THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Bayesian Cluster Analysis

Citation for published version:
Wade, SK 2023, 'Bayesian Cluster Analysis', *Philosophical Transactions A: Mathematical, Physical and Engineering Sciences*, vol. 381, no. 2247, 20220149.

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Philosophical Transactions A: Mathematical, Physical and Engineering Sciences

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 23. Nov. 2023

PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:
xxxxx, xxxxx, xxxx

Keywords:
Bayesian analysis, Clustering, Ensembles, Mixture models, Model misspecification

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

THE ROYAL SOCIETY PUBLISHING

Bayesian Cluster Analysis

S. Wade[1]

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

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.

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 (Cheeseman et al., 1988; Kuhn and Feigelson, 2019); geosciences to detect minefields or seismic faults from spatial data (Dasgupta and Raftery, 1998); natural language processing where topic models employ clustering to infer latent topics across documents for information retrieval (Blei et al., 2003); biomedicine to discover groups of individuals or genes with similar patterns in gene expression or omics data (Chauvel et al., 2019; Oyelade et al., 2016), 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

© The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

density estimation or regression (Titterington et al., 1985).

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 (Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990), and results crucially depend on the choice of dissimilarity and linkage. Partition-based algorithms, including k-means (Hartigan and Wong, 1979) and k-mediods (Kaufman and Rousseeuw, 1990), 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 (Jain, 2010), 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 (Ester et al., 1996; Schubert et al., 2017), 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 utilizes mixture models, where each (non-empty) mixture component corresponds to a cluster (Fraley and Raftery, 2002; Fruhwirth-Schnatter et al., 2019; McLachlan and Peel, 2004). 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 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 (Section 2), along with an illustrative application to highlight the advantages of the Bayesian approach in Section 3. In addition, Section 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 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,

$\pi(\boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{z})\pi(\boldsymbol{z})$, where $\pi(\boldsymbol{z})$ represents the prior over the space of clusterings and $p(\boldsymbol{y} \mid \boldsymbol{z})$ can be defined through a model-based (Section (a)) or a loss-based (Section (b)) 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 (Hennig, 2015). 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 optimum 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 combing multiple data sources through data integration (Gabasova et al., 2017; Kirk et al., 2012; Lock and Dunson, 2013), hierarchical Bayesian frameworks for partially-exchangeable or nested data (Beraha et al., 2021; Camerlenghi et al., 2019; Denti et al., 2021; Lijoi et al., 2014, 2022; Müller et al., 2004; Rodriguez et al., 2006; Teh et al., 2006), hidden Markov models and other extensions for temporal data (Kaufmann, 2019; Maheu and Zamenjani, 2019), accounting for spatially-indexed data (Fernández and Green, 2002; Forbes, 2019; Green and Richardson, 2002), incorporating general covariate information (Gormley and Frühwirth-Schnatter, 2019; Masoudnia and Ebrahimpour, 2014; Müller et al., 2011; Quintana et al., 2022), and more.

(a) Model-based approach

The most popular approach to Bayesian clustering employs a model-based framework through mixture models (Bouveyron et al., 2019; Grün, 2019). In this case, the data is assumed to be conditionally i.i.d. from a convex combination of parametric components:

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

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 $\boldsymbol{z}$:

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

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, $f(\cdot|\theta, \psi) = \text{N}(\cdot \mid \mu_j, \Sigma_j)$. 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 $\text{N}(\cdot \mid \mu_j, \sigma^2 I)$ (Kulis and Jordan, 2012; Kurihara and Welling, 2009). 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 centers 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 (Banfield and Raftery, 1993; Fraley and Raftery, 2002). Other

types of constraints on the covariance matrices can also be considered, such as mixtures of factor analysers (Ghahramani et al., 1996; McLachlan et al., 2011) and mixtures of Gaussian graphical models within a casual framework (Castelletti and Consonni, 2021; Rodriguez et al., 2011).

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 (Frühwirth-Schnatter and Pyne, 2010; Lee and McLachlan, 2014), shifted asymmetric Laplace distributions (Franczak et al., 2013), and normal inverse Gaussian distributions (O'Hagan et al., 2016). For directional data on the unit sphere, examples include mixtures of Kent distributions (Peel et al., 2001), von-Mises-Fisher distributions (Banerjee et al., 2005), or Gaussian distributions in distinct tangent spaces (Straub et al., 2015). For discrete data, mixtures of Bernoulli or multinomial distributions, known as latent class models, are appropriate for categorical data (Blei et al., 2003; Goodman, 1974); latent variable approaches using a logistic or probit transformation are employed for ordinal data (DeYoreo and Kottas, 2018; Kottas et al., 2005); and mixtures of Plackett-Luce models are used for rankings (Mollica and Tardella, 2017). For count data, examples include mixtures of Poisson distributions (Karlis and Xekalaki, 2005; Krnjajić et al., 2008), negative-binomial distributions (Liu et al., 2022), and rounded continuous kernels (Canale and Dunson, 2011; Canale and Prünster, 2017), as well as zero-inflated Poisson or negative binomial distributions for sparse counts (Wu and Luo, 2022). 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 (Cai et al., 2011; Norets and Pelenis, 2020). Moreover, the kernels may be themselves mixtures for increased flexibility (Malsiner-Walli et al., 2017; Stephenson et al., 2019).

In high-dimensional settings, challenges arise both from a computational (Celeux et al., 2019b) and theoretical (Chandra et al., 2020) perspective. In particular, Chandra et al. (2020) shows 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 (Doo and Kim, 2021; Kim et al., 2006; Tadesse et al., 2005; White et al., 2016) or shrinkage priors (Malsiner-Walli et al., 2016; Yau and Holmes, 2011). 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 (Murphy et al., 2020) or parsimonious Gaussian mixtures (McNicholas and Murphy, 2008), as well as approaches that conduct clustering directly on the lower dimensional space (Chandra et al., 2020). While approaches mainly focus on incorporating linear dimension reduction, extensions based on non-linear dimension reduction can also be considered (Iwata et al., 2013).

(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. (2020) employs a generalized Bayesian framework through the use of Gibbs posteriors (Bissiri et al., 2016). 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. (2020)). 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 (Argiento et al., 2014); Bayesian hierarchical clustering which builds a tree of hierarchical clustering solutions based on Bayesian nonparametric mixture models (Heller and Ghahramani, 2005; Knowles and Ghahramani, 2014; Neal, 2003; Savage et al., 2009); and Bayesian distance-based clustering based on pairwise distances between observations (Dahl et al., 2021, 2017; Duan and Dunson, 2021; Natarajan et al., 2021).

(c) Priors

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 $j \in \{1, \dots, J\}$. 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 criterion can be used to compare the mixture model under different choices of J (Celeux et al., 2019a); 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) (Miller and Harrison, 2018; Nobile and Fearnside, 2007; Richardson and Green, 1997) 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 (Frühwirth-Schnatter et al., 2021; Malsiner-Walli et al., 2016; Rousseau and Mengersen, 2011; Van Havre et al., 2015).
(iv) Bayesian nonparametric (BNP) mixtures (Müller, 2019) assume $J = \infty$ and can be viewed as a limiting case of overfitted mixtures, with the Dirichlet process (DP) mixture (Lo, 1984) 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.

Weights. To specify the prior on the weights, the standard choice for finite mixtures is the symmetric Dirichlet distribution, $(w_1, \dots, w_J) \sim \text{Dir}(\alpha, \dots, \alpha)$, 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 (Rousseau and Mengersen, 2011). The parameter α has an influential role, and Van Havre et al. (2015) 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 (Connor and Mosimann, 1969), multinomial Pitman-Yor process (Lijoi et al., 2020), BFRY priors (Lee et al., 2016), normalized jumps of a finite point process (Argiento and De Iorio, 2022), or non-informative Jeffreys priors (Bernardo and Girón, 1988).

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 (Ishwaran and James, 2001; Sethuraman, 1994), with a discussion on choice of hyperparameters in Giordano et al. (2022). Alternatively, the weights can be marginalized with a prior defined directly on the partition of data points into clusters. Exchangeable partition probability functions (EPPFs) (Pitman, 1995) 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 (Ferguson, 1973) 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 (De Blasi et al., 2013; Gnedin and Pitman, 2006), which contain the EPPF of the DP, Pitman-Yor (PY) process (Pitman and Yor, 1997), and MFM (Miller and Harrison, 2018) as special cases. Other proposals (Lee and Sang, 2022; Lu et al., 2018; Wallach et al., 2010) 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 centering around prior information on the clustering structure (Paganin et al., 2021; Smith and Allenby, 2019). 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 and Prünster (2011) for an overview of priors beyond the DP.

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. $\theta_j \overset{iid}{\sim} H_0$. 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 (Rousseau et al., 2019). For example, consider the Gaussian scale-location mixture with kernel $\mathrm{N}(y \mid \mu_j, \sigma_j^2)$. The conjugate prior is the normal-inverse gamma:

$$\mu_j \mid \sigma_j^2 \sim \mathrm{N}(\mu_0, \sigma_j^2/c), \quad \sigma_j^2 \sim \mathrm{IG}(\nu/2, \delta^2/2).$$

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 and Raftery (2007) 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 and Raftery (2007) suggest a value of $c = 0.01$, or assigned a hyperprior. Other examples of data-dependent priors can be found in Diebolt and Robert (1994); Richardson and Green (1997); Wasserman (2000). 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 (Jeffreys,

1939) for the atoms often lead to improper posteriors, non-informative priors can be combined with hierarchical hyperpriors to produce proper posteriors (Grazian and Robert, 2018).

In order to favour components that are well-separated, the independence assumption on the atoms can be relaxed through the use of repulsive priors (Beraha et al., 2022; Petralia et al., 2012; Xie and Xu, 2020), determinantal point processes (Xu et al., 2016), or non-local priors (Fúquene et al., 2019). In particular, this form of prior regularization helps to improve interpretation and encourages more meaningful clustering structures, however 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 (Chen et al., 2014; Huang et al., 2021) to ensure well-separated clusters.

3. Example: discovering cell subtypes

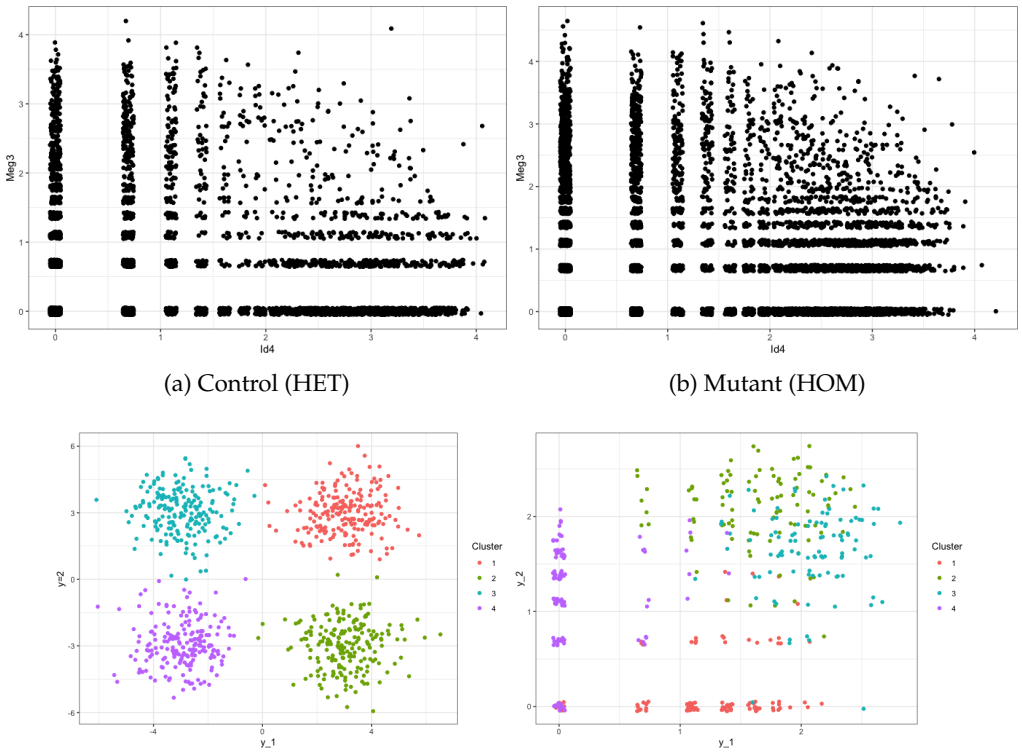
The rise in single-cell RNA sequencing (scRNA-seq) technology allows researchers to go beyond bulk RNA measurements and understand gene expression patterns at the single-cell level. Cells are heterogeneous in nature, and with existing technology able to record measurements on thousands of cells across thousands of genes, clustering has become an important tool to characterize latent cell types with similar expression patterns and summarize the data (Kiselev et al., 2019; Petegrosso et al., 2020).

To highlight the advantages of Bayesian cluster analysis, we consider an experimental scRNA-seq datasets (Manuel et al., 2022)[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 (Liu et al., 2022), 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 Section 2(a), 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 is 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. (2016); Vallejos et al. (2017), 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. (2022) follows more recent proposals which integrate normalization and clustering in a combined model-based framework (Duan et al., 2019; Prabhakaran et al., 2016; Sun et al., 2018; Wu and Luo, 2022). 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

[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.



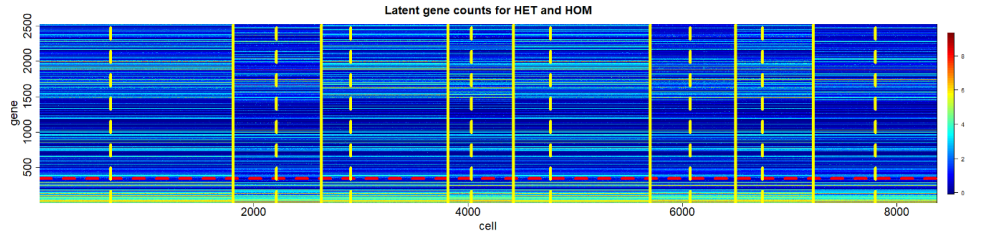
(a) Control (HET) (b) Mutant (HOM)

(c) Simulated data from GMM (d) Simulated data from NBMM

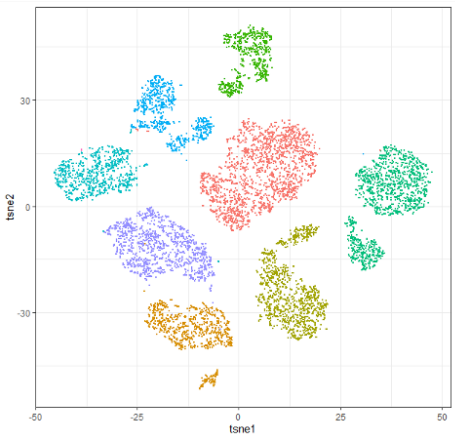
Figure 1: In order to highlight limitations of the standard workflow for scRNA-seq data, which firsts log-transforms data and then applies tools, such as k-means for clustering, we plot in (a) and (b) the log-transformed counts across all cells for two genes, Id4 and Meg3, and in (c) data simulated from 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.

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 $\text{NB}(y \mid \beta\mu_j, \phi_j)$, 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 (Eling et al., 2018). Borrowing of strength and shared clustering across the mutant and control conditions is permitted through a hierarchical Bayesian framework, namely the hierarchical Dirichlet process (Teh et al., 2006). For full details, see (Liu et al., 2022).

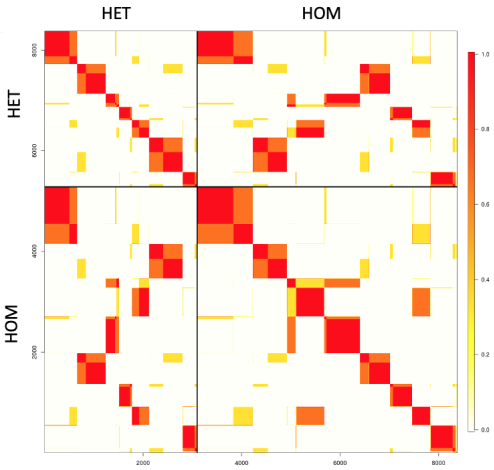
The Bayesian approach permits us to produce a range of graphical tools and tables to visualize and summarize not only point estimates but also uncertainty in the clustering structure and all parameters. The posterior estimated latent counts (corrected by the posterior capture efficiencies) are shown in the top of Figure 2. 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 (bottom left of Figure 2). In addition to the clustering estimate, we can also visualize uncertainty in the clustering structure, for example through the *posterior similarity matrix*, whose



(a) Estimated latent counts



(b) Clustering estimate



(c) Uncertainty in clustering

Figure 2: Highlights of the analysis of (Liu et al., 2022). Top: 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. Bottom left: visualization of the clustering estimate in the two-dimensional space obtained through t-distributed stochastic neighbor embedding (t-SNE Van der Maaten and Hinton, 2008) of the high-dimensional data. Bottom right: uncertainty in clustering characterized by the posterior similarity matrix.

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 and Ghahramani (2018). 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 (Liu et al., 2022).

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 is collected. With advancements in computing and general inference schemes such as Markov chain Monte Carlo

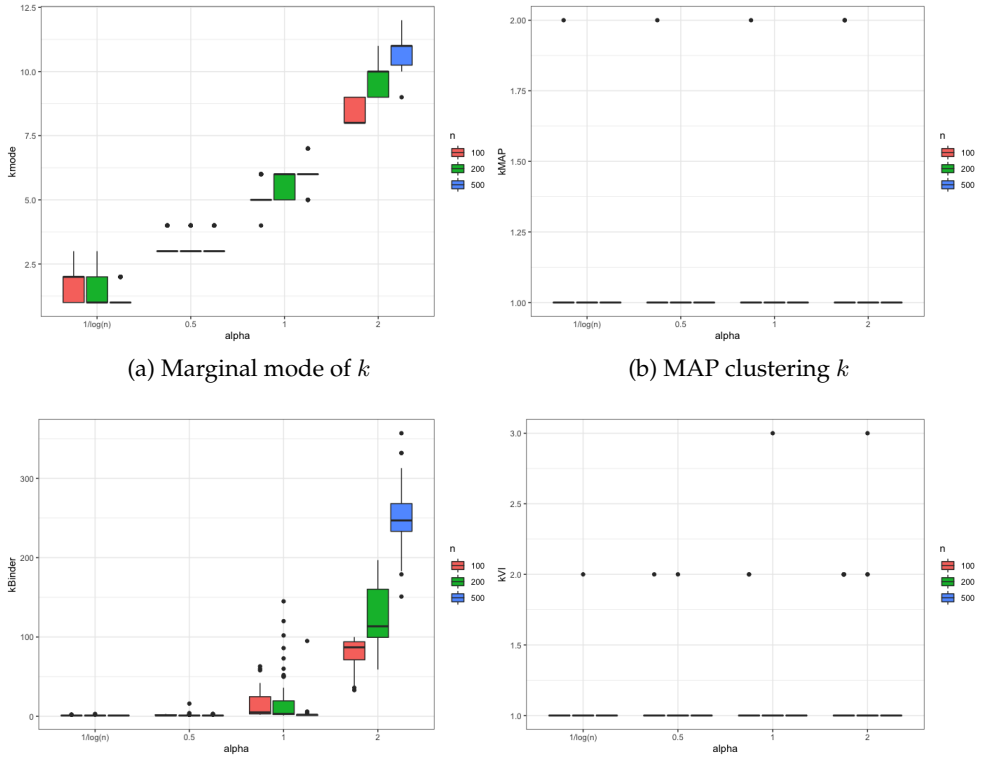
(MCMC), BNP mixtures can be easily implemented. Moreover, well-established theory validates the use of BNP mixtures for asymptotically optimal density estimation (Ghosal et al., 1999, 2000; Ghosal and Van der Vaart, 2017; Wu and Ghosal, 2010). Together these properties and developments have led to the huge growth and adoption of BNP mixtures, especially DP mixtures, for a variety of applications in statistics and machine learning in the 21st century.

However, this enthusiasm was dampened by the negative results of Miller and Harrison (2013, 2014) 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 (Rousseau et al., 2019; Rousseau and Mengersen, 2011), and MFM which yield consistent estimates for the number of components (Guha et al., 2021; Miller, 2022). 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 (Guha et al., 2021). Indeed, it is well-known that DP mixtures can introduce many small extra clusters. To overcome this, Guha et al. (2021) 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 (Ascolani et al., 2022; Ohn and Lin, 2020). Furthermore, Frühwirth-Schnatter and Malsiner-Walli (2019) 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 prior's support) (Kleijn and van der Vaart, 2006, 2012). However, Guha et al. (2021) 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. (2021) 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 (Miller and Dunson, 2018) as well as modal-based clustering (Rodríguez and Walker, 2014). 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 (2019) demonstrates that the MAP clustering has desirable asymptotic properties in the simple example of Miller and Harrison (2013), 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 (Medvedovic and



(a) Marginal mode of k (b) MAP clustering k

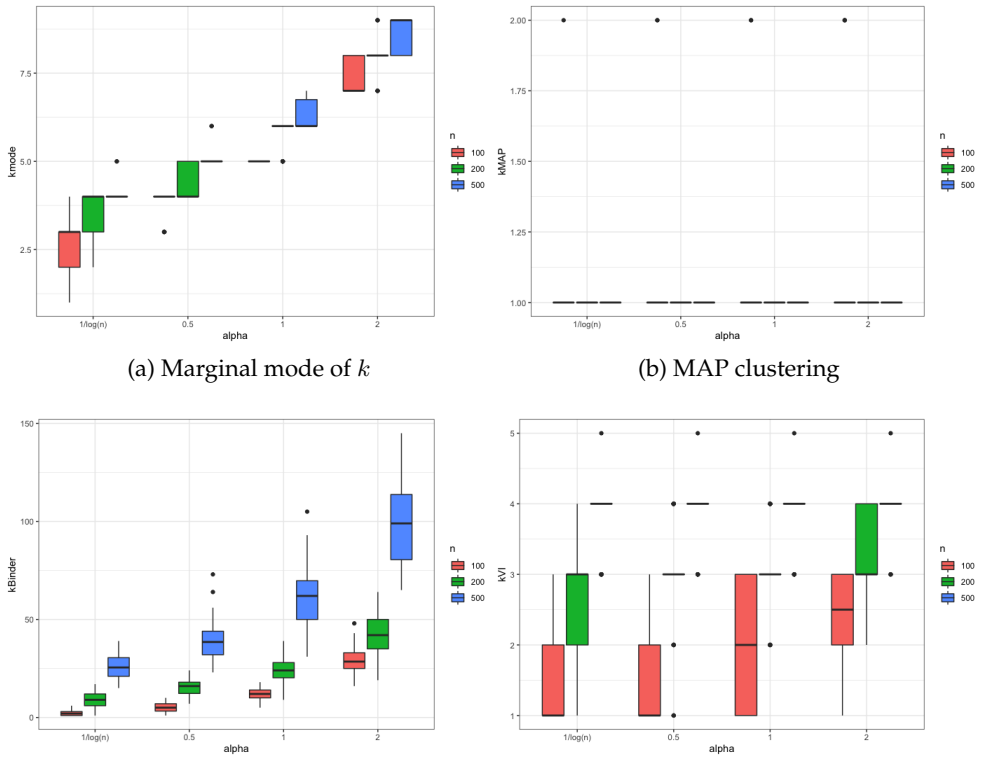
(c) Binder clustering k (d) VI clustering k

Figure 3: Comparison of different estimators for the number of clusters in the example of Miller and Harrison (2013), 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.

Sivaganesan, 2002; Medvedovic et al., 2004; Molitor et al., 2010; Rasmussen et al., 2009). 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 (Dahl, 2009; Heard et al., 2006; Heller and Ghahramani, 2005; Raykov et al., 2016). Alternative loss functions were considered in Fritsch and Ickstadt (2009); Lau and Green (2007); Quintana and Iglesias (2003); Wade and Ghahramani (2018). Two widely used loss functions, which are considered below, are Binder's loss[2] (Binder, 1978) and the variation of information (VI) (Meilă, 2007). General algorithms to optimize the posterior expected loss can be found in (Lau and Green, 2007; Wade and Ghahramani, 2018), with more recent schemes in Dahl and Müller (2017); Dahl et al. (2022); Rastelli and Friel (2018) 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 1) a standard normal distribution and 2) 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 and Harrison (2013) corresponds to 1) and the marginal posterior on the number of clusters was demonstrated to be severely inconsistent. Rajkowski

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



(a) Marginal mode of k (b) MAP clustering

(c) Binder clustering (d) VI clustering

Figure 4: Comparison of different estimators for the number of clusters in the misspecified example of Rajkowski (2019), 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.

(2019) 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 $\alpha = 0.5, 1, 2$ 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 obtained through a hierarchical clustering algorithm (with dissimilarity equal to one minus the posterior similarity), and then, we perform a greedy search (Wade and Ghahramani, 2018) starting from this clustering for any possible further improvements.

Figure 3 compares the different estimators for the example of Miller and Harrison (2013). Focusing on the mode of the marginal posterior on k (Figure 3a), we empirically observe the inconsistency shown by Miller and Harrison (2013). 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. (2022); Ohn and Lin (2020)). Instead, the MAP clustering (Figure 3b) contains only a single cluster (as proved by Rajkowski (2019)) 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 and Ghahramani (2018), Binder's loss has a preference to split off small clusters over merging, and thus, when α is too large, the Binder clustering (Figure 3c) extremely overestimates the number of clusters. The VI is a more symmetric metric in this regard, and therefore, the VI clustering (Figure 3d) only contains a single cluster, in almost all replicates, that is also robust to the choice of α.

Results for the example of Rajkowski (2019) 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 (2019) 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, we however empirically observe a different behaviour when incorporating uncertainty on the within cluster variance. In fact, the MAP clustering (Figure 4b) contains only a single cluster in almost all replicates and is robust to the choice of α. Again, the marginal posterior on k (Figure 4a) and the Binder clustering (Figure 4c) 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 4d) 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 effect 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 (2019) 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. (2022) 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 (mixtures of finite mixtures, sparse mixtures, and infinite mixtures beyond the DP), as well as quantifying uncertainty, e.g. through empirical coverage of credible balls around the estimators (Wade and Ghahramani, 2018). 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 (Paulon et al., 2018; Schnell et al., 2016) and earthquake studies (Natvig and Tvete, 2007).

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 (Meilă, 2007), and stability indices (Koepke and Clarke, 2013). 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 (Liu et al., 2022).

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 (Paulon et al., 2018).

Competing Interests. The author declares that she has 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 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

Argiento, R., Cremaschi, A., and Guglielmi, A. (2014). A "density-based" algorithm for cluster analysis using species sampling Gaussian mixture models. *Journal of Computational and Graphical Statistics*, 23(4):1126–1142.

Argiento, R. and De Iorio, M. (2022). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Annals of Statistics*. To appear.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2022). Clustering consistency with Dirichlet process mixtures. *arXiv preprint arXiv:2205.12924*.

Banerjee, A., Dhillon, I. S., Ghosh, J., Sra, S., and Ridgeway, G. (2005). Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(9).

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821.

Beraha, M., Argiento, R., Møller, J., and Guglielmi, A. (2022). MCMC computations for Bayesian mixture models using repulsive point processes. *Journal of Computational and Graphical Statistics*, pages 1–14.

Beraha, M., Guglielmi, A., and Quintana, F. A. (2021). The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions. *Bayesian Analysis*, 16(4):1187–1219.

Bernardo, J. and Girón, F. (1988). A Bayesian analysis of simple mixture problems. *Bayesian Statistics*, 3(3):67–78.

Binder, D. (1978). Bayesian cluster analysis. *Biometrika*, 65:31–38.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Cai, D., Campbell, T., and Broderick, T. (2021). Finite mixture models do not reliably learn the number of components. In *International Conference on Machine Learning*, pages 1158–1169.

Cai, J.-H., Song, X.-Y., Lam, K.-H., and Ip, E. H.-S. (2011). A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Computational Statistics & Data Analysis*, 55(11):2889–2907.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019). Latent nested nonparametric priors (with discussion). *Bayesian Analysis*, 14(4):1303–1356.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539.

Canale, A. and Prünster, I. (2017). Robustifying Bayesian nonparametric mixtures for count data. *Biometrics*, 73(1):174–184.

Castelletti, F. and Consonni, G. (2021). Bayesian graphical modelling for heterogeneous causal effects. *arXiv preprint arXiv:2106.03252*.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. (2019a). Model selection for mixture models–perspectives and strategies. In *Handbook of Mixture Analysis*, pages 117–154. Chapman and Hall/CRC.

Celeux, G., Kamary, K., Malsiner-Walli, G., Marin, J.-M., and Robert, C. P. (2019b). Computational solutions for Bayesian inference in mixture models. In *Handbook of Mixture Analysis*, pages 73–96. Chapman and Hall/CRC.

Chandra, N. K., Canale, A., and Dunson, D. B. (2020). Escaping the curse of dimensionality in Bayesian model based clustering. *arXiv preprint arXiv:2006.02700*.

Chauvel, C., Novoloaca, A., Veyre, P., Reynier, F., and Becker, J. (2019). Evaluation of integrative clustering methods for the analysis of multi-omics data. *Briefings in Bioinformatics*, 21(2):541–552.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D. (1988). Autoclass: A Bayesian classification system. In *Machine Learning Proceedings 1988*, pages 54–64. Elsevier.

Chen, C., Zhu, J., and Zhang, X. (2014). Robust Bayesian max-margin clustering. *Advances in Neural Information Processing Systems*, 27.

Connor, R. and Mosimann, J. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64:194–206.

Dahl, D. (2009). Modal clustering in a class of product partition models. *Bayesian Analysis*, 4:243–264.

Dahl, D. and Müller, P. (2017). sdols: Summarizing distributions of latent structures. *R package version*, 1:591.

Dahl, D. B., Andros, J., and Carter, J. B. (2021). Cluster analysis via random partition distributions. *arXiv preprint arXiv:2106.02760*.

Dahl, D. B., Day, R., and Tsai, J. W. (2017). Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association*, 112(518):721–732.

Dahl, D. B., Johnson, D. J., and Müller, P. (2022). Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, pages 1–13.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, 93(441):294–302.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229.

Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2021). A common atoms model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association*, pages 1–12.

DeYoreo, M. and Kottas, A. (2018). Bayesian nonparametric modeling for multivariate ordinal regression. *Journal of Computational and Graphical Statistics*, 27(1):71–84.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B*, 56(2):363–375.

Doo, W. and Kim, H. (2021). Bayesian variable selection in clustering high-dimensional data via a mixture of finite mixtures. *Journal of Statistical Computation and Simulation*, 91(12):2551–2568.

Duan, L. L. and Dunson, D. B. (2021). Bayesian distance clustering. *J. Mach. Learn. Res.*, 22:224–1.

Duan, T., Pinto, J. P., and Xie, X. (2019). Parallel clustering of single cell transcriptomic data with split-merge sampling on Dirichlet process mixtures. *Bioinformatics*, 35(6):953–961.

Eling, N., Richard, A. C., Richardson, S., Marioni, J. C., and Vallejos, C. A. (2018). Correcting the mean-variance dependency for differential variability testing using single-cell RNA sequencing data. *Cell systems*, 7(3):284–294.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.

Fernández, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society: Series B*, 64(4):805–826.

Forbes, F. (2019). Mixture models for image analysis. In *Handbook of Mixture Analysis*, pages 385–405. Chapman and Hall/CRC.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2013). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1149–1157.

Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4:367–392.

Fruhwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2019). *Handbook of Mixture Analysis*. CRC Press.

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). From here to infinity: sparse finite versus Dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13(1):33–64.

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279–1307.

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2):317–336.

Fúquene, J., Steel, M., and Rossell, D. (2019). On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society: Series B*, 81(5):809–837.

Gabasova, E., Reid, J., and Wernisch, L. (2017). Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. *PLoS Computational Biology*, 13(10).

Ghahramani, Z., Hinton, G. E., et al. (1996). The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2022). Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. *Bayesian Analysis*, 1(1):1–34.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Gormley, I. C. and Frühwirth-Schnatter, S. (2019). Mixture of experts models. *Handbook of Mixture Analysis*, pages 271–307.

Grazian, C. and Robert, C. P. (2018). Jeffreys priors for mixture estimation: Properties and alternatives. *Computational Statistics & Data Analysis*, 121:149–163.

Green, P. J. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070.

Grün, B. (2019). Model-based clustering. In *Handbook of mixture analysis*, pages 157–192. Chapman and Hall/CRC.

Guha, A., Ho, N., and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188.

Hartigan, J. and Wong, M. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C*, 28:100–108.

Heard, N., Holmes, C., and 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. *Journal of the American Statistical Association*, 101:18–29.

Heller, K. and Ghahramani, Z. (2005). Bayesian hierarchical clustering. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 297–304.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, 64:53–62.

Huang, W., Ng, T. L. J., Laitonjam, N., and Hurley, N. J. (2021). Posterior regularisation on Bayesian hierarchical mixture clustering. *arXiv preprint arXiv:2105.06903*.

Ishwaran, H. and James, L. (2001). Gibbs campling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173.

Iwata, T., Duvenaud, D., and Ghahramani, Z. (2013). Warped mixtures for nonparametric cluster shapes. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 311–320.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Cata*. Prentice-Hall, Inc.

Jeffreys, H. (1939). *The Theory of Probability*. Oxford University Press.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 35–58.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Kaufmann, S. (2019). Hidden Markov models in time series, with applications in economics. In *Handbook of Mixture Analysis*, pages 309–341. Chapman and Hall/CRC.

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4):877–893.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24):3290–3297.

Kiselev, V., Andrews, T., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Genetics*, 20:273–282.

Kleijn, B. J. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Kleijn, B. J. and van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381.

Knowles, D. A. and Ghahramani, Z. (2014). Pitman Yor diffusion trees for Bayesian hierarchical clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):271–289.

Koepke, H. and Clarke, B. (2013). A Bayesian criterion for cluster stability. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(4):346–374.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics*, 14(3):610–625.

Krnjajić, M., Kottas, A., and Draper, D. (2008). Parametric and nonparametric Bayesian model specification: A case study involving models for count data. *Computational Statistics & Data Analysis*, 52(4):2110–2128.

Kuhn, M. A. and Feigelson, E. D. (2019). Applications in astronomy. In *Handbook of Mixture Analysis*, pages 463–489. Chapman and Hall/CRC.

Kulis, B. and Jordan, M. I. (2012). Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1131–1138.

Kurihara, K. and Welling, M. (2009). Bayesian k-means as a "Maximization-Expectation" algorithm. *Neural Computation*, 21(4):1145–1172.

Lau, J. and Green, P. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16:526–558.

Lee, C. J. and Sang, H. (2022). Why the rich get richer? On the balancedness of random partition models. In *Proceedings of the 39th International Conference on Machine Learning*.

Lee, J., James, L. F., and Choi, S. (2016). Finite-dimensional BFRY priors and variational Bayesian inference for power law models. *Advances in Neural Information Processing Systems*, 29.

Lee, S. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202.

Lijoi, A., Nipoti, B., and Prünster, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260 – 1291.

Lijoi, A. and Prünster, I. (2011). Models beyond the Dirichlet process. In Hjort, N., Holmes, C., Müller, P., and Walker, S., editors, *Bayesian Nonparametrics*, pages 80–136, Cambridge, UK. Cambridge University Press.

Lijoi, A., Prünster, I., and Rebaudo, G. (2022). Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics*.

Lijoi, A., Prünster, I., and Rigon, T. (2020). The Pitman-Yor multinomial process for mixture modelling. *Biometrika*, 107(4):891–906.

Liu, J., Wade, S., and Bochkina, N. (2022). Shared differential clustering across single-cell RNA sequencing datasets with the hierarchical Dirichlet process. arXiv.

Lo, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12:351–357.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

Lu, J., Li, M., and Dunson, D. (2018). Reducing over-clustering via the powered Chinese restaurant process. *arXiv preprint arXiv:1802.05392*.

Maheu, J. M. and Zamenjani, A. S. (2019). Applications in finance. In *Handbook of Mixture Analysis*, pages 407–437. Chapman and Hall/CRC.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, 26(1):303–324.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, 26(2):285–295.

Manuel, M. N., Tan, K. B., Kozic, Z., Molinek, M., Marcos, T. S., Razak, M. F. A., Dobolyi, D., Dobie, R., Henderson, B., Hendserson, N., Chan, W. K., Daw, M., Mason, J., and Price, D. (2022). PAX6: limits the competence of developing cerebral cortical cells to respond to inductive intercellular signals. *PLOS Biology*.

Masoudnia, S. and Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293.

McLachlan, G. J., Baek, J., and Rthnayake, S. (2011). Mixtures of factor analysers for the analysis of high-dimensional data. *Mixtures: Estimation and application*, pages 189–212.

McLachlan, G. J. and Peel, D. (2004). *Finite Mixture Models*. John Wiley & Sons.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18:1194–1206.

Medvedovic, M., Yeung, K., and Bumgarner, R. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20:1222–1232.

Meilă, M. (2007). Comparing clusterings – an information based distance. *J. Multivar. Anal.*, 98:873–895.

Miller, J. W. (2022). Consistency of mixture models with a prior on the number of components. *arXiv preprint arXiv:2205.03384*.

Miller, J. W. and Dunson, D. B. (2018). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114:1113–1125.

Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *Advances in Neural Information Processing Systems*, 26.

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15(1):3333–3370.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). Bayesian profile regression with an application to the national survey of children's health. *Biostatistics*, 11:484–498.

Mollica, C. and Tardella, L. (2017). Bayesian Plackett–Luce mixture models for partially ranked data. *Psychometrika*, 82(2):442–458.

Müller, P. (2019). Bayesian nonparametric mixture models. *Handbook of Mixture Analysis*, pages 97–116.

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B*, 66(3):735–749.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278.

Murphy, K., Viroli, C., and Gormley, I. C. (2020). Infinite mixtures of infinite factor analysers. *Bayesian Analysis*, 15(3):937–963.

Natarajan, A., De Iorio, M., Heinecke, A., Mayer, E., and Glenn, S. (2021). Cohesion and repulsion in Bayesian distance clustering. *arXiv preprint arXiv:2107.05414*.

Natvig, B. and Tvete, I. F. (2007). Bayesian hierarchical space–time modeling of earthquake data. *Methodology and Computing in Applied Probability*, 9(1):89–114.

Neal, R. M. (2003). Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629.

Nobile, A. and Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162.

Norets, A. and Pelenis, J. (2020). Adaptive Bayesian estimation of mixed discrete-continuous distributions under smoothness and sparsity. *Journal of Econometrics*.

Ohn, I. and Lin, L. (2020). Optimal Bayesian estimation of Gaussian mixtures with growing number of components. *arXiv preprint arXiv:2007.09284*.

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., and Adebiyi, E. (2016). Clustering algorithms: their application to gene expression data. *Bioinformatics and Biology Insights*, 10:237–253.

O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P. D., and Karlis, D. (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics & Data Analysis*, 93:18–30.

Paganin, S., Herring, A. H., Olshan, A. F., and Dunson, D. B. (2021). Centered partition processes: Informative priors for clustering (with discussion). *Bayesian Analysis*, 16(1):301–370.

Paulon, G., Trippa, L., and Müller, P. (2018). Invited comment on Article by Wade and Ghahramani. *Bayesian Analysis*, 13(2):559–626.

Peel, D., Whiten, W. J., and McLachlan, G. J. (2001). Fitting mixtures of Kent distributions to aid in joint set identification. *Journal of the American Statistical Association*, 96(453):56–63.

Petegrosso, R., Li, Z., and Kuang, R. (2020). Machine learning and statistical methods for clustering single-cell RNA-sequencing data. *Bioinformatics*, 21(4):1209–1223.

Petralia, F., Rao, V., and Dunson, D. (2012). Repulsive mixtures. *Advances in Neural Information Processing Systems*, 25.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

Prabhakaran, S., Azizi, E., Carr, A., and Pe'er, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, pages 1070–1079. PMLR.

Quintana, F. and Iglesias, P. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B*, 65:557–574.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). The dependent Dirichlet process and related models. *Statistical Science*, 37(1):24–41.

Rajkowski, Ł. (2019). Analysis of the maximal a posteriori partition in the Gaussian Dirichlet process mixture model. *Bayesian Analysis*, 14(2):477–494.

Rasmussen, C., De la Cruz, B., Ghahramani, Z., and Wild, D. (2009). Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:615–628.

Rastelli, R. and Friel, N. (2018). Optimal Bayesian estimators for latent variable cluster models. *Statistics and Computing*, 28(6):1169–1186.

Raykov, Y. P., Boukouvalas, A., and Little, M. A. (2016). Simple approximate MAP inference for Dirichlet processes mixtures. *Electronic Journal of Statistics*, 10(2):3548–3578.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B*, 59(4):731–792.

Rigon, T., Herring, A. H., and Dunson, D. B. (2020). A generalized Bayes framework for probabilistic clustering. *arXiv preprint arXiv:2006.05451*.

Rodriguez, A., Dunson, D., and Gelfand, A. (2006). The nested Dirichlet process. *Journal of the American Statistical Association*, 103:1131–1154.

Rodriguez, A., Lenkoski, A., and Dobra, A. (2011). Sparse covariance estimation in heterogeneous samples. *Electronic Journal of Statistics*, 5:981.

Rodríguez, C. E. and Walker, S. G. (2014). Univariate Bayesian nonparametric mixture modeling with unimodal kernels. *Statistics and Computing*, 24(1):35–49.

Rousseau, J., Grazian, C., and Lee, J. E. (2019). Bayesian mixture models: Theory and methods. In *Handbook of Mixture Analysis*, pages 53–72. Chapman and Hall/CRC.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B*, 73(5):689–710.

Savage, R. S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W. M., Grant, M., Denby, K. J., and Wild, D. L. (2009). R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC bioinformatics*, 10(1):1–9.

Schnell, P. M., Tang, Q., Offen, W. W., and Carlin, B. P. (2016). A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*, 72(4):1026–1036.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Smith, A. N. and Allenby, G. M. (2019). Demand models with random partitions. *Journal of the American Statistical Association*.

Stephenson, B. J., Herring, A. H., and Olshan, A. (2019). Robust clustering with subpopulation-specific deviations. *Journal of the American Statistical Association*.

Straub, J., Chang, J., Freifeld, O., and Fisher III, J. (2015). A Dirichlet process mixture model for spherical data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 930–938.

Sun, Z., Wang, T., Deng, K., Wang, X.-F., Lafyatis, R., Ding, Y., Hu, M., and Chen, W. (2018). DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics*, 34(1):139–146.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet process. *Journal of the American Statistical Association*, 101:1566–1581.

Titterington, D. M., Afm, S., Smith, A. F., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons Incorporated.

Vallejos, C., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

Van Havre, Z., White, N., Rousseau, J., and Mengersen, K. (2015). Overfitting Bayesian mixture models with an unknown number of components. *PloS One*, 10(7).

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626.

Wallach, H., Jensen, S., Dicker, L., and Heller, K. (2010). An alternative prior process for nonparametric Bayesian clustering. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 892–899.

Wasserman, L. (2000). Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society: Series B*, 62(1):159–180.

White, A., Wyse, J., and Murphy, T. B. (2016). Bayesian variable selection for latent class analysis using a collapsed Gibbs sampler. *Statistics and Computing*, 26(1):511–527.

Wu, Q. and Luo, X. (2022). Nonparametric Bayesian two-level clustering for subject-level single-cell expression data. *Statistica Sinica*, 32:1–22.

Wu, Y. and Ghosal, S. (2010). The L_1-consistency of Dirichlet mixtures in multivariate density estimation. *Journal of Multivariate Analysis*, 101:2411–2419.

Xie, F. and Xu, Y. (2020). Bayesian repulsive Gaussian mixture model. *Journal of the American Statistical Association*, 115(529):187–203.

Xu, Y., Müller, P., and Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics*, 72(3):955–964.

Yau, C. and Holmes, C. (2011). Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, 6(2):329.