Cohesion and Repulsion in Bayesian Distance Clustering

Abhinav Natarajan
University of Oxford

Maria De Iorio
National University of Singapore

Andreas Heinecke
Yale-NUS College

Emanuel Mayer
Yale-NUS College

Simon Glenn
University of Oxford

Abstract

Clustering in high-dimensions poses many statistical challenges. While traditional distance-based clustering methods are computationally feasible, they lack probabilistic interpretation and rely on heuristics for estimation of the number of clusters. On the other hand, probabilistic model-based clustering techniques often fail to scale and devising algorithms that are able to effectively explore the posterior space is an open problem. Based on recent developments in Bayesian distance-based clustering, we propose a hybrid solution that entails defining a likelihood on pairwise distances between observations. The novelty of the approach consists in including both cohesion and repulsion terms in the likelihood, which allows for cluster identifiability. This implies that clusters are composed of objects which have small *dissimilarities* among themselves (cohesion) and similar dissimilarities to observations in other clusters (repulsion). We show how this modelling strategy has interesting connection with existing proposals in the literature. The proposed method is computationally efficient and applicable to a wide variety of scenarios. We demonstrate the approach in simulation and an application in digital numismatics. Supplementary Material with code is available online.

Keywords: Bayesian high-dimensional clustering; microclustering; digital numismatics; likelihood without likelihood; random partition models; composite likelihood.

Abhinav Natarajan is a doctoral student at the Mathematical Institute, University of Oxford, United Kingdom OX1 2JD (E-mail: *natarajan@maths.ox.ac.uk*). Maria De Iorio is Professor at Yong Loo Lin School of Medicine, National University of Singapore, Singapore 138527, Professor of Biostatistics at University College London and Principal Investigator at A*STAR, Singapore 138632 (E-mail: *mdi@nus.edu.sg*). Andreas Heinecke is Assistant Professor of Mathematics at Yale-NUS College, Singapore 138527 (E-mail: *andreas.heinecke@yale-nus.edu.sg*). Emanuel Mayer is Associate Professor of History at Yale-NUS College, Singapore 138527 (E-mail: *emanuel.mayer@yale-nus.edu.sg*). Simon Glenn is Research Fellow at the Ashmolean Museum, University of Oxford, United Kingdom OX1 2PH (E-mail: *simon.glenn@ashmus.ox.ac.uk*).

1 Introduction

Multidimensional clustering has been a fruitful line of research in statistics for a long time. The surge in the availability of data in recent years poses new challenges to clustering methods and the scalability of the associated computational algorithms, particularly in high dimensions. There are two main classes of clustering methods: those based on probabilistic models (model-based clustering), and constructive approaches based on *dissimilarities* between observations (distance-based clustering). The first class of methods includes popular tools such as mixture models (Mclachlan and Basford, 1988; Dasgupta and Raftery, 1998), product partition models (PPMs) (Hartigan, 1990; Barry and Hartigan, 1992), and nonparametric models like the Dirichlet process or more general species sampling models (Pitman, 1996; Ishwaran and James, 2003). An overview can be found in the article by Quintana (2006). The second class of methods includes the popular hierarchical clustering, k-means, and its variants like k-medoids.

Distance-based clustering algorithms, although computationally accessible and scalable to high dimensions, are often less interpretable, and do not quantify clustering uncertainty because of the lack of a probabilistic foundation. Out-of-sample prediction is challenging with these algorithms, and inference on the number of clusters relies on heuristics such as the elbow method. Moreover, there are theoretical limitations to the results produced by any distance-based clustering algorithm; in particular, they cannot simultaneously satisfy constraints about scale-invariance and consistency while also exploring all possible partitions (Kleinberg, 2002). On the other hand the drawbacks of model-based clustering methods are their analytic intractability and computational burden arising when working with high dimensional observations. To add to this, a fundamental difficulty with both types of clustering methods is that there is no consensus on what constitutes a true cluster (Hennig, 2015), and that the aims of clustering should be application-specific.

The focus of this paper is high dimensional clustering, in particular when point-wise evaluation of the likelihood is computationally intractable and posterior inference is infeasible. Our approach builds on recent proposals by Duan and Dunson (2021) and Rigon *et al.* (2023) that bridge the gap between model-based and distance-based clustering. The main idea behind this research is to specify a probability model on the distances between observations instead of the observations themselves, reducing a multidimensional problem to a low-dimensional one. An early reference for Bayesian clustering based on distances can be found in Lau and Green (2007).

Let $\rho_n = \{C_1, \ldots, C_K\}$ denote a partition of the set $[n] = \{1, \ldots, n\}$, and let $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ be a set of observations in $\mathbb{R}^l$. It is convenient to represent a clustering through cluster allocation indicators z_i, where $z_i = j$ when $i \in C_j$. Rigon *et al.* (2023) reformulate the clustering problem in terms of decision theory. They show that a large class of distance-based clustering methods based on loss-functions, including k-means and k-medoids, are equivalent to maximum a posteriori estimates in a probabilistic model with appropriately defined likelihood on the distances. Explicitly, they consider product partition models where the likelihood decomposes into cluster-wise *cohesions*

$$\pi(\mathbf{X} \mid \lambda, \rho_n) = \prod_{k=1}^{K} \exp\left(-\lambda \sum_{i \in C_k} D(x_i, C_k)\right)$$

where $D(x_i, C_k)$ measures the dissimilarity of observation i from cluster C_k and λ is a parameter that controls the posterior dependence on the distances between observations. A major drawback in this approach is that the number of clusters K must be pre-specified, whereas K is an object of inference in many practical scenarios. Inference on K is also problematic in the method proposed by Duan and Dunson (2021). This is due to identifiability issues that arise when working with distances. The starting point of their approach is an overfitted mixture model. By noting that in high dimensions the contribution of the cluster centres to the likelihood is negligible compared to the contribution from pairwise distances within the cluster, they specify a *partial likelihood* on the pairwise distances between observations

$$\pi(\mathbf{X} \mid \rho_n, \alpha, \beta) = \prod_{k=1}^{K} \prod_{i,j \in C_k} g(d(x_i, x_j); \alpha, \beta)^{1/n_k}$$

where g is a Gamma(α, β) density and n_k is the size of the kth cluster. Although this approach allows for estimation of K, it often relies on the specification of the maximum number of clusters in the sense that the clustering allocation significantly changes with this parameter.

We propose a model for high-dimensional clustering based on pairwise distances that combines cluster-wise cohesions with a *repulsive* term that imposes a strong identifiability constraint in the likelihood by penalising clusters that are not well-separated. To this end, we borrow ideas from machine learning such as the cross-cluster penalty in the calculation of a silhouette coefficient, and from the literature on repulsive distributions. The idea of repulsive distributions has been previously studied in the context of mixture models (Petralia *et al.*, 2012; Quinlan *et al.*, 2017; Xu *et al.*, 2016) to separate the location and scale parameters of the mixture kernels. We discuss the connection of repulsive distributions to our model in more detail in Section 2.1.

There are other instances of model-based clustering methods which exploit pairwise distances for cluster estimation. One example is the framework of Voronoi tesselations, a partition strategy that has found application in Bayesian statistics and partition models (Denison and Holmes, 2001; Møller and Øivind Skare, 2001; Corander *et al.*, 2008). In this approach a set of centres is sampled from a prior and the sample space is partitioned into the associated Voronoi cells. When the centres are chosen from the observations themselves, the implied prior on partitions depends on the pairwise distances between the observations. In the Bayesian random partition model literature, there have been various proposals to include covariate information in cluster allocation probabilities. Most notably, Müller *et al.* (2011) use a *similarity function* defined on sets of covariates belonging to all experimental units from a given cluster to modify the cohesions of a product partition model. Their similarity function is the marginal density of the covariates from an auxiliary probability model, which can also be interpreted as the marginal density on the distances of the covariates from a latent centre in the auxilliary probability space. This approach incorporates information about the distances between covariates into the partition prior. In high-dimensional settings, i.e., when the number of covariates is large, the covariate information dominates the clustering and the influence of the response is relatively inconsequential. See for example the work by Barcella *et al.* (2017). Alternatively, Dahl (2008); Dahl *et al.* (2017) propose random partition models through different modifications of the Dirichlet process cluster allocation probability: in the first case of the full conditional $\Pr(z_i \mid z_{-i})$, and in the second case on the sequential conditional probabilities $\Pr(z_i \mid z_1, \ldots, z_{i-1})$.

All these methods are linked through the use of pairwise dissimilarities, often in the form of distances, to define a partition prior for flexible Bayesian modelling. Here, we use the same strategy to define the likelihood on pairwise distances while using standard partition priors such as the Dirichlet process or the recently proposed *microclustering priors* (Zanella *et al.*, 2016; Betancourt *et al.*, 2022). In this respect, our model is strongly related to composite likelihood methods which will be discussed in Section 2.1.

The paper is structured as follows. In Section 2 we introduce the model and the computational strategy. In Section 3 we apply the proposed methodology to a problem from digital numismatics. We conclude the paper in Section 4. In Supplementary Material we present details of the computational algorithm, extensive simulation studies, and further results from the data application.

2 Model

In this section we describe the pairwise distance-based likelihood, and we present a justification for our modelling approach. The proposed strategy can accommodate different partition priors. In particular, we discuss a microclustering prior (Betancourt *et al.*, 2022) as it is the most relevant for our application in digital numismatics. We conclude the section with a discussion on the choice of hyperparameters, and of the MCMC algorithm.

2.1 Likelihood specification

We specify the likelihood on pairwise distances between observations instead of directly on the observations. This strategy falls naturally into the framework of composite likelihood. In its most general form, a composite likelihood is obtained by multiplying together a collection of component functions, each of which is a valid conditional or marginal density (Lindsay, 1988). The utility of composite likelihoods is in their computational tractability when a full likelihood is difficult to specify or computationally challenging to work with. In this context, the working assumption is the conditional independence of the individual likelihood components. Key examples of composite likelihood approaches include pseudolikelihood methods for approximate inference in spatial processes (Besag, 1975), posterior inference in population genetics models (Li and Stephens, 2003; Larribe and Fearnhead, 2011), pairwise difference likelihood and maximum composite likelihood in the analysis of dependence structure (Lele and Taper, 2002), and the use of independence loglikelihood for inference on clustered data (Chandler and Bate, 2007). See Varin *et al.* (2011) for an overview. Other approaches to overcome likelihood intractability include specifying the likelihood on summary statistics of the data (Beaumont *et al.*, 2002), or comparing simulated data from the model with the observed data (Fearnhead and Prangle, 2012). Both these ideas underlie *approximate Bayesian computation* (Marjoram *et al.*, 2003).

Combining ideas from composite likelihood methods and distance-based clustering, our strategy is to specify a likelihood on the distances that decomposes into a contribution from within-cluster distances and cross-cluster distances:

$$\pi(\mathbf{D} \mid \theta, \lambda, \rho_n) = \left[\prod_{k=1}^{K} \prod_{\substack{i,j \in C_k \\ i<j}} f(D_{ij} \mid \lambda_k) \right] \left[\prod_{(k,t) \in A} \prod_{\substack{i \in C_k \\ j \in C_t}} g(D_{ij} \mid \theta_{kt}) \right] \tag{1}$$

where $\mathbf{D} = [d(x_i, x_j)]_{ij}$ is the matrix of all pairwise distances, $A = \{(k,t) : 1 \leq k < t \leq K\}$, and f and g are probability densities. Note that this formulation does not result in a valid probability model on the data, but rather on a space $\mathcal{X}$ that is obtained as follows: let G be the group of isometries of $\mathbb{R}^l$ (with respect to the chosen distance metric), and let $H = \{(g, \dots, g) \in G^n : g \in G\}$ be the diagonal subgroup of G^n. Then $\mathcal{X}$ is the orbit space $\mathbb{R}^{l \times n}/H$ (for a reference see Klaus 1995). In Section 2.3 we discuss the choice of f and g in Equation (1). The first term in Equation (1) is similar to the cohesions of Duan and Dunson (2021); Rigon *et al.* (2023) and quantifies how similar the observations within each cluster are to each other; we call this the *cohesive* part of the likelihood.

The second multiplicative term in the likelihood, which we call the *repulsive* term, is related to the idea of repulsive mixtures. Typical mixture models associate with each cluster a location parameter ϕ_j, and these are assumed to be i.i.d. from a fixed prior. Petralia *et al.* (2012), and Quinlan *et al.* (2017) relax the i.i.d. assumption and use a repulsive joint prior of the form

$$\pi(\phi) \propto \prod_{i,j} h(d(\phi_i, \phi_j)) \tag{2}$$

where d is a distance measure and h decays to 0 for small values of its input. They do this to penalise clusters that are too close to each other, inducing parsimony. We generalise this idea by using a repulsive distribution on the observations themselves, i.e., by setting g in Equation (1) to a density that decays as its input approaches 0. This form of repulsion is important to our application because it encourages the formation of clusters from points that are not only close to each other but also have similar distances to points in other clusters. Moreover the repulsion allows for cross-clustering distances of different magnitude for different pairs of clusters. Consequently, this strategy allows for estimation of the number of clusters. Using repulsion on the observations instead of the cluster centres is also a viable strategy when the location parameters are not of interest or when posterior inference on the location parameters is computationally difficult, as is usually the case in high-dimensional clustering (Johnstone and Titterington, 2009). In doing so, we relax the assumption of conditional independence between clusters given their cluster-specific parameters. In Supplementary Material we further investigate the role of the repulsion term, showing its importance in identifying the number of clusters. This is consistent with the work by Fúquene *et al.* (2019) that shows that repulsion leads to faster learning of K in model-based settings.

2.2 Posterior Uncertainty

The distance based likelihood in Equation (1) has sharper peaks and flatter tails than the model-based likelihood from the raw data, as is typical in composite likelihood or pseudo-likelihood frameworks due to the artificial independence assumption. Intuitively, our model assumes $O(n^2)$ independent pieces of information whereas the data generation process may only produce $O(n)$ independent pieces of information. Consequently, well-separated clusters are associated to high posterior probability with corresponding underestimation of uncertainty, while poorly separated clusters will often be split into smaller clusters due to the artificially increased uncertainty. Although the estimation of uncertainty is inaccurate, localisation of modes in the posterior is satisfactory (as is typically the case for composite likelihood methods) and the model is able to provide some measure of uncertainty even when direct approaches fail to recover the clustering structure. This is demonstrated in our simulations in Supplementary Material. Moreover, our model can be used to guide more direct model-based approaches with better prior information. We believe that the major drawback of our approach is the choice of distance, as we remark in the discussion section. Finally we note that a common strategy in composite likelihood models to counteract the underestimation of uncertainty is to artificially flatten the likelihood, raising it to the power $1/n$. We cannot employ the same approach as this would flatten both the within-cluster terms and the cross-cluster terms, making them overlap significantly and making the clusters unidentifiable. We have nevertheless tried this approach, and as expected we obtained poor inference results (not shown).

2.3 Choice of distance densities

The likelihood in Equation (1) results in a monotonically decreasing density on the within-cluster distances if the cohesive term is chosen as the exponential of a loss function, as suggested by Rigon *et al.* (2023). This choice might be too restrictive in application, as more flexible distributions are required to accommodate the complexity in the data. As a consequence of such a restrictive choice, more dispersed clusters may be broken up into smaller clusters. Rigon *et al.* (2023) alleviate this problem by fixing the number of clusters. We instead propose a more flexible choice of f and g motivated by the following commonly-encountered scenario.

Assume that the original data have a multivariate Normal distribution such that each cluster is defined by a Normal kernel

$$y_i \mid z_i = k, \mu_k, \sigma_k^2 \sim \mathcal{N}(\mu_k, \sigma_k^2 I_l)$$

Hence the within-cluster differences are distributed as $\mathcal{N}(0, 2\sigma_k^2 I_l)$ and the corresponding squared Euclidean distances have a Gamma$(l/2, 1/(2\sigma_k^2))$ distribution. On the other hand, inter-cluster squared distances are distributed as a 3-parameter non-central χ^2:

$$g\left(\|x_i - x_j\|_2^2 \mid z_i = k, z_j = t, k \neq t, \theta_{kt} = (\theta_{kt}^{(1)}, \theta_{kt}^{(2)})\right) = \\ \sum_{m=0}^{\infty} \frac{\left(\theta_{kt}^{(1)}\right)^m \exp(-\theta_{kt}^{(1)})}{m!} h\left(\|x_i - x_j\|_2^2; l/2 + m, \theta_{kt}^{(2)}\right)$$

where $\theta_{kt}^{(1)}$ is the noncentrality parameter and corresponds to the squared distance between the cluster centres μ_k, $\theta_{kt}^{(2)}$ is a scale parameter related to the within-cluster variances of the two clusters, l is the dimension of the original data, and $h(\cdot; a, b)$ is a Gamma(a, b) density. This setup would cover many real world applications, but posterior inference on the parameters of a non-central χ^2 is unnecessarily complicated. Moreover the non-central χ^2 is defined on the squared Euclidean distance, which will lead to a non-central χ distribution on the distances. Indeed when we know that the data generation process coincides with a Normal mixture, we should use the correct distribution but this is not often the case. As such when the data generating process is unknown we work with Gamma distributions on the distances mainly for computational convenience. We propose setting f to be a Gamma(δ_1, λ_k) as in Duan and Dunson (2021):

$$f(x \mid \lambda_k) = \frac{\lambda_k^{\delta_1} x^{\delta_1 - 1} \exp(-\lambda_k x)}{\Gamma(\delta_1)}$$

where x is a pairwise distance and δ_1 is a fixed shape parameter that controls the cluster dispersion. When $\delta_1 < 1$, f is a monotonically decreasing density. We set g in Equation (1) to be a Gamma(δ_2, θ_{kt}) density, where $\delta_2 > 1$ is a fixed shape parameter that controls the shape of the decay of g towards the origin. We note that the constraint $\delta_2 > 1$ ensures that g is a repulsive density by forcing $g(0 \mid \theta_{kt}) = 0$. An appropriate choice of δ_1 and δ_2 is application-specific, and we discuss possible alternatives in Section 2.4.

In our experiments we find that using a Gamma distribution directly on the distances does not have an appreciable effect on posterior inference, suggesting that the methodology is robust. This strategy is also followed by Duan and Dunson (2021) but for different reasons.

2.4 Prior specification

Here we discuss the choice of priors for the cluster-specific parameters and the partition.

2.4.1 Prior Cluster-Specific Parameters

For computational convenience, we choose conjugate Gamma priors $\lambda_k \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$ and $\theta_{kt} \overset{\text{iid}}{\sim} \text{Gamma}(\zeta, \gamma)$. To set $\delta_1, \delta_2, \alpha, \beta, \zeta$ and γ, we follow a procedure in the spirit of empirical Bayes methods, as straightforward application of empirical Bayes is hindered by an often flat marginal likelihood of the parameters in question. We summarise our method in Algorithm 1.

Algorithm 1 Choosing $\alpha, \beta, \zeta, \gamma, \delta_1$, and δ_2

1. Compute a heuristic initial value for K, say K_{elbow}, via the elbow method, fitting k-means clustering for a range of values of K and with the within-cluster-sum-of-squares (WSS) score as the objective function.

2. Use k-means clustering with K_{elbow} to obtain an initial clustering configuration.

3. Split the pairwise distances into two groups A and B that correspond to the within-cluster and inter-cluster distances in this initial configuration.

4. Fit a Gamma distribution to the values in A using maximum likelihood estimation and set δ_1 to be the shape parameter of this distribution.

5. Set $\alpha = \delta_1 n_A$ and $\beta = \sum_{a \in A} a$, where n_A is the cardinality of the set A. This corresponds to the conditional posterior of λ obtained by specifying an improper prior $\pi(\lambda) \propto I(\lambda > 0)$ and treating A as a weighted set of observations from a Gamma(δ_1, λ) distribution.

6. Repeat steps 4 and 5 to obtain values for δ_2, ζ and γ by considering the values in B.

The range for K in step 1 of the algorithm can be chosen to be quite broad, for example from one to $n - 1$. When only pairwise distances or dissimilarities are available and not the raw data, k-medoids and the within-cluster-sum-of-dissimilarities can be used instead.

The proposed method depends on the choice of K obtained by the elbow method. In Supplementary Material we show that posterior inference is robust to the choice of K obtained, as long as this choice lies within a sensible range. We also propose an alternative method to fit the prior that results in a mixture prior on possible values for K.

2.4.2 Prior on partitions

The model can accommodate any prior on partitions of the observations, which is equivalent to specifying a prior on the partitions of $[n] = \{1, \ldots, n\}$. Let $\rho_n = \{C_1, \ldots, C_K\}$ denote a partition of $[n]$ where the C_j are pairwise disjoint and $K \leq n$. A common choice is to use a product partition model (PPM) as the prior for ρ_n; see for example the paper by Hartigan (1990) or Barry and Hartigan (1992). In a PPM there is a non-negative function $c(C_j)$, usually referred to as a *cohesion function*, which is used to define the prior

$$\Pr(\rho_n) = M \prod_{j=1}^{K} c(C_j)$$

where M is a normalising constant. This prior includes as special cases the Dirichlet Process (Quintana and Iglesias, 2003) as well as Gibbs-type priors. Alternatively one can consider the implied prior on partitions derived from a species sampling model (Pitman, 1996); in this case it can be shown that $\Pr(\rho_n = \{C_1, \ldots, C_K\}) = p(n_1, \ldots, n_K)$ where $n_j = |C_j|$ is the number of elements in C_j and p is a symmetric function of its arguments called the *exchangeable partition probability function* (EPPF).

In our application, we opt for a prior that has the microclustering property (Miller *et al.*, 2015; Zanella *et al.*, 2016; Betancourt *et al.*, 2022); that is, cluster sizes grow sublinearly in the number of observations n. This property is appropriate for die analysis in numismatics where each die is represented by a very limited number of samples. We use a class of random partition models described in Betancourt *et al.* (2022) called *Exchangeable Sequence of Clusters (ESC)*. In this model a generative process gives rise to a prior on partitions, which we describe briefly. A random distribution ν is drawn from the set $\mathcal{P}$ of distributions on the positive integers; ν is distributed according to P_ν. The cluster sizes n_j are sampled from ν, conditional upon the following event

$$E_n = \left\{ \text{there exists } K \in \mathbb{N} \text{ such that } \sum_{j=1}^{K} n_j = n \right\}.$$

We require that $\nu(1) > 0$ for all $\nu \in \mathcal{P}$ to ensure that $\Pr(E_n \mid \nu) > 0$ for all ν. A random partition with cluster sizes $\{n_1, \ldots, n_K\}$ is drawn by allocating cluster labels from a uniform permutation of

$$(\underbrace{1, \ldots, 1}_{n_1 \text{ times}}, \underbrace{2, \ldots, 2}_{n_2 \text{ times}}, \ldots, \underbrace{K, \ldots, K}_{n_K \text{ times}})$$

The resulting partition model is denoted $ESC_n(P_\nu)$. Here we give details on the clustering structure implied by the microclustering prior in Equations (3) and (4), as well as on the prior predictive distribution of the cluster label for a new observation in Equation (5). Betancourt *et al.* (2022) derive the conditional and marginal EPPF for the class of microclustering priors as well as the conditional allocation probabilities. Let $(z_1, \ldots, z_n)$ be the cluster allocation labels for $\rho_n \sim ESC_n(P_\nu)$. Then for any $i \in [n]$ Betancourt *et al.* (2022) show that:

$$\Pr(\rho_n \mid \nu) = \Pr(n_1, \ldots, n_K \mid \nu) = \frac{K!}{n! \Pr(E_n \mid \nu)} \prod_{j=1}^{K} n_j! \nu(n_j) \tag{3}$$

$$\Pr(\rho_n) = \Pr(n_1, \ldots, n_K) = \frac{1}{\Pr(E_n)} \mathbb{E}_{\nu \sim P_\nu} \left[\frac{K!}{n!} \prod_{j=1}^{K} n_j! \nu(n_j) \right] \tag{4}$$

$$\pi(z_i = j \mid \mathbf{z}_{-i}, \nu) \propto \begin{cases} (n_{j,-i} + 1) \dfrac{\nu(n_{j,-i} + 1)}{\nu(n_{j,-i})} & j = 1, \ldots, K_{-i} \\ (K_{-i} + 1)\nu(1) & j = K_{-i} + 1 \end{cases} \tag{5}$$

where $\mathbf{z}_{-i}$ is the set of cluster labels excluding z_i, $n_{j,-i}$ is the numerosity of $C_{j,-i} = C_j \setminus \{i\}$, and K_{-i} is the number of clusters in the induced partition of $[n] \setminus \{i\}$. Betancourt *et al.* (2022) suggest setting ν to a negative binomial truncated to the positive integers and show that the resulting model, which they call the ESC-NB model, exhibits the microclustering property. We use a variant of the ESC-NB model by setting ν to a shifted negative binomial as it aids the choice of hyperparameters. We set $\nu = \text{NegBin}(r, p) + 1$ where $r \sim \text{Gamma}(\eta, \sigma)$ and $p \sim \text{Beta}(u, v)$. To set the hyperparameters σ, η, u and v, one can use the conditional distribution on the number of clusters K and a prior guess on the number of clusters. In general posterior inference is not sensitive to the choice of the hyperparameters in the prior for

r and p. Nevertheless, the marginal and conditional distributions of K can be analytically calculated or approximated as in the following proposition. The proposition can be used for setting the hyperparameters in the priors for r and p, especially when relevant prior information on K is available.

Proposition 1. *The conditional distribution on the number of clusters K in the ESC model with a shifted negative binomial is given by*

$$\pi(K \mid r, p) = \frac{1}{\Pr(E_n \mid r, p)} \begin{cases} \dfrac{(1-p)^{rK} p^{n-K}}{(n-K)B(rK, n-K)} & K < n \\ (1-p)^{rn} & K = n \end{cases} \tag{6}$$

where $B(\cdot,\cdot)$ is the Beta function. The marginal distribution of K is approximated by

$$\pi(K) \approx \tilde{\pi}(K) \propto \frac{\Gamma(n-K+u)}{\Gamma(n-K+1)} \times \begin{cases} \sigma^{\eta}\Gamma(\eta+v)K^{-\eta}(n-K)^{\eta-u}\Psi(v+\eta, \eta-u+1, \sigma/\omega_K) & K < n \\ \sigma^{u}\Gamma(\eta-u)K^{-u} & K = n \end{cases} \tag{7}$$

where $\Psi(\cdot,\cdot,\cdot)$ is the confluent hypergeometric function of the second kind. If $u = v = 1$, the marginal distribution of K is exactly given by

$$\pi(K) \propto \omega_K'^{-\eta}\Psi\left(\eta, \eta, \frac{\sigma}{\omega_K'}\right) - \mathbf{1}_{\{K<n\}}\omega_K^{-\eta}\Psi\left(\eta, \eta, \frac{\sigma}{\omega_K}\right) \tag{8}$$

where $\omega_K = \frac{K}{n-K}$ and $\omega_K' = \frac{K}{n-K+1}$.

Proof. See Supplementary Material. □

In our simulation studies and real data analyses, we opt for an empirical Bayes approach (see Algorithm 2) to set the hyperparameters for r and p, which is consistent with our method for setting the hyperparameters for $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$.

We conclude this section by noting that the model lends itself to any choice of partition prior. Particular choices could favour a different clustering structure and should be tailored to the application in question. For instance, Pitman-Yor (Pitman and Yor, 1997) or

Algorithm 2 Choosing values for η, σ, u, and v.

1. Fix the cluster labels at the initial clustering configuration obtained in Step 2 of Algorithm 1.
2. Sample r and p from their conditional posteriors in the model using a Gamma$(1, 1)$ prior for r and Beta$(1, 1)$ prior for p.
3. Use MLE to fit a Gamma(η, σ) distribution to the posterior samples of r and a Beta(u, v) distribution to the posterior samples of p.

Gibbs-type priors (Gnedin and Pitman, 2006) could be used as drop-in replacements for the microclustering prior and cover a wide range of partition priors such as mixture with random number of components (Miller and Harrison, 2018; Argiento and De Iorio, 2022). When prior information on the partition is available a more direct approach could be employed; see for example Paganin *et al.* (2021). We also note that as the number of observations n grows larger than the number of clusters K (i.e., when not all clusters are singletons), posterior inference quickly becomes less sensitive to the choice of partition prior as the likelihood will dominate the posterior.

2.5 Posterior inference

Posterior inference is performed through an MCMC scheme. Cluster allocations can be updated either through a Gibbs update of individual cluster labels, or through a split-merge algorithm as proposed by Jain and Neal (2004). The split-merge algorithm is more efficient for large n as it leads to better mixing of the chain. In our applications we combine a Gibbs step and a split-merge step in each iteration as suggested by Jain and Neal (2004). In Supplementary Material we show that the time complexity of the cluster reallocation step is $O(n^2)$. We note that the pre-computation of the $\binom{n}{2}$ pairwise distances is typically not a bottleneck, as the distances are computed only once. Posterior inference for r and

p are performed through a Metropolis step and a Gibbs update respectively. We do not sample the λ_k and θ_{kt} as we marginalise over them. If required, they can be sampled by conditioning on the cluster allocation and sampling from a Gamma-Gamma conjugate model. See Supplementary Material for derivations of the posterior conditionals and the full details of the Gibbs and split-merge algorithms.

We note that the MCMC scheme can be modified as necessary to accommodate different choice of partition priors, as efficient algorithms are available for most Bayesian nonparametric processes.

3 Application to Digital Numismatics

3.1 Description of the data

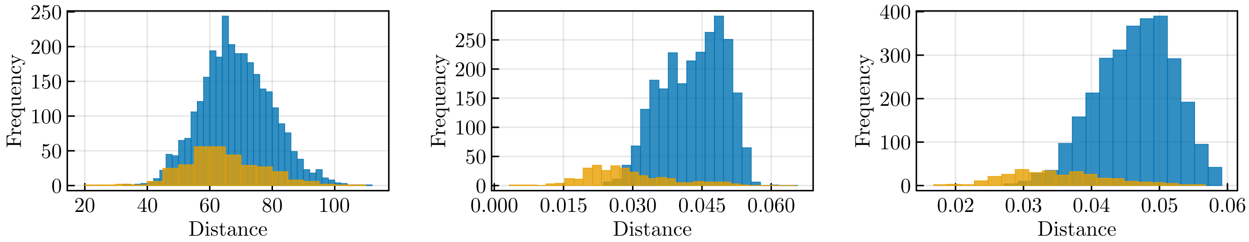
Die studies determine the number of dies used to mint a discreet issue of coinage. With almost no exceptions, dies were destroyed after they wore out, which is why die studies rely on an analysis of the coins struck by them. From a statistical viewpoint the first task to be accomplished is clustering the coins with the goal of identifying if they were cast from the same die.

Die studies are an indispensable tool for pre-modern historical chronology and economic and political history. They are used for putting coins (and by extension rulers and events) into chronological order, to identify mints, and for estimating the output of a mint over time. While digital technology has made large amounts of coinage accessible, numismatic research still requires meticulous and time-consuming manual work. Conducting a die study is time consuming because each coin has to be compared visually to every other coin at least once to determine whether their obverse (front face) and their reverse (back face) were struck from the same dies. For example, a study of 800 coin obverses would require more than 300 000 visual comparisons and could take an expert numismatist approximately 450 hours from scratch. This makes it practically impossible to conduct large scale die studies of coinages like that of the Roman Empire, which would be historically more valuable than the small-scale die studies done today. The practical difficulties of manual die-studies calls for computer-assisted die studies.

We consider here silver coins from one of several issues minted between late 64 C.E and mid 66 C.E., immediately after the great fire of Rome. Pressed for funds, Nero reduced the weight of gold and silver coins by c. 12%, so that he could produce more coinage out of the available bullion stock. Determining the number of dies used to strike this coinage will make it possible to come up with reasonable estimates of how many gold and silver coins Nero minted during this period, and help to determine how much bullion he may have saved in the immediate aftermath of the great fire. This type of numismatic work would require time-consuming effort by highly trained experts if performed manually. Here we demonstrate the potential in digital numismatics of our strategy by clustering a dataset of 81 coins, which requires a few hours for pre-processing of the images and a few minutes to fit out model. The distance computation is straightforward to parallelise, further speeding up computation. The data consists of 81 high-resolution images of obverses taken from a forthcoming die study on Nero's coinage. To test the performance of our model, die analysis is firstly performed by visual inspection by a numismatic expert to provide the ground truth. This analysis identifies ten distinct die groups. The images were standardised to 380×380 pixels to compute the pairwise distances.

3.2 Computing pairwise distances

Fitting the model in Equation (1) requires the definition of a distance between images that has the potential to differentiate between images of coins minted from different dies and to capture the similarity of images of coins minted from the same die. The pixelwise Euclidean distance between the digital images cannot be used to obtain such information about the semantic dissimilarity of images. Due to the high dimensionality of the ambient image space the data set of images is sparse, with little separation between the largest and smallest pairwise Euclidean distances in the data set (Beyer *et al.*, 1999). Figure 1a illustrates this for our dataset. In contrast, numismatists rely on domain knowledge and often years of experience to identify few key feature points in images of coins to aid comparisons. This essentially coincides with disregarding irrelevant features and performing dimension reduction. When defining the distance between images, our goal is to automate expert knowledge acquisition and focus on extraction of key features. This is a common strategy in many tasks in computer vision (Szeliski, 2010) and, more generally, in statistical shape analysis (Dryden and Mardia, 2016; Gao *et al.*, 2019a). Taylor (2020) uses landmarking to define a distance between images of ancient coins with the ultimate goal of die-analysis using simple hierarchical clustering. They do not provide an estimate of the number of dies represented in the sample or an overall subdivision of coins into die groups.



(a) Pixelwise distances

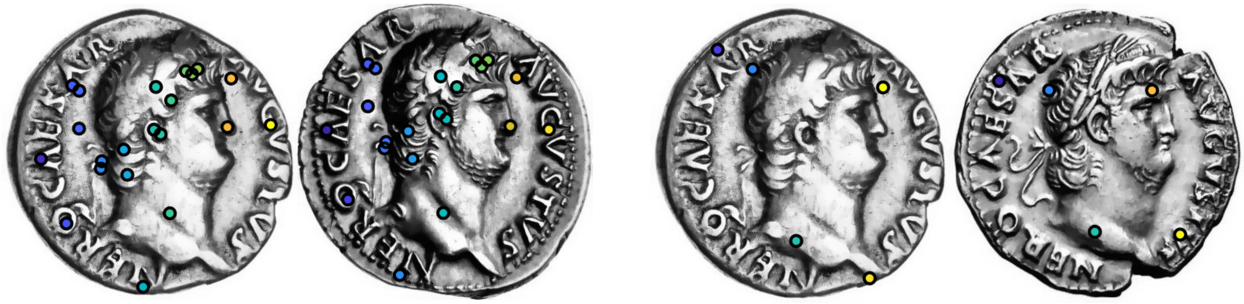
(b) Distances computed using our method

(c) Distances after MDS embedding

Figure 1. Coin data: Histogram of within-cluster distances (orange) and inter-cluster distances (blue). The clusters correspond to the true clusters obtained by a die study conducted by an expert numismatist.

To identify comparable key features across pairs of coin images, we find sets of matched landmark pairs between images by exploiting the Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) and a low-distortion correspondence filtering procedure (Lipman *et al.*, 2014).

We use the landmark ranking algorithm of Gao *et al.* (2019b) and Gao *et al.* (2019a), which we extend for ranking pairs of landmarks. We define a dissimilarity score between images using these ranked landmark pairs. Figure 2 shows an example of matched landmark sets for images from the same die group and from different die groups. Details of the pipeline are provided in the Supplementary Material. These dissimilarity scores are used as input for our algorithm. The pairwise dissimilarities are shown in Figure 1b, and in Supplementary Material we compare the prior predictive distribution on dissimilarities as implied by by our data-driven prior specification process to the kernel density estimate of the dissimilarities.



(a) Matched landmarks on coins belonging to the same die group.

(b) Matched landmarks on coins belonging to different die groups.

Figure 2. Matched sets of landmarks are used to construct a dissimilarity measure between images of coins. The number of landmarks is one of the components of this dissimilarity measure.

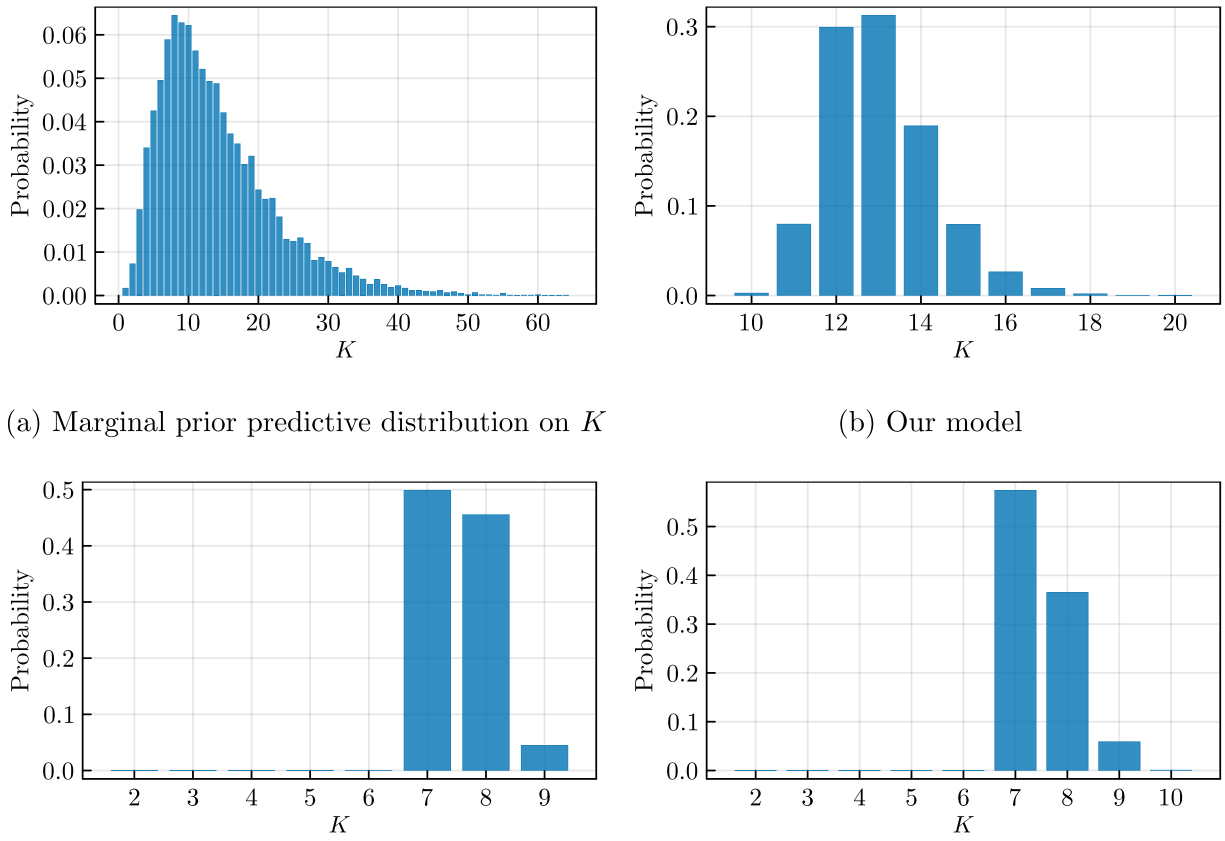
3.3 Results

We run our model on the coin data for 50000 iterations, discarding the first 10000 iterations as burnin. We compare our model to the Mixture of Finite Mixtures (MFM) model proposed by Miller and Harrison (2018) and a Dirichlet Process Mixture (DPM), both as implemented in the `Julia` package `BayesianMixtures.jl` (Miller, 2020), using Normal mixture kernels with diagonal covariance matrix and conjugate priors. For the purposes of the comparison, we embed the coins as points in Euclidean space by applying Multi-Dimensional Scaling (Kruskal, 1964) (as implemented by the `MultivariateStats.jl` package in `Julia`) to the dissimilarity scores between the coins. We run the MFM and DPM samplers on the MDS output, which is 80-dimensional.

To evaluate algorithm performance, we compute the *co-clustering matrix* whose entries s_{ij} are given by $s_{ij} = \Pr(z_i = z_j \mid X)$. Each s_{ij} can be estimated from the MCMC output and its estimate is not affected by the label-switching phenomenon (Stephens, 2000; Fritsch and Ickstadt, 2009). In Figure 4 we compare the true adjacency matrix to the co-clustering matrices obtained by our model, MFM, and DPM. In Figure 3 we show the marginal prior predictive distribution on the number of clusters K implied by our choice of prior hyperparameters, as well as the posterior distribution on K for each method. In Supplementary Material we show the posterior co-clustering matrix for our model without repulsion, the posterior distributions of r and p, and we provide convergence diagnostics for the sampler.

For each method a clustering point-estimate is obtained via the SALSO algorithm (Dahl *et al.*, 2022). This algorithm takes as its input the posterior samples of cluster allocations and searches for a point estimate that minimises the posterior expectation of the Variation of Information distance (Meilă, 2007; Wade and Ghahramani, 2018). Point estimates are also obtained via k-means (on the MDS output) and k-medoids (on the dissimilarities), as implemented in the `Clustering.jl` package in `Julia`, using the value of K obtained by the elbow method as in Section 2.4.1. Table 1 shows the Binder loss, Normalised Variation of Information (NVI) distance, Adjusted Rand Index (ARI), and Normalised Mutual Information (NMI) of these point estimates with respect to the true clustering. In Supplementary Material we show the adjacency matrices for the various point estimates.

The findings from this application suggest that clustering with our model on the distances alone can produce sensible results in terms of the original data, providing a viable strategy for high-dimensional settings. We further demonstrate the effectiveness of our model through simulation studies in Supplementary Material. These studies (1) show the effect of dimensionality and cluster separation on posterior inference, (2) demonstrate the robustness of our method to choice of prior hyperparameters, and (3) demonstrate the importance of the repulsion term in our likelihood. We remind the reader that our model underestimates uncertainty as discussed in section 2.2.



(a) Marginal prior predictive distribution on K

(b) Our model

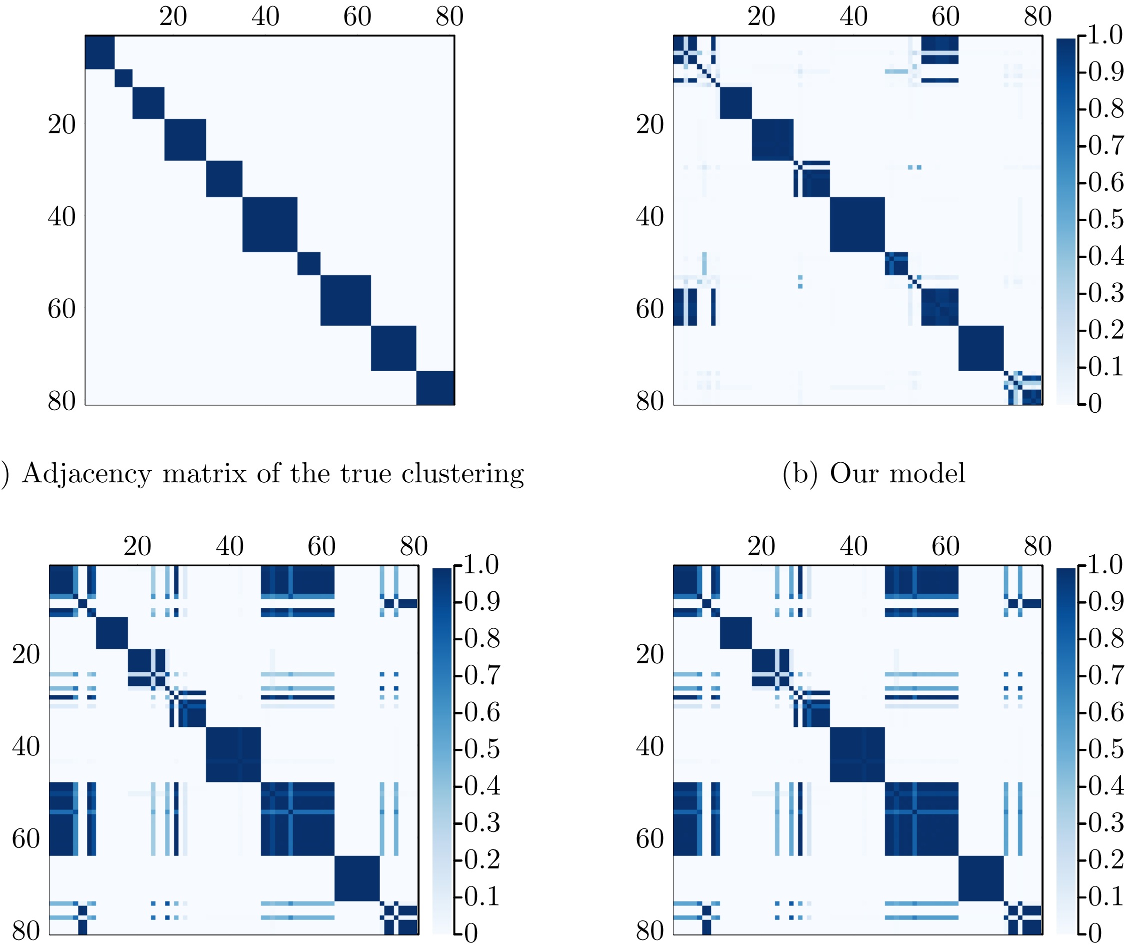
(c) MFM

(d) DPM

Figure 3. Coins data: Posterior distribution of the number of clusters K.

	Our Model	MFM	DPM	k-means	k-medoids
Binder loss	**0.04**	0.12	0.12	0.08	0.10
NVI distance	**0.12**	0.19	0.19	0.28	0.38
ARI	**0.78**	0.52	0.52	0.52	0.41
NMI	**0.88**	0.79	0.79	0.74	0.64
K	17	7	7	12	12

Table 1. Coins data: Comparison of point estimates with the true clustering. We highlight the best value for each measure in bold.



(a) Adjacency matrix of the true clustering

(b) Our model

(c) MFM

(d) DPM

Figure 4. Coins data: Posterior co-clustering matrices

4 Discussion

It is the curse of dimensionality, a malediction that has plagued the scientist from the earliest days.

— Richard Bellman, 1961

When clustering in large dimensions, the main statistical challenges include (i) accommodating for the sparsity of points in large dimensions, (ii) capturing the complexity of the data generating process in high-dimensions, including the interdependence of features which might be cluster-specific, (iii) estimating model parameters, (iv) developing methods which are robust to different underlying generative processes and cluster characteristics, (v) devising computational algorithms that are scalable to large datasets, (vi) producing interpretable results, (vii) assessing the validity of the cluster allocation, and (viii) determining the degree to which different features contribute to clustering.

We propose a hybrid clustering method for high-dimensional problems which is essentially model-based clustering on pairwise distances between the original observations. This method is also applicable in settings where the likelihood is not computationally tractable. The strategy allows us to overcome many of the aforementioned challenges, bypassing the specification of a model on the original data. The main contribution of our work is to combine cohesive and repulsive components in the likelihood, and we provide theoretical justifications for our model choices. Our method is robust to different generative processes, and computationally more efficient than model-based approaches for high dimensions because it reduces the multi-dimensional likelihood on each data point to a unidimensional likelihood on each distance. Our method also leads to interpretable results as clusters are defined in terms of the original observations. The model can be easily extended to categorical variables by considering (for example) the Hamming distance or cross entropy and specifying appropriate distributions f and g in Equation (1). The main drawbacks of our methodology is that the role of each feature is embedded in the distances and model performance is dependent on the definition of the distances. We do not advise the use of a distance-based approach when the dimension is small because in that case standard model-based approaches work well, and using distances as a summary of the data causes loss of information.

From our application in digital numismatics, it is clear that the definition of the distances plays a crucial role and a future direction of this work is to develop landmark estimation methods better able to capture the distinguishing features of images.

Finally, there is an interesting connection between the likelihood in Equation (1) and the likelihood for a stochastic blockmodel in the $p1$ family (Snijders and Nowicki, 1997; Nowicki and Snijders, 2001; Schmidt and Morup, 2013) where every block of nodes can be thought of as a cluster of similar objects. This connection is a topic of further research.

5 Supplementary Material

Additional results: Proof of Proposition 1, detailed description of the MCMC algorithm, computational complexity of the MCMC algorithm, details of the method to compute distances between coin images, additional plots and results from the numismatic example, alternative method to choose prior hyperparameters, and simulation studies are available online in a supplementary PDF document.

Code: Julia code to perform posterior inference is available on Github at `https://github.com/abhinavnatarajan/RedClust.jl`, and also as a package in the default Julia package registry ("General"). The code used to run the numismatic and simulated examples and generate the corresponding figures is available at `https://github.com/abhinavnatarajan/RedClust.jl/tree/examples`.

6 Acknowledgements

This work was supported by the National University of Singapore through NUS HSS Seed Fund R-607-000-449-646, by Yale-NUS College through Yale-NUS IG20-RA001, and the Singapore Ministry of Education through the Academic Research Fund Tier 2 under Grant MOE-T2EP40121-0021.

7 Conflicts of Interest

We report that there are no competing interests to declare.

References

Argiento R, De Iorio M (2022). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." *Annals of Statistics*, **50**(5), 2641–2663. doi:10.1214/22-AOS2201.

Barcella W, De Iorio M, Baio G (2017). "A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models." *Canadian Journal of Statistics*, **45**(3), 254–273. doi:10.1002/cjs.11323.

Barry D, Hartigan JA (1992). "Product Partition Models for Change Point Problems." *The Annals of Statistics*, **20**(1), 260 – 279. doi:10.1214/aos/1176348521.

Beaumont MA, Zhang W, Balding DJ (2002). "Approximate Bayesian Computation in Population Genetics." *Genetics*, **162**(4), 2025–2035. ISSN 0016-6731. doi:10.1093/genetics/162.4.2025.

Besag J (1975). "Statistical Analysis of Non-Lattice Data." *Journal of the Royal Statistical Society. Series D (The Statistician)*, **24**(3), 179–195. ISSN 00390526, 14679884. doi:10.2307/2987782.

Betancourt B, Zanella G, Steorts RC (2022). "Random Partition Models for Microclustering Tasks." *Journal of the American Statistical Association*, **117**(539), 1215–1227. doi:10.1080/01621459.2020.1841647.

Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999). "When Is "Nearest Neighbor" Meaningful?" In C Beeri, P Buneman (eds.), "Database Theory — ICDT'99," pp. 217–235. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-49257-3. doi:10.1007/3-540-49257-7_15.

Chandler RE, Bate S (2007). "Inference for Clustered Data Using the Independence Log-likelihood." *Biometrika*, **94**(1), 167–183. ISSN 00063444, 14643510. doi:10.1093/biomet/asm015.

Corander J, Sirén J, Arjas E (2008). "Bayesian spatial modeling of genetic population structure." *Computational Statistics*, **23**(1), 111–129. ISSN 1613-9658. doi:10.1007/s00180-007-0072-x.

Dahl DB (2008). "Distance-Based Probability Distribution for Set Partitions with Applications to Bayesian Nonparametrics." In "2008 JSM proceedings: Papers presented at the Joint Statistical Meetings, Denver, Colorado, August 3-7, 2008, and other ASA-sponsored conferences ; communicating statistics: speaking out and reaching out," American Statistical Association, Alexandria, Virginia, USA.

Dahl DB, Day R, Tsai JW (2017). "Random Partition Distribution Indexed by Pairwise Information." *Journal of the American Statistical Association*, **112**(518), 721–732. doi:10.1080/01621459.2016.1165103. PMID: 29276318.

Dahl DB, Johnson DJ, Müller P (2022). "Search Algorithms and Loss Functions for Bayesian Clustering." *Journal of Computational and Graphical Statistics*, **31**(4), 1189–1201. doi:10.1080/10618600.2022.2069779.

Dasgupta A, Raftery AE (1998). "Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering." *Journal of the American Statistical Association*, **93**(441), 294–302. ISSN 01621459. doi:10.2307/2669625.

Denison DGT, Holmes CC (2001). "Bayesian Partitioning for Estimating Disease Risk." *Biometrics*, **57**(1), 143–149. doi:10.1111/j.0006-341X.2001.00143.x.

Dryden IL, Mardia KV (2016). *Statistical Shape Analysis, with Application in R*. New York, NY: John Wiley & Sons, Ltd. ISBN 9781119072492. doi:10.1002/9781119072492.

Duan LL, Dunson DB (2021). "Bayesian Distance Clustering." *Journal of Machine Learning Research*, **22**(224), 1–27. ISSN 1532-4435. URL `http://jmlr.org/papers/v22/20-688.html`.

Fearnhead P, Prangle D (2012). "Constructing Summary Statistics for Approximate Bayesian Computation: Semi-Automatic Approximate Bayesian Computation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(3), 419–474. doi:10.1111/j.1467-9868.2011.01010.x.

Fritsch A, Ickstadt K (2009). "Improved criteria for clustering based on the posterior similarity matrix." *Bayesian Analysis*, **4**(2), 367 – 391. doi:10.1214/09-BA414.

Fúquene J, Steel M, Rossell D (2019). "On choosing mixture components via non-local priors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **81**(5), 809–837. doi:10.1111/rssb.12333.

Gao T, Kovalsky SZ, Boyer DM, Daubechies I (2019a). "Gaussian process landmarking for three-dimensional geometric morphometrics." *SIAM Journal on Mathematics of Data Science*, **1**(1), 237–267. doi:10.1137/18M1203481.

Gao T, Kovalsky SZ, Daubechies I (2019b). "Gaussian Process Landmarking on Manifolds." *SIAM Journal on Mathematics of Data Science*, **1**(1), 208–236. doi:10.1137/18M1184035.

Gnedin AV, Pitman J (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences*, **138**, 5674–5685. doi:10.1007/s10958-006-0335-z.

Hartigan J (1990). "Partition models." *Communications in Statistics - Theory and Methods*, **19**(8), 2745–2756. doi:10.1080/03610929008830345.

Hennig C (2015). "What are the true clusters?" *Pattern Recognition Letters*, **64**, 53–62. ISSN 0167-8655. doi:10.1016/j.patrec.2015.04.009.

Ishwaran H, James LF (2003). "Generalized Weighted Chinese Restaurant Processes for Species Sampling Mixture Models." *Statistica Sinica*, **13**(4), 1211–1235. ISSN 10170405, 19968507. URL `https://www.jstor.org/stable/24307169`.

Jain S, Neal RM (2004). "A Split-Merge Markov chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *Journal of Computational and Graphical Statistics*, **13**(1), 158–182. doi:10.1198/1061860043001.

Johnstone IM, Titterington DM (2009). "Statistical challenges of high-dimensional data." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **367**(1906), 4237–4253. doi:10.1098/rsta.2009.0159.

Klaus J (1995). *Topology.* Springer-Verlag, 2nd edition. ISBN 978-0387908922.

Kleinberg J (2002). "An Impossibility Theorem for Clustering." In "Proceedings of the 15th International Conference on Neural Information Processing Systems," NIPS'02, p. 463–470. MIT Press, Cambridge, MA, USA.

Kruskal JB (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika*, **29**, 1–27. doi:10.1007/BF02289565.

Larribe F, Fearnhead P (2011). "On Composite Likelihoods in Statistical Genetics." *Statistica Sinica*, **21**(1), 43–69. ISSN 10170405, 19968507. URL `https://www.jstor.org/stable/24309262`.

Lau JW, Green PJ (2007). "Bayesian Model-Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, **16**(3), 526–558. doi:10.1198/106186007X238855.

Lele S, Taper ML (2002). "A composite likelihood approach to (co)variance components estimation." *Journal of Statistical Planning and Inference*, **103**(1), 117–135. ISSN 0378-3758. doi:10.1016/S0378-3758(01)00215-4.

Li N, Stephens M (2003). "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data." *Genetics*, **165**(4), 2213–2233. ISSN 0016-6731. doi:10.1093/genetics/165.4.2213. PMID: 14704198.

Lindsay BG (1988). "Composite likelihood methods." In "Statistical inference from stochastic processes (Ithaca, NY, 1987)," volume 80 of *Contemporary Mathematics*, pp. 221–239. American Mathematical. Society, Providence, RI. doi:10.1090/conm/080/999014.

Lipman Y, Yagev S, Poranne R, Jacobs DW, Basri R (2014). "Feature Matching with Bounded Distortion." *ACM Transactions on Graphics*, **33**(3). ISSN 0730-0301. doi:10.1145/2602142.

Lowe D (2004). "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision*, **60**, 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Marjoram P, Molitor J, Plagnol V, Tavaré S (2003). "Markov Chain Monte Carlo without likelihoods." *Proceedings of the National Academy of Sciences*, **100**(26), 15324–15328. ISSN 0027-8424. doi:10.1073/pnas.0306899100.

Mclachlan G, Basford K (1988). *Mixture Models: Inference and Applications to Clustering*, volume 38. Marcel Dekker. doi:10.2307/2348072.

Meilă M (2007). "Comparing clusterings—an information based distance." *Journal of Multivariate Analysis*, **98**(5), 873–895. ISSN 0047-259X. doi:10.1016/j.jmva.2006.11.013.

Miller J (2020). "`BayesianMixtures`." Version 0.1.1, URL `https://github.com/jwmi/BayesianMixtures.jl`.

Miller J, Betancourt B, Zaidi A, Wallach H, Steorts RC (2015). "Microclustering: When the Cluster Sizes Grow Sublinearly with the Size of the Data Set." doi:10.48550/arXiv.1512.00792.

Miller JW, Harrison MT (2018). "Mixture Models With a Prior on the Number of Components." *Journal of the American Statistical Association*, **113**(521), 340–356. doi:10.1080/01621459.2016.1255636. PMID: 29983475.

Møller J, Øivind Skare (2001). "Coloured Voronoi tessellations for Bayesian image analysis and reservoir modelling." *Statistical Modelling*, **1**(3), 213–232. doi:10.1177/1471082X0100100304.

Müller P, Quintana F, Rosner GL (2011). "A Product Partition Model With Regression on Covariates." *Journal of Computational and Graphical Statistics*, **20**(1), 260–278. doi:10.1198/jcgs.2011.09066.

Nowicki K, Snijders TAB (2001). "Estimation and Prediction for Stochastic Blockstructures." *Journal of the American Statistical Association*, **96**(455), 1077–1087. doi:10.1198/016214501753208735.

Paganin S, Herring AH, Olshan AF, Dunson DB (2021). "Centered Partition Processes: Informative Priors for Clustering (with Discussion)." *Bayesian Analysis*, **16**(1), 301 – 670. doi:10.1214/20-BA1197.

Petralia F, Rao V, Dunson D (2012). "Repulsive Mixtures." In F Pereira, CJC Burges, L Bottou, KQ Weinberger (eds.), "Advances in Neural Information Processing Systems," volume 25. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper/2012/file/8d6dc35e506fc23349dd10ee68dabb64-Paper.pdf`.

Pitman J (1996). “Some Developments of the Blackwell-MacQueen URN Scheme.” *Lecture Notes-Monograph Series*, **30**, 245–267. ISSN 07492170. doi:10.1214/lnms/1215453576.

Pitman J, Yor M (1997). “The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator.” *The Annals of Probability*, **25**(2), 855–900. ISSN 00911798. URL `http://www.jstor.org/stable/2959614`.

Quinlan JJ, Quintana FA, Page GL (2017). “Parsimonious Hierarchical Modeling Using Repulsive Distributions.” doi:10.48550/arXiv.1701.04457.

Quintana FA (2006). “A predictive view of Bayesian clustering.” *Journal of Statistical Planning and Inference*, **136**(8), 2407–2429. ISSN 0378-3758. doi:10.1016/j.jspi.2004.09.015.

Quintana FA, Iglesias PL (2003). “Bayesian clustering and product partition models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(2), 557–574. doi:10.1111/1467-9868.00402.

Rigon T, Herring AH, Dunson DB (2023). “A generalized Bayes framework for probabilistic clustering.” *Biometrika*. ISSN 1464-3510. doi:10.1093/biomet/asad004.

Schmidt MN, Morup M (2013). “Nonparametric Bayesian modeling of complex networks: an introduction.” *IEEE Signal Processing Magazine*, **30**(3), 110–128. doi:10.1109/MSP.2012.2235191.

Snijders TAB, Nowicki K (1997). “Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure.” *Journal of the Classification*, **14**(455), 75–100. doi:10.1198/016214501753208735.

Stephens M (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**(4), 795–809. doi:10.1111/1467-9868.00265.

Szeliski R (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media. ISBN 978-3-030-34372-9. doi:10.1007/978-3-030-34372-9.

Taylor ZM (2020). *The Computer-Aided Die Study (CADS): A Tool for Conducting Numismatic Die Studies with Computer Vision and Hierarchical Clustering*. Bachelor's thesis, Trinity Uni. URL `https://digitalcommons.trinity.edu/compsci_honors/54/`.

Varin C, Reid N, Firth D (2011). "An Overview of Composite Likelihood Methods." *Statistica Sinica*, **21**(1), 5–42. ISSN 10170405, 19968507. URL `https://www.jstor.org/stable/24309261`.

Wade S, Ghahramani Z (2018). "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)." *Bayesian Analysis*, **13**(2), 559 – 626. doi:10.1214/17-BA1073.

Xu Y, Müller P, Telesca D (2016). "Bayesian inference for latent biologic structure with determinantal point processes (DPP)." *Biometrics*, **72**(3), 955–964. doi:10.1111/biom.12482.

Zanella G, Betancourt B, Wallach H, Miller J, Zaidi A, Steorts RC (2016). "Flexible Models for Microclustering with Application to Entity Resolution." In "Proceedings of the 30th International Conference on Neural Information Processing Systems," NIPS'16, p. 1425–1433. Curran Associates Inc., Red Hook, NY, USA. ISBN 9781510838819.