–405. New York, NY: Chapman and Hall/CRC.
33. Green PJ, Richardson S. 2002 Hidden Markov models and disease mapping. *J. Am. Stat. Assoc.* **97**, 1055–1070. (doi:10.1198/016214502388618870)
34. Gormley IC, Frühwirth-Schnatter S. 2019 Mixture of experts models. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 271–307. New York, NY: Chapman and Hall/CRC.
35. Masoudnia S, Ebrahimpour R. 2014 Mixture of experts: a literature survey. *Artif. Intell. Rev.* **42**, 275–293. (doi:10.1007/s10462-012-9338-y)
36. Müller P, Quintana F, Rosner GL. 2011 A product partition model with regression on covariates. *J. Comput. Graph. Stat.* **20**, 260–278.
37. Quintana FA, Müller P, Jara A, MacEachern SN. 2022 The dependent Dirichlet process and related models. *Stat. Sci.* **37**, 24–41. (doi:10.1214/20-STS819)
38. Bouveyron C, Celeux G, Murphy TB, Raftery AE. 2019 *Model-based clustering and classification for data science: with applications in R*, **vol. 50**. Cambridge, UK: Cambridge University Press.
39. Grün B. 2019 Model-based clustering. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 157–192. New York, NY: Chapman and Hall/CRC.
40. Kulis B, Jordan MI. 2012 Revisiting *k*-means: new algorithms via Bayesian nonparametrics. In *Proc. of the 29th Int. Conf. on Machine Learning, Edingburgh, UK, 27 June–1 July 2012*, pp. 1131–1138. Madison, WI: Omnipress.
41. Kurihara K, Welling M. 2009 Bayesian *k*-means as a 'Maximization-Expectation' algorithm. *Neural Comput.* **21**, 1145–1172. (doi:10.1162/neco.2008.12-06-421)
42. Banfield JD, Raftery AE. 1993 Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821. (doi:10.2307/2532201)
43. Ghahramani Z, Hinton GE. 1996 The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.
44. McLachlan GJ, Baek J, Rthnayake S. 2011 Mixtures of factor analysers for the analysis of high-dimensional data. In *Mixtures: estimation and application* (eds KL Mengersen, CP Robert, D Titterington), pp. 189–212. Hoboken, NJ: John Wiley & Sons.
45. Castelletti F, Consonni G. 2021 Bayesian graphical modelling for heterogeneous causal effects. (https://arxiv.org/abs/2106.03252).
46. Rodriguez A, Lenkoski A, Dobra A. 2011 Sparse covariance estimation in heterogeneous samples. *Electron. J. Stat.* **5**, 981.
47. Frühwirth-Schnatter S, Pyne S. 2010 Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-*t* distributions. *Biostatistics* **11**, 317–336.
48. Lee S, McLachlan GJ. 2014 Finite mixtures of multivariate skew *t*-distributions: some recent and new results. *Stat. Comput.* **24**, 181–202. (doi:10.1007/s11222-012-9362-4)
49. Franczak BC, Browne RP, McNicholas PD. 2013 Mixtures of shifted asymmetric Laplace distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1149–1157. (doi:10.1109/TPAMI.2013.216)
50. O'Hagan A, Murphy TB, Gormley IC, McNicholas PD, Karlis D. 2016 Clustering with the multivariate normal inverse Gaussian distribution. *Comput. Stat. Data Anal.* **93**, 18–30.
51. Peel D, Whiten WJ, McLachlan GJ. 2001 Fitting mixtures of Kent distributions to aid in joint set identification. *J. Am. Stat. Assoc.* **96**, 56–63. (doi:10.1198/016214501750332974)
52. Banerjee A, Dhillon IS, Ghosh J, Sra S, Ridgeway G. 2005 Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.* **6**, 1345–1382.
53. Straub J, Chang J, Freifeld O, Fisher III J. 2015 A Dirichlet process mixture model for spherical data. In *Proc. of the Eighteenth Int. Conf. on Artificial Intelligence and Statistics, San Diego, CA, 9–12 May 2015*, pp. 930–938. Cambridge, MA: PMLR.

54. Goodman LA. 1974 Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231. (doi:10.1093/biomet/61.2.215)
55. DeYoreo M, Kottas A. 2018 Bayesian nonparametric modeling for multivariate ordinal regression. *J. Comput. Graph. Stat.* **27**, 71–84. (doi:10.1080/10618600.2017.1316280)
56. Kottas A, Müller P, Quintana F. 2005 Nonparametric Bayesian modeling for multivariate ordinal data. *J. Comput. Graph. Stat.* **14**, 610–625. (doi:10.1198/106186005X63185)
57. Mollica C, Tardella L. 2017 Bayesian Plackett–Luce mixture models for partially ranked data. *Psychometrika* **82**, 442–458. (doi:10.1007/s11336-016-9530-0)
58. Karlis D, Xekalaki E. 2005 Mixed Poisson distributions. *Int. Stat. Rev.* **73**, 35–58. (doi:10.1111/j.1751-5823.2005.tb00250.x)
59. Krnjajić M, Kottas A, Draper D. 2008 Parametric and nonparametric Bayesian model specification: a case study involving models for count data. *Comput. Stat. Data Anal.* **52**, 2110–2128.
60. Liu J, Wade S, Bochkina N. 2022 Shared differential clustering across single-cell RNA sequencing datasets with the hierarchical Dirichlet process. ArXiv.
61. Canale A, Dunson DB. 2011 Bayesian kernel mixtures for counts. *J. Am. Stat. Assoc.* **106**, 1528–1539. (doi:10.1198/jasa.2011.tm10552)
62. Canale A, Prünster I. 2017 Robustifying Bayesian nonparametric mixtures for count data. *Biometrics* **73**, 174–184. (doi:10.1111/biom.12538)
63. Wu Q, Luo X. 2022 Nonparametric Bayesian two-level clustering for subject-level single-cell expression data. *Stat. Sin.* **32**, 1–22.
64. Cai J-H, Song X-Y, Lam K-H, Ip EH-S. 2011 A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Comput. Stat. Data Anal.* **55**, 2889–2907. (doi:10.1016/j.csda.2011.05.011)
65. Norets A, Pelenis J. 2020 Adaptive Bayesian estimation of mixed discrete-continuous distributions under smoothness and sparsity. *J. Econom.* **90**, 1355–1377.
66. Malsiner-Walli G, Frühwirth-Schnatter S, Grün B. 2017 Identifying mixtures of mixtures using Bayesian estimation. *J. Comput. Graph. Stat.* **26**, 285–295. (doi:10.1080/10618600.2016.1200472)
67. Stephenson BJ, Herring AH, Olshan A. 2019 Robust clustering with subpopulation-specific deviations. *J. Am. Stat. Assoc.* **115**, 521–537. (doi:10.1080/01621459.2019.1611583)
68. Celeux G, Kamary K, Malsiner-Walli G, Marin J-M, Robert CP. 2019 Computational solutions for Bayesian inference in mixture models. In H*andbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 73–96. New York, NY: Chapman and Hall/CRC.
69. Chandra NK, Canale A, Dunson DB. 2020 Escaping the curse of dimensionality in Bayesian model based clustering. (https://arxiv.org/abs/2006.02700).
70. Doo W, Kim H. 2021 Bayesian variable selection in clustering high-dimensional data via a mixture of finite mixtures. *J. Stat. Comput. Simul.* **91**, 2551–2568. (doi:10.1080/00949655.2021.1902526)
71. Kim S, Tadesse MG, Vannucci M. 2006 Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93**, 877–893. (doi:10.1093/biomet/93.4.877)
72. Tadesse MG, Sha N, Vannucci M. 2005 Bayesian variable selection in clustering high-dimensional data. *J. Am. Stat. Assoc.* **100**, 602–617. (doi:10.1198/016214504000001565)
73. White A, Wyse J, Murphy TB. 2016 Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler. *Stat. Comput.* **26**, 511–527. (doi:10.1007/s11222-014-9542-5)
74. Malsiner-Walli G, Frühwirth-Schnatter S, Grün B. 2016 Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26**, 303–324. (doi:10.1007/s11222-014-9500-2)
75. Yau C, Holmes C. 2011 Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Anal.* **6**, 329. (doi:10.1214/11-BA612)
76. Murphy K, Viroli C, Gormley IC. 2020 Infinite mixtures of infinite factor analysers. *Bayesian Anal.* **15**, 937–963. (doi:10.1214/19-BA1179)
77. McNicholas PD, Murphy TB. 2008 Parsimonious Gaussian mixture models. *Stat. Comput.* **18**, 285–296. (doi:10.1007/s11222-008-9056-0)
78. Iwata T, Duvenaud D, Ghahramani Z. 2013 Warped mixtures for nonparametric cluster shapes. In *Proc. of the Twenty-Ninth Conf. on Uncertainty in Artificial Intelligence, Bellevue, WA, 11–15 August 2013*, pp. 311–320. Portland, OR: AUAI Press.
79. Rigon T, Herring AH, Dunson DB. 2020 A generalized Bayes framework for probabilistic clustering. (https://arxiv.org/abs/2006.05451).

80. Bissiri PG, Holmes CC, Walker SG. 2016 A general framework for updating belief distributions. *J. R. Stat. Soc. B* **78**, 1103–1130. (doi:10.1111/rssb.12158)
81. Argiento R, Cremaschi A, Guglielmi A. 2014 A 'density-based' algorithm for cluster analysis using species sampling Gaussian mixture models. *J. Comput. Graph. Stat.* **23**, 1126–1142. (doi:10.1080/10618600.2013.856796)
82. Heller K, Ghahramani Z. 2005 Bayesian hierarchical clustering. In *Proc. of the 22nd Int. Conf. on Machine Learning, Bonn, Germany, 7–11 August 2005*, pp. 297–304. New York, NY: Association for Computing Machinery.
83. Knowles DA, Ghahramani Z. 2014 Pitman Yor diffusion trees for Bayesian hierarchical clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 271–289. (doi:10.1109/TPAMI.2014.2313115)
84. Neal RM. 2003 Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Stat.* **7**, 619–629.
85. Savage RS, Heller K, Xu Y, Ghahramani Z, Truman WM, Grant M, Denby KJ, Wild DL. 2009 R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC Bioinf.* **10**, 1–9. (doi:10.1186/1471-2105-10-242)
86. Dahl DB, Andros J, Carter JB. 2021 Cluster analysis via random partition distributions. (https://arxiv.org/abs/2106.02760).
87. Dahl DB, Day R, Tsai JW. 2017 Random partition distribution indexed by pairwise information. *J. Am. Stat. Assoc.* **112**, 721–732. (doi:10.1080/01621459.2016.1165103)
88. Duan LL, Dunson DB. 2021 Bayesian distance clustering. *J. Mach. Learn. Res.* **22**, 10 228–10 254.
89. Natarajan A, De Iorio M, Heinecke A, Mayer E, Glenn S. 2021 Cohesion and repulsion in Bayesian distance clustering. (https://arxiv.org/abs/2107.05414).
90. Celeux G, Frühwirth-Schnatter S, Robert CP. 2019 Model selection for mixture models–perspectives and strategies. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 117–154. New York, NY: Chapman and Hall/CRC.
91. Miller JW, Harrison MT. 2018 Mixture models with a prior on the number of components. *J. Am. Stat. Assoc.* **113**, 340–356. (doi:10.1080/01621459.2016.1255636)
92. Nobile A, Fearnside AT. 2007 Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Stat. Comput.* **17**, 147–162. (doi:10.1007/s11222-006-9014-7)
93. Richardson S, Green PJ. 1997 On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. B* **59**, 731–792. (doi:10.1111/1467-9868.00095)
94. Frühwirth-Schnatter S, Malsiner-Walli G, Grün B. 2021 Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Anal.* **16**, 1279–1307.
95. Rousseau J, Mengersen K. 2011 Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. B* **73**, 689–710. (doi:10.1111/j.1467-9868.2011.00781.x)
96. Van Havre Z, White N, Rousseau J, Mengersen K. 2015 Overfitting Bayesian mixture models with an unknown number of components. *PLoS ONE* **10**, e0131739. (doi:10.1371/journal.pone.0131739)
97. Müller P. 2019 Bayesian nonparametric mixture models. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 97–116. New York, NY: Chapman and Hall/CRC.
98. Lo A. 1984 On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.* **12**, 351–357.
99. Connor R, Mosimann J. 1969 Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Stat. Assoc.* **64**, 194–206. (doi:10.1080/01621459.1969.10500963)
100. Lijoi A, Prünster I, Rigon T. 2020 The Pitman-Yor multinomial process for mixture modelling. *Biometrika* **107**, 891–906. (doi:10.1093/biomet/asaa030)
101. Lee J, James LF, Choi S. 2016 Finite-dimensional BFRY priors and variational Bayesian inference for power law models. *Adv. Neural Inf. Process. Syst.* **29,** 3170–3178.
102. Argiento R, De Iorio M. 2022 Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Ann. Stat.* **50**, 2641–2663. (doi:10.1214/22-AOS2201)
103. Bernardo J, Girón F. 1988 A Bayesian analysis of simple mixture problems. *Bayesian Stat.* **3**, 67–78.

104. Ishwaran H, James L. 2001 Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* **96**, 161–173. (doi:10.1198/016214501750332758)
105. Sethuraman J. 1994 A constructive definition of Dirichlet priors. *Stat. Sin.* **4**, 639–650.
106. Giordano R, Liu R, Jordan MI, Broderick T. 2022 Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. *Bayesian Anal.* **1**, 1–34. (doi:10.1214/22-BA1309)
107. Pitman J. 1995 Exchangeable and partially exchangeable random partitions. *Probab. Theory Relat. Fields* **102**, 145–158. (doi:10.1007/BF01213386)
108. Ferguson T. 1973 A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230. (doi:10.1214/aos/1176342360)
109. De Blasi P, Favaro S, Lijoi A, Mena RH, Prünster I, Ruggiero M. 2013 Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 212–229. (doi:10.1109/TPAMI.2013.217)
110. Gnedin A, Pitman J. 2006 Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.* **138**, 5674–5685. (doi:10.1007/s10958-006-0335-z)
111. Pitman J, Yor M. 1997 The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900. (doi:10.1214/aop/1024404422)
112. Lee CJ, Sang H. 2022 Why the rich get richer? On the balancedness of random partition models. In *Proc. of the 39th Int. Conf. on Machine Learning, Baltimore, MD, 17–23 July 2022*, pp. 12 521–12 541. Cambridge, MA: PMLR.
113. Lu J, Li M, Dunson D. 2018 Reducing over-clustering via the powered Chinese restaurant process. (https://arxiv.org/abs/1802.05392).
114. Wallach H, Jensen S, Dicker L, Heller K. 2010 An alternative prior process for nonparametric Bayesian clustering. In *Proc. of the Thirteenth Int. Conf. on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010*, pp. 892–899. Cambridge, MA: PMLR.
115. Paganin S, Herring AH, Olshan AF, Dunson DB. 2021 Centered partition processes: informative priors for clustering (with discussion). *Bayesian Anal.* **16**, 301–370. (doi:10.1214/20-BA1197)
116. Smith AN, Allenby GM. 2019 Demand models with random partitions. *J. Am. Stat. Assoc.* **115**, 47–65. (doi:10.1080/01621459.2019.1604360)
117. Lijoi A, Prünster I. 2011 Models beyond the Dirichlet process. In *Bayesian nonparametrics* (eds N Hjort, C Holmes, P Müller, S Walker), pp. 80–136, Cambridge, UK: Cambridge University Press.
118. Rousseau J, Grazian C, Lee JE. 2019 Bayesian mixture models: theory and methods. In *Handbook of mixture analysis* (eds S Fruhwirth-Schnatter, G Celeux, CP Robert), pp. 53–72. New York, NY: Chapman and Hall/CRC.
119. Fraley C, Raftery AE. 2007 Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classif.* **24**, 155–181. (doi:10.1007/s00357-007-0004-5)
120. Diebolt J, Robert CP. 1994 Estimation of finite mixture distributions through Bayesian sampling. *J. R. Stat. Soc. B* **56**, 363–375.
121. Wasserman L. 2000 Asymptotic inference for mixture models by using data-dependent priors. *J. R. Stat. Soc. B* **62**, 159–180. (doi:10.1111/1467-9868.00226)
122. Jeffreys H. 1939 *The theory of probability*. Oxford, UK: Oxford University Press.
123. Grazian C, Robert CP. 2018 Jeffreys priors for mixture estimation: properties and alternatives. *Comput. Stat. Data Anal.* **121**, 149–163. (doi:10.1016/j.csda.2017.12.005)
124. Beraha M, Argiento R, Møller J, Guglielmi A. 2022 MCMC computations for Bayesian mixture models using repulsive point processes. *J. Comput. Graph. Stat.* **31**, 422–435. (doi:10.1080/10618600.2021.2000424)
125. Petralia F, Rao V, Dunson D. 2012 Repulsive mixtures. *Adv. Neural Inf. Process. Syst.* **25**, 1889–1897.
126. Xie F, Xu Y. 2020 Bayesian repulsive Gaussian mixture model. *J. Am. Stat. Assoc.* **115**, 187–203. (doi:10.1080/01621459.2018.1537918)
127. Xu Y, Müller P, Telesca D. 2016 Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics* **72**, 955–964. (doi:10.1111/biom.12482)
128. Fúquene J, Steel M, Rossell D. 2019 On choosing mixture components via non-local priors. *J. R. Stat. Soc. B* **81**, 809–837.
129. Chen C, Zhu J, Zhang X. 2014 Robust Bayesian max-margin clustering. *Adv. Neural Inf. Process. Syst.* **27**, 532–540.

130. Huang W, Ng TLJ, Laitonjam N, Hurley NJ. 2021 Posterior regularisation on Bayesian hierarchical mixture clustering. (https://arxiv.org/abs/2105.06903).
131. Kiselev V, Andrews T, Hemberg M. 2019 Challenges in unsupervised clustering of single-cell RNA-seq data. *Genetics* **20**, 273–282.
132. Petegrosso R, Li Z, Kuang R. 2020 Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Bioinformatics* **21**, 1209–1223.
133. Manuel MN *et al.* 2022 PAX6: limits the competence of developing cerebral cortical cells to respond to inductive intercellular signals. *PLoS Biol.* **20**, e3001563. (doi:10.1371/journal.pbio.3001563)
134. Prabhakaran S, Azizi E, Carr A, Pe'er D. 2016 Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *Int. Conf. on Machine Learning, New York, NY, 20–22 June 2016*, pp. 1070–1079. Cambridge, MA: PMLR.
135. Vallejos C, Risso D, Scialdone A, Dudoit S, Marioni JC. 2017 Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14**, 565–571. (doi:10.1038/nmeth.4292)
136. Duan T, Pinto JP, Xie X. 2019 Parallel clustering of single cell transcriptomic data with split-merge sampling on Dirichlet process mixtures. *Bioinformatics* **35**, 953–961. (doi:10.1093/bioinformatics/bty702)
137. Sun Z, Wang T, Deng K, Wang X-F, Lafyatis R, Ding Y, Hu M, Chen W. 2018 DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* **34**, 139–146. (doi:10.1093/bioinformatics/btx490)
138. Eling N, Richard AC, Richardson S, Marioni JC, Vallejos CA. 2018 Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell Syst.* **7**, 284–294. (doi:10.1016/j.cels.2018.06.011)
139. Van der Maaten L, Hinton G. 2008 Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605.
140. Wade S, Ghahramani Z. 2018 Bayesian cluster analysis: point estimation and credible balls (with discussion). *Bayesian Anal.* **13**, 559–626. (doi:10.1214/17-BA1073)
141. Ghosal S, Ghosh JK, Ramamoorthi R. 1999 Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Stat.* **27**, 143–158. (doi:10.1214/aos/1018031105)
142. Ghosal S, Ghosh JK, Van Der Vaart AW. 2000 Convergence rates of posterior distributions. *Ann. Stat.* **28**, 500–531.
143. Ghosal S, Van der Vaart A. 2017 *Fundamentals of nonparametric Bayesian inference*. Cambridge, UK: Cambridge University Press.
144. Wu Y, Ghosal S. 2010 The L_1-consistency of Dirichlet mixtures in multivariate density estimation. *J. Multivar. Anal.* **101**, 2411–2419. (doi:10.1016/j.jmva.2010.06.012)
145. Miller JW, Harrison MT. 2013 A simple example of Dirichlet process mixture inconsistency for the number of components. *Adv. Neural Inf. Process. Syst.* **26**, 199–206.
146. Miller JW, Harrison MT. 2014 Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15**, 3333–3370.
147. Guha A, Ho N, Nguyen X. 2021 On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli* **27**, 2159–2188. (doi:10.3150/20-BEJ1275)
148. Miller JW. 2022 Consistency of mixture models with a prior on the number of components. (https://arxiv.org/abs/2205.03384).
149. Ascolani F, Lijoi A, Rebaudo G, Zanella G. 2022 Clustering consistency with Dirichlet process mixtures. (https://arxiv.org/abs/2205.12924).
150. Ohn I, Lin L. 2020 Optimal Bayesian estimation of Gaussian mixtures with growing number of components. (https://arxiv.org/abs/2007.09284).
151. Frühwirth-Schnatter S, Malsiner-Walli G. 2019 From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Adv. Data Anal. Classif.* **13**, 33–64.
152. Kleijn BJ, van der Vaart AW. 2006 Misspecification in infinite-dimensional Bayesian statistics. *Ann. Stat.* **34**, 837–877. (doi:10.1214/009053606000000029)
153. Kleijn BJ, van der Vaart AW. 2012 The Bernstein-von-Mises theorem under misspecification. *Electron. J. Stat.* **6**, 354–381. (doi:10.1214/12-EJS675)

154. Cai D, Campbell T, Broderick T. 2021 Finite mixture models do not reliably learn the number of components. In *Int. Conf. on Machine Learning*, Online, 18–24 July 2021, pp. 1158–1169. Cambridge, MA: PMLR.
155. Miller JW, Dunson DB. 2018 Robust Bayesian inference via coarsening. *J. Am. Stat. Assoc.* **114**, 1113–1125. (doi:10.1080/01621459.2018.1469995)
156. Rodríguez CE, Walker SG. 2014 Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Stat. Comput.* **24**, 35–49.
157. Rajkowski Ł. 2019 Analysis of the maximal a posteriori partition in the Gaussian Dirichlet process mixture model. *Bayesian Anal.* **14**, 477–494. (doi:10.1214/18-BA1114)
158. Medvedovic M, Sivaganesan S. 2002 Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206. (doi:10.1093/bioinformatics/18.9.1194)
159. Medvedovic M, Yeung K, Bumgarner R. 2004 Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics* **20**, 1222–1232. (doi:10.1093/bioinformatics/bth068)
160. Molitor J, Papathomas M, Jerrett M, Richardson S. 2010 Bayesian profile regression with an application to the national survey of children's health. *Biostatistics* **11**, 484–498. (doi:10.1093/biostatistics/kxq013)
161. Rasmussen C, De la Cruz B, Ghahramani Z, Wild D. 2009 Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **6**, 615–628. (doi:10.1109/TCBB.2007.70269)
162. Dahl D. 2009 Modal clustering in a class of product partition models. *Bayesian Anal.* **4**, 243–264.
163. Heard N, Holmes C, Stephens D. 2006 A quantitative study of gene regulation involved in the immune response of anopheline mosquitos: an application of Bayesian hierarchical clustering of curves. *J. Am. Stat. Assoc.* **101**, 18–29. (doi:10.1198/016214505000000187)
164. Raykov YP, Boukouvalas A, Little MA. 2016 Simple approximate MAP inference for Dirichlet processes mixtures. *Electron. J. Stat.* **10**, 3548–3578. (doi:10.1214/16-EJS1196)
165. Fritsch A, Ickstadt K. 2009 Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **4**, 367–392. (doi:10.1214/09-BA414)
166. Lau J, Green P. 2007 Bayesian model-based clustering procedures. *J. Comput. Graph. Stat.* **16**, 526–558. (doi:10.1198/106186007X238855)
167. Quintana F, Iglesias P. 2003 Bayesian clustering and product partition models. *J. R. Stat. Soc. B* **65**, 557–574. (doi:10.1111/1467-9868.00402)
168. Binder D. 1978 Bayesian cluster analysis. *Biometrika* **65**, 31–38. (doi:10.1093/biomet/65.1.31)
169. Meilă M. 2007 Comparing clusterings—an information based distance. *J. Multivar. Anal.* **98**, 873–895.
170. Dahl D, Müller P. 2017 Sdols: summarizing distributions of latent structures. *R package version*, 1:591.
171. Dahl DB, Johnson DJ, Müller P. 2022 Search algorithms and loss functions for Bayesian clustering. *J. Comput. Graph. Stat.* **31**, 1189–1201. (doi:10.1080/10618600.2022.2069779)
172. Rastelli R, Friel N. 2018 Optimal Bayesian estimators for latent variable cluster models. *Stat. Comput.* **28**, 1169–1186. (doi:10.1007/s11222-017-9786-y)
173. Paulon G, Trippa L, Müller P. 2018 Invited comment on article by Wade and Ghahramani. *Bayesian Anal.* **13**, 559–626.
174. Schnell PM, Tang Q, Offen WW, Carlin BP. 2016 A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics* **72**, 1026–1036. (doi:10.1111/biom.12522)
175. Natvig B, Tvete IF. 2007 Bayesian hierarchical space–time modeling of earthquake data. *Methodol. Comput. Appl. Probab.* **9**, 89–114. (doi:10.1007/s11009-006-9008-0)
176. Koepke H, Clarke B. 2013 A Bayesian criterion for cluster stability. *Stat. Anal. Data Min.: ASA Data Sci. J.* **6**, 346–374. (doi:10.1002/sam.11176)
177. Wade S. 2023 Bayesian cluster analysis. Figshare. (doi:10.6084/m9.figshare.c.6423927)