Adaptive Dirichlet Process mixture model with unknown concentration parameter and variance: Scaling high dimensional clustering via collapsed variational inference

Annesh Pal, Aguirre Mimoun, Rodolphe Thiébaut, Boris P. Hejblum

Biostatistics, YEAR, pp. 1–19
DOI added during production
Published: Date added during production
Paper

Adaptive Dirichlet Process mixture model with unknown concentration parameter and variance: Scaling high dimensional clustering via collapsed variational inference

Annesh Pal[1,2] Aguirre Mimoun[3] Rodolphe Thiébaut[1,2,4] and Boris P. Hejblum[1,2,*]

[1]Université de Bordeaux, INSERM, INRIA, Bordeaux Population Health, U1219, SISTM, 33000 Bordeaux, France
[2]Vaccine Research Institute, 94010 Cretéil, France
[3]Centre Hospitalier Universitaire de Bordeaux, Laboratoire d'Hématologie, 33000 Bordeaux, France
[4]Centre Hospitalier Universitaire de Bordeaux, Service d'Information Médicale, 33000 Bordeaux, France

[*]Corresponding author. boris.hejblum@u-bordeaux.fr

Abstract

Dirichlet process mixture models (DPMM) provide a principled framework for density estimation and data clustering, offering a robust alternative to parametric mixture models by estimating cluster number K from an infinite choice of partitions induced by a Dirichlet Process $DP(\alpha, G_0)$ prior on the data. Posterior consistency is guaranteed by an adaptive estimation of the DP concentration parameter α, and the estimated K jointly depends on α as well as the variance hyper-parameter of DP base distribution G_0. Although Markov Chain Monte Carlo (MCMC) methods have bridged the gap between theory and application for such models, they scale poorly to high-dimensional data (e.g. with several hundred or even thousands of features) with limited sample size and suffer from slow and difficult convergence. Variational Inference (VI) exist as an alternative for faster convergence, but lack implementation with increasing model fidelity. Collapsed VI is implemented to integrate out hierarchical parameters beyond mean field VI's independence assumptions, however, the resulting complexity imposes an implementational barrier to a fully adaptive DPMM inference for high dimensional data. We propose a novel method that performs adaptive clustering with DPMM using collapsed VI, while incorporating weakly-informative priors for α and G_0. We illustrate the importance of G_0 covariance structure and prior choice by considering different parameterisations of the data covariance matrix. On high-dimensional Gaussian simulations, our model demonstrates substantially faster convergence than a state-of-the-art MCMC splice sampler. We further evaluate performances on Negative Binomial simulations and conduct sensitivity analyses to assess robustness on realistic data conditions. Application to a publicly available leukemia transcriptomic data set comprising 72 samples and 2,194 gene expression successfully recovers every known sub-type, all while identifying additional gene expression-based sub-clusters with meaningful biological interpretation.

Keywords: Clustering, Bayesian Nonparametrics, Dirichlet process mixture model, Variational inference, Unstructured covariance, Concentration parameter, High-dimension, Gene expression data

Received: Date Month Year. **Revised received:** Date Month Year. **Accepted:** Date Month Year

1. Introduction

Clustering is an exploratory data analysis technique, where the goal is to group the data based on measured or estimated inherent similarities (Jain, 2010). Since the early emergence of K-means in 1950s (Steinhaus et al., 1956; Forgy, 1965; MacQueen, 1967; Lloyd, 1982), a large number of clustering methods have been deployed for diverse interdisciplinary applications that deal with high dimensional data sets (Dinh et al., 2025). These methods include density based approaches (Ester et al., 1996; Campello et al., 2013), graph based techniques (Ertoz et al., 2002; Traag et al., 2019) and probabilistic models that define clusters based on mixture distributions (Bock (1996), Bouveyron et al. (2007)). However, a prerequisite for standard clustering techniques and finite mixture models is the number of clusters K (or analogous parameters for density based methods) present *a priori* (Cai et al., 2021). Even dimensionality reduction techniques like PCA (Pearson, 1901) or UMAP (McInnes et al., 2018) fail to capture underlying structure for high-dimensional data and require strong domain-specific knowledge to select an appropriate K (Wani, 2025). To circumvent this limitation, Bayesian non-parametric mixture models (NPMM) have emerged as an efficient alternative to perform adaptive clustering based on probability distributions without pre-specifying K (Orbanz and Teh, 2010; Gershman and Blei, 2012).

Introduction of Dirichlet process $DP(\alpha, G_0)$ by Ferguson (1973) enabled the development of DP mixture models (DPMM) as a class of Bayesian NPMM for density estimation, clustering as well as regression (Müller and Mitra, 2013). DPMM assume an infinite choice of unknown K corresponding to partitions of the mixture components based on a DP prior (Frühwirth-Schnatter and Malsiner-Walli, 2019), and corresponding posterior inference facilitates the estimate of K (a non-empty partition corresponds to a cluster $k \in K$). Although DPMM have been scrutinized for inconsistent posterior estimation (Miller and Harrison, 2014, 2013), Ascolani et al. (2023) have successfully demonstrated that consistency can be obtained by varying the concentration parameter α, for instance, using a Gamma prior to implement a fully Bayesian scheme to update α (West, 1992).

DPMM often results in analytically intractable likelihoods and posteriors. Markov Chain Monte Carlo (MCMC) methods are a popular choice to sample from such analytically intractable distributions for parameter inference (Escobar and West, 1995; Neal, 2000). However, these methods exhibit slow convergence rates with difficulty in assessment (Gershman and Blei, 2012), especially for high dimensions with several hundreds or even thousands of features per sample and limited sample size. An alternative is to use variational inference (VI) that approximates the true posterior $p(\cdot|data)$ with a tractable density function $q(\cdot)$ from a (known) family of probability distributions while deterministically optimising an information-theoretic criterion like KL divergence (Blei et al., 2017). Rooted in early estimation methods for graphical models (Jordan et al., 1999), VI has gained prominence for its computational efficiency and scalability, particularly in modern machine learning applications involving large datasets (Bernardi et al., 2024; Loya et al., 2025).

A general scheme of DPMM implementation would involve a stick-breaking representation of the DP prior (Sethuraman, 1994) with hyper-priors defined for α and base distribution G_0. To estimate model parameters, mean-field VI is the most widely adopted approach that assumes a factorized variational distribution and iteratively updates the parameters using conjugate priors (Blei and Jordan, 2006; Bishop and Nasrabadi, 2006). Although the independent factorization among variational parameters induces analytical tractability, latent allocation variables in DPMM are inherently dependent through their shared mixture proportions across samples. This fails to capture the relationships between hierarchical parameters and results in inconsistency between cluster numbers estimated from α and the latent allocation variables respectively (details in Supplementary material). Kurihara et al. (2007) have provided a collapsed VI framework by integrating out (collapsing) the mixture proportions, thus defining a prior distribution for the latent allocation vectors that depends only on α (1). This integration, however, constructs a non-standard prior with no standard conjugate hyper-prior for α. Hence, the collapsed VI approach lacks a current working implementation with an adaptive inference on α.

Apart from α that has a directly proportional effect on the inferred number of clusters (Li et al., 2019), the model variance also plays a significant role in estimating K (Hejblum et al., 2019). Incorporating G_0 variance as an

unknown variable becomes important for coherent estimation, which is often ignored while implementing DPMM (including Blei and Jordan (2006)). The cluster variance particularly impacts high-dimensional data, which is often characterized by reduced signal-to-noise ratio, high variability due to limited sample size (Johnstone and Titterington, 2009) and reduced separation between data points with increasing dimensions (Clarke et al., 2008). The commonly used Inverse-Wishart IW prior for covariance matrix receives criticism for a rigid structure, strongly coupled relations between variance and covariance (O'Hagan and Forster, 2004; Tokuda et al., 2025) and high sensitivity to hyper-parameter choice (Hennig et al., 2015). Factorized covariance representations (like Cholesky decomposed factors) exhibit similar variance-covariance dependency as the resulting matrix have linear combination of constituent factor elements that are common in variance as well as covariance terms. Approaches like block diagonal covariance matrices, latent factor models (Chandra et al., 2023) and element-wise distributions (Jing et al., 2024) focus on introducing conditional independence within the covariance matrices. A fundamental issue with Gaussian mixture models or GMM (both finite and infinite) is their assumption of Gaussian distributed clusters. This assumption is frequently violated in real world data sets, where cluster distributions may exhibit skewness, heavy tails, multimodality, or other deviations from normality (Hejblum et al., 2019; Zhang et al., 2022). Paired with the '*curse of dimensionality*', data sets like gene expression data often show less separated or overlapped clusters (Yu et al., 2017) and using GMMs lead to misspecified clustering (Kasa and Rajan, 2023). In biological applications, the cluster overlapping can also arise from factors like multimodality and high correlation (Clarke et al., 2008). In order to retrieve clustering estimates from such data, it is important to have cluster-specific distributions that regulate within-cluster spread as well as inter-cluster separability for robust outcomes.

In this paper, we propose a novel collapsed VI approach that explicitly incorporates the concentration parameter α as well as model covariance in a DPMM. Combining Euler-Maclaurin and Taylor series approximations, we obtain a closed form variational distribution for α with a conjugate Gamma hyper-prior. Following a selection between different covariance parameterisations, the analysis evaluates effect of hyper-prior choice through their corresponding performances on Gaussian simulated data. We establish the model with sparsity inducing hyper-prior (namely, Sparse DPMM) as the best choice (a desirable property for high-dimensional as well as highly-clustered data (Jing et al., 2024)) and further evaluate the sensitivity based on the hyper-parameters with high dimensional ($d = 1000$) Negative Binomial simulations. The optimal model is compared to an MCMC approach as implemented in the R package **NPflow** (Hejblum et al., 2019) in terms of convergence speed. We apply Sparse DPMM to cluster leukemia sub-types based on gene expression data set (Armstrong et al., 2002; De Souto et al., 2008) with 2194 genes and 72 samples, and further evaluated the performance with existing clustering techniques. In addition, we provide a computational implementation through the R-package **vimixr**, available on CRAN.

The rest of the paper continues as follows. Section 2 develops the statistical methodology of our proposed model, focusing on formulating variational updates for different G_0 covariance hyper-priors and model implementation (Table 1). Section 3 highlights the theoretical merits and illustrates the inference results on simulated data for optimal model choice and sensitivity analysis with an emphasis on hyper-parameter tuning, while comparing computational performance with a splice sampling MCMC approach (Hejblum et al., 2019). It is followed by an application of Sparse DPMM on gene expression data to cluster acute lymphoblastic (ALL), mixed lineage (MLL) and acute myeloid (AML) leukemia sub-types (Armstrong et al., 2002) and compare performance with K-means (MacQueen, 1967), density-based DBSCAN (Ester et al., 1996) and HDBSCAN (Campello et al., 2013), graph-based shared nearest neighbourhood (Ertoz et al., 2002) and Leiden (Kelly, 2023), and model-based high dimensional clustering (Bouveyron et al., 2007) techniques. Section 4 concludes with discussions on the benefits and short comings of the proposed model, with possible extensions for further research and development.

2. Methods

2.1. Problem set-up

Consider $X_n \in \mathbb{R}^d$ random variables sampled from an unknown mixture of distributions F

$$X_n | G \overset{\text{i.i.d}}{\sim} F \quad \text{for } n = 1, \ldots, N$$

where $F(\mathbf{X}) = \int_\Theta f_\eta(\mathbf{X})G(d\eta)$ (Hejblum et al., 2019). We assume a multivariate Gaussian distribution for $f_\eta(.)$ with mean μ and covariance matrix Σ ($\eta = \{\mu, \Sigma\}$). G is the unknown mixing distribution that characterizes the mixture components over the parameter space Θ, thus inducing clustering among X_n. A DP prior on $G \sim DP(\alpha, G_0)$ with concentration parameter α and base distribution G_0 (Ferguson, 1973) gives us a non-parametric mixing distribution

$$G(.) = \sum_{k=1}^{\infty} \pi_k \delta_{\eta_k}(.) \quad \text{where } \eta_k \in G_0$$

The mixing proportions π_k's are drawn from a stick-breaking scheme (Sethuraman, 1994), which further introduce latent allocation variables Z_n in the model that follows a categorical distribution with parameters $\{\pi_k\}$'s such that $p(Z_n = k) = \pi_k$ for every $n \in N$. The likelihood of X_n hence follows

$$X_n \sim \prod_k MVN(\mu_k, \Sigma_k)^{\mathbb{I}[Z_n = k]}$$

(Blei and Jordan, 2006). Our objective is to estimate Z_n as well as $\eta_k = \{\mu_k, \Sigma_k\}$ to quantitatively evaluate the partitioning of the parameter space $\mathbf{\Theta}$.

2.2. Dirichlet process mixture model

A hierarchical Bayesian model follows

$$\begin{aligned} V_k &\sim Beta(1, \alpha) \\ \pi_k &:= V_k \prod_{j<k}(1 - V_j) \\ \mu_k, \Sigma_k &\sim G_0 \\ z_n &\sim Categorical\left(\{\pi_k\}\right) \\ x_n | z_n &\sim MVN(\mu_{z_n}, \Sigma_{z_n}) \end{aligned}$$

This is an illustration of a Dirichlet process mixture model (DPMM), introduced by Antoniak (1974). The concentration parameter α has a direct impact on the posterior expectation of the number of non-empty clusters (Teh et al., 2010). Escobar and West (1995) have proposed a data augmentation scheme that leverages a Gamma hyper-prior $\alpha \sim Gamma(a, b)$ yielding a posterior distribution that adjust to the actual number of clusters observed in the data with consistency (Ascolani et al., 2023). With a conjugate choice for G_0, the model provides an estimate for the unknown parameters of interest.

2.3. Choice of G_0

The underlying structure of $\eta = \{\mu, \Sigma\}$ governs the distributional choice for G_0. We use a multivariate Gaussian $MVN(\mathbf{0}, \Sigma_\mu)$ as the prior for μ_k. To illustrate the effect of variance parameter on a clustering model, several parameterisation for Σ are presented in Table 1.

Assumption	**Structure**	**Prior**
Fixed	$M_1 : \frac{1}{\sigma} I_{d\times d}$	–
	$M_2 : \Sigma$	–
Unknown (global)	$M_3 : \frac{1}{\sigma} I_{d\times d}$	$\sigma \sim \Gamma(g_1, g_2)$
	$M_4 : \Sigma$	$\Sigma \sim IW(\nu_0, V_0)$
	$M_5 : \Sigma^{-1} = LL^t$	$L_{ij} \sim N(\mu_0, \sigma_0), L_{ii}^2 \sim \Gamma(a_0, b_0)$
Unknown (cluster-specific)	$M_6 : \Sigma_k$	$\Sigma_k \sim IW(\nu_0, V_0)$
	$M_7 : \Sigma_k$	$\Sigma_{k_{ij}}^{-1} \sim Lap(0, c_0), \Sigma_{k_{ii}}^{-1} \sim \Gamma(a_0, b_0)$
	$M_8 : \Sigma_k$	$\Sigma_{k_{ij}}^{-1} \sim N(c_0, 10^{-6}), \Sigma_{k_{ii}}^{-1} \sim \Gamma(a_0, b_0)$

Table 1 Model structure based on choice of Σ; σ is a scalar quantity, L is the Cholesky factorized lower triangular matrix for $\Sigma^{-1} = LL^t$ and $i, j \in 1, \ldots, d$

We introduce gradual complexity in terms of covariance matrix choices, starting from known or fixed variance (M_1, M_2). For unknown variance, global refers to a common covariance matrix Σ for all cluster choices (M_3, M_4, M_5), whereas cluster-specific defines each cluster with a unique covariance matrix Σ_k, along with the mean vectors (M_6, M_7, M_8). The choice of prior distributions maintain the conjugacy of the models, while considering hyper-parameters that establish weakly informative priors.

2.4. Posterior estimation using collapsed VI

We implement a collapsed variational inference (CVI) approach, where we integrate out (or collapse) the stick-breaking parameters $\{V_k\}$, following Kurihara et al. (2007). This yields a distribution of the latent allocation vectors $\{z_n\}$ depending only on the DP concentration parameter α:

$$\begin{aligned} p(z_n|\alpha) &= \prod_k \int_{V_k} p(z_n|V_k)p(V_k|\alpha)dV_k \\ &= \prod_k \alpha \frac{\Gamma(\mathbf{1}[z_n = k] + 1)\Gamma(\mathbf{1}[z_n > k] + \alpha)}{\Gamma(1 + \mathbf{1}[z_n \geq k] + \alpha)} \end{aligned} \tag{1}$$

where $\mathbf{1}[.]$ represents the indicator function.
We can reconstruct our DPMM as:

$$\begin{aligned} \alpha &\sim Gamma(a, b) \\ \mu_k, \Sigma_k &\sim G_0 \\ z_n &\sim p(z_n|\alpha) \ (1) \\ x_n|z_n &\sim MVN(\mu_{z_n}, \Sigma_{z_n}). \end{aligned}$$

The posterior distribution is approximated by a mean-field variational distribution for the DPMM parameters

$$q(\alpha, \{\mu_k\}, \{\Sigma_k\}, \{z_n\}) = q(\alpha) \prod_{k \leq K} q(\mu_k)q(\Sigma_k) \prod_n q(z_n)$$

(Blei et al., 2017; Zhang et al., 2018). Although having theoretically "infinite" prior choices for k, truncated variational distributions with an upper limit $k = K$ enable practical implementation (Ishwaran and James, 2001).

For posterior estimation, we use exponential families of distributions for $q(\cdot)$ and apply coordinate ascent algorithm to update the variational hyper-parameters (Bishop and Nasrabadi, 2006). It ensures guarantee of convergence and provides a closed form update $q^*(\cdot)$ that maximizes the Evidence lower bound (ELBO) (or equivalently minimizing the information-theoretic Kullbach-Leibler (KL) divergence)

$$\begin{aligned} q^*(\theta_i) &\propto \exp(\mathbb{E}_{q_{\theta^{-i}}} [\log\, p(X, \theta)]) \\ &\propto \exp(\mathbb{E}_{q_{\theta^{-i}}} [\log\, p(X|\theta) + \log\, p(\theta)]) \end{aligned} \tag{2}$$

where $\theta_j \in \mathbf{\Theta} = \{\alpha, \{\mu_k\}, \{\Sigma_k\}, \{z_n\}\}$ and $\theta^{-j} = \mathbf{\Theta} \backslash \theta_j$ represents the set of all except the j^{th} parameter (Blei and Jordan, 2006; Bishop and Nasrabadi, 2006). Appropriate prior as well as variational distribution choices facilitate the update of $q^*(\cdot)$ hyper-parameters. These updates are used to perform posterior inference for the unknown $\{z_n\}, \{\theta_i\}$, thus estimating data cluster as well as distributional partitions corresponding to the cluster allocations.

2.4.1. Latent allocations $\{z_n\}$

After collapsing the stick-breaking parameters, the prior conditional distribution of z_n can be written as

$$p(z_n = k|\mathbf{z}^{-n}) = \frac{1+N_k^{-n}}{1+\alpha+N_{\geq k}^{-n}} \prod_{j<k} \frac{\alpha+N_{>j}^{-n}}{1+\alpha+N_{\geq j}^{-n}}$$

(Kurihara et al., 2007). So, the variational distribution updates as

$$q^*(z_n) \propto \exp\Bigg(\sum_{k \leq K} \mathbf{1}[z_n = k] \mathbb{E}_{q^{-z_n}} \Big[\log\, p(z_n = k | \mathbf{z}^{-n}) + log(p(x_n | \mu_k, \Sigma_k) \Big] \Bigg). \tag{3}$$

$q^*(z_n)$ follows a Categorical distribution with probability hyper-parameters q_{nk}:

$$\begin{aligned} q_{nk} \propto \exp\Bigg[& \log(1 + \mathbb{E}_q[N_k^{-n}]) - \frac{\mathbb{V}_q[N_k^{-n}]}{(1 + \mathbb{E}_q[N_k^{-n}])^2} \\ & - \log(1 + \mathbb{E}_q[N_{\geq k}^{-n}] + \frac{w_1}{w_2}) + \frac{\mathbb{V}_q[N_{\geq k}^{-n}] + \frac{w_1}{w_2^2}}{(1 + \mathbb{E}_q[N_{\geq k}^{-n}] + \frac{w_1}{w_2})^2} \\ & + \sum_{j<k} \Bigg(\log(\frac{w_1}{w_2} + \mathbb{E}_q[N_{>j}^{-n}]) - \frac{\mathbb{V}_q[N_{>j}^{-n}] + \frac{w_1}{w_2^2}}{(\frac{w_1}{w_2} + \mathbb{E}_q[N_{>j}^{-n}])^2} \\ & - \log(1 + \frac{w_1}{w_2} + \mathbb{E}_q[N_{\geq j}^{-n}]) + \frac{\mathbb{V}_q[N_{\geq j}^{-n}] + \frac{w_1}{w_2^2}}{(1 + \frac{w_1}{w_2} + \mathbb{E}_Q[N_{\geq j}^{-n}])^2} \Bigg) \\ & + \mathbb{E}_q[\log\, p(x_n | \mu_k, \Sigma_k)] \Bigg] \end{aligned} \tag{4}$$

where $N_k = \sum_n \mathbf{1}[z_n = k]$, $N_{>k} = \sum_n \mathbf{1}[z_n > k]$ and $N_{\geq k} = N_k + N_{>k}$. The calculations of their corresponding expectations and variances follow from Kurihara et al. (2007) (details in Supplementary material).

2.4.2. Distributional parameters $\{\mu_k, \Sigma_k\}$

We apply multivariate Gaussian distributions, both as prior and variational $q(\cdot)$, for μ_k such that

$$\begin{aligned} p(\mu_k) &\sim MVN(\mu_0 = \mathbf{0}, \Sigma_\mu) \\ q(\mu_k) &\sim MVN(\phi_k, \Lambda_k). \end{aligned}$$

The updated $q^*(\mu_k)$ depends on $p(\mu_k)$ as well as $p(X_n = x_n | z_n)$, and hence the structure of Σ_k plays an important role. For instance, when we assume Σ_k to be fixed and consider $\Sigma_k = \Sigma = \frac{1}{\sigma} I_{d \times d}$, the variational update of μ_k follows Blei and Jordan (2006). On the other hand, if we consider $\Sigma_k = \Sigma$ as a full matrix, the updated hyper-parameters of μ_k can be written as

$$\begin{aligned} \Lambda_k &= \left(\Sigma_\mu^{-1} + (\sum_n q_{nk}) \Sigma^{-1} \right)^{-1} \\ \phi_k &= \Lambda_k \Sigma^{-1} \left(\sum_n q_{nk} x_n \right) \text{(details in Supplementary material)} \end{aligned} \tag{5}$$

Analogous results are observed for unknown Σ with global structures. Owing to the mean-field variational family, the expression of Σ^{-1} in Eq:(5) is replaced by $\mathbb{E}_q[\Sigma^{-1}]$ with conjugate choices for $q(\Sigma)$ (details in Supplementary material).

For cluster-specific unknown Σ_k, the prior distribution of μ_k is conditionally dependent on Σ_k so that $p(\mu_k | \Sigma_k) \sim MVN(\mu_0 = \mathbf{0}, \frac{1}{k_0} \Sigma_k)$, where k_0 is a fixed scaling factor and influences the scale of μ_k's. For full

matrix structure, we apply $q(\Sigma_k) \sim IW(\nu_k, V_k)$ and the hyper-parameters are updated as

$$\begin{aligned}
\nu_k &= \nu_0 + 1 + \sum_n q_{nk} \\
V_k &= V_0 + \sum_n q_{nk} x_n x_n^t \\
\Lambda_k &= \frac{\mathbb{E}_q[\Sigma_k^{-1}]^{-1}}{k_0 + \sum_n q_{nk}} \\
\phi_k &= \frac{\sum_n q_{nk} x_n}{k_0 + \sum_n q_{nk}}
\end{aligned} \tag{6}$$

An alternative is to use element wise distributions on the precision matrix Σ_k^{-1} as mentioned in Table:1. A simplification of the identity $det(e^{\Sigma_k^{-1}}) = e^{tr(\Sigma_k^{-1})}$ (details in Supplementary material) facilitates the use of Gamma distributions for diagonal elements of Σ_k^{-1}. If we consider Laplace distribution for the off-diagonal elements, we obtain the following update

$$\begin{aligned}
&p(\Sigma_{k_{ii}}^{-1}) \sim \Gamma(a_0, b_0), q(\Sigma_{k_{ii}}^{-1}) \sim \Gamma(a_{k_i}, b_{k_i}) \\
&p(\Sigma_{k_{ij}}^{-1}) \sim Lap(0, c_0), q(\Sigma_{k_{ij}}^{-1}) \sim Lap(0, c_{k_{ij}}) \\
&a_{k_i} = a_0 + \sum_n q_{nk} + 1 \\
&b_{k_i} = b_0 + \frac{1}{2} \sum_n q_{nk} x_{ni}^2 \\
&c_{k_{ij}} = \left(\frac{1}{c_0} + \frac{1}{2} \sum_n q_{nk} |x_{ni} x_{nj}| \right)^{-1}
\end{aligned} \tag{7}$$

and for Gaussian distributed off-diagonal elements, we get

$$\begin{aligned}
&p(\Sigma_{k_{ij}}^{-1}) \sim N(c_0, 10^{-6}), q(\Sigma_{k_{ij}}^{-1}) \sim N(c_{k_{ij}}, 10^{-6}) \\
&c_{k_{ij}} = c_0 - \frac{10^{-6}}{2} \sum_n q_{nk} x_{ni} x_{nj}
\end{aligned} \tag{8}$$

where $i, j \in \{1, 2, ..., d\}$ and $i > j$ (details in Supplementary material). Update of μ_k remains as shown in Eq:(6) for both the cases.

2.4.3. Concentration parameter α

We use Gamma distributions as the prior and variational distribution for α

$$\begin{aligned}
p(\alpha) &\sim \Gamma(a, b) \\
q(\alpha) &\sim \Gamma(w_1, w_2)
\end{aligned}$$

so that

$$q^*(\alpha) \propto \exp(\mathbb{E}_{q^{-\alpha}}[\log\, p(\alpha)] + \mathbb{E}_{q^{-\alpha}}[\log\, p(\mathbf{Z}|\alpha)])$$

where $\mathbf{Z} = \{z_n\}$. Applying Taylor series approximation and Euler-Maclaurin formula, we obtain the estimates of w_1 and w_2 (details in Supplementary material).

$$w_1 = a + t - 1 \tag{9a}$$

$$\begin{aligned}
w_2 = b + \sum_{k<t} \Bigg[&\log(a_0 + \mathbb{E}_q[N_{\geq k}]) - \frac{\mathbb{V}_q[N_{\geq k}]}{(a_0 + \mathbb{E}_q[N_{\geq k}])^2} - \log(a_0 + \mathbb{E}_q[N_{>k}]) \\
&+ \frac{\mathbb{V}_q[N_{>k}]}{(a_0 + \mathbb{E}_q[N_{>k}])^2} \Bigg] + \log(a_0 + \mathbb{E}_q[N_t]) - \frac{\mathbb{V}_q[N_t]}{(a_0 + \mathbb{E}_q[N_t])^2} - \log(a_0 + 1)
\end{aligned} \tag{9b}$$

where t is the position of last non-zero cluster.

2.5. Implementation

Starting with an initial choice of latent allocation probabilities q°_{nk}, the DPMM iteratively updates VI parameters based on the coordinate ascent algorithm (Eq: (2)) until there is no significant change in ELBO (Blei and Jordan, 2006)

$$\begin{aligned}\text{ELBO} &= \mathbb{E}_q\Big[\log\, p(X, Z, \{\mu_k, \Sigma_k\}, \alpha)\Big] - \mathbb{E}_q\Big[\log\, q(Z, \{\mu_k, \Sigma_k\}, \alpha)\Big] \\ &= \mathbb{E}_q\Big[\log\, p(X|Z, \{\mu_k, \Sigma_k\}, \alpha) + \log\, p(Z, \{\mu_k, \Sigma_k\}, \alpha)\Big] \\ &- \mathbb{E}_q\Big[\log\, q(Z, \{\mu_k, \Sigma_k\}, \alpha)\Big].\end{aligned} \tag{10}$$

ELBO acts as a lower bound of log evidence $logp(X)$ which minimizes reverse KL divergence between the target and variational distributions. The objective of VI is to update the variational parameters such that ELBO is maximized, thus approximating $p(\cdot)$ with $q(\cdot)$.

Due to non-convexity of the ELBO, VI is sensitive to the choice of initial parameters (Blei et al., 2017). We circumvent this by choosing random initialization, and select the optimum starting values based on variational log likelihood (VLL) term $\mathbb{E}_q[\log\, p(X|Z, \{\mu_k, \Sigma_k\}, \alpha)](= \mathbb{E}_q[\log\, p(X|Z)])$ from Eq: (10). Although ELBO explicitly regularizes model complexity by including negative entropy of $q(\cdot)$ via $-\mathbb{E}_q[\log\, q(Z, \{\mu_k, \Sigma_k\}, \alpha)]$, it is effected numerically especially for higher dimensions and can over-influence the ELBO values (details in Supplementary material). The VLL, on the other hand, implicitly handles model complexity through expectation over $q(\cdot)$, while providing a numerical fit of the data based on updated variational distribution. Thus, it makes sense to select the initial VI parameters that maximizes VLL.

3. Results

The proposed DPMM integrates a hierarchical approach in terms of varying the DP concentration parameter α in a collapsed VI method. The variational Gamma hyper-parameter w_1 depends on the number of non-zero clusters t present in $\mathbf{Z}$ (9a), instead of total variational clusters K. For a Mean field VI (Blei and Jordan, 2006), the distribution of $\mathbf{Z}$ depends on α via the mixing proportions $\{\pi_k\}$'s, which can be non-zero for empty clusters (smaller proportions for unlikely clusters). Owing to the hierarchical dependency between α and $\{\pi_k\}$'s, the variational update of α thus considers all the plausible clusters based on non-zero $\{\pi_k\}$'s (which is often equal to K), rather than only the non-zero clusters t. This leads to an inconsistency between the number of clusters estimated between α and Z (details in Supplementary material). However, for a collapsed VI, the distribution of $\mathbf{Z}$ directly depends on α (1). And with the theoretical approximations (2.4.3), our approach is able to maintain consistent estimates of cluster number through the variational updates of α hyper-parameters (9a), (9b).

The cluster estimates also depend on the structure as well as prior choice of G_0 variance (Hejblum et al., 2019). A common covariance structure (models M_3, M_4 and M_5 from Table 1) can prove inadequate to recover clusters with heterogeneous covariances (Biernacki et al., 2002). Alternatively, a cluster-specific covariance structure (models M_6, M_7 and M_8) is favoured, which integrates within-cluster variance in the model and complements the effect of α (Fig: 1). For cluster-specific covariance structure, an Inverse-Wishart prior often incorporates rigidity in the model (M_6) (O'Hagan and Forster, 2004; Tokuda et al., 2025). We introduced independence between variance and covariance terms by assigning element-wise prior distributions (models M_7 and M_8) and applied conjugate $\Gamma(.)$ distributions for the diagonal elements. For the off-diagonal elements, we considered Gaussian distributions with arbitrarily small fixed variance (model M_8) and 0-centered Laplace distribution (model M_7).

3.1. Effect of G_0 Covariance Structure and Prior Choice

To compare the different model choices given in Table:1, we generated random $N = 100$ multivariate Gaussian variables of dimension $d = 2$ and grouped them into $K_{true} = 2$ clusters. Using our developed R-package **vimixr**, we ran the models over generated data and calculated posterior number of clusters K_{post} as well as adjusted Rand index (ARI) scores for 100 simulation runs.

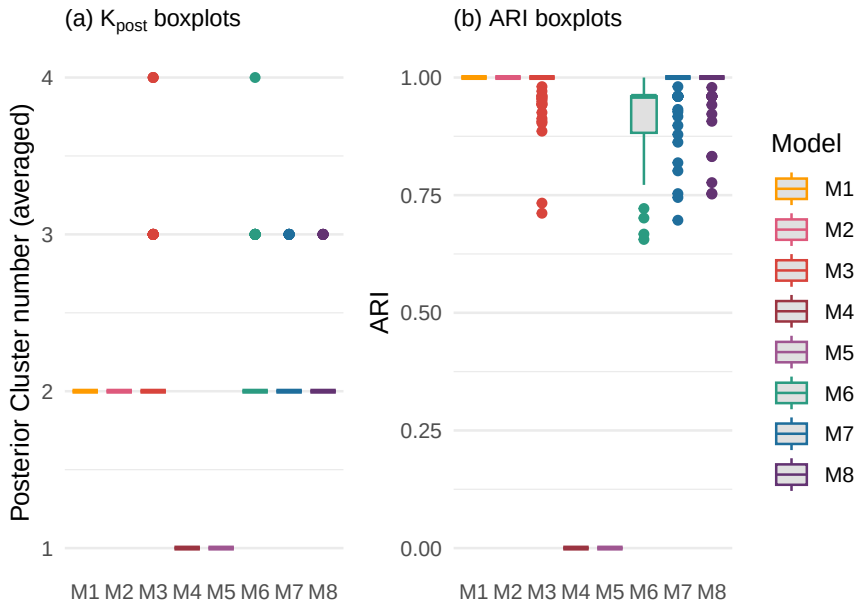

Figure 1 K_{post} and ARI scores for different model choice

In Figure 1, the fixed variance models M_1 with diagonal and M_2 with full covariance matrices, respectively, act as known parameters and set the baseline for performance comparison. For unknown variance, global covariance structure shows comparative performance with diagonal covariance model $M3$ and significantly poor performance with both full covariance model $M4$ and Cholesky-decomposed covariance model $M5$. $M3$ employs uncorrelated mixture distribution for all the variational clusters, thus reducing rigidity and variance-covariance coupling. Compared to global covariance, cluster-specific covariance structure provides better results.

For different prior choices of the cluster-specific covariance structure, we further illustrated the effect of dimensionality and higher clusters by increasing dimension to $d = 100$ and calculating K_{post} and ARI for different $K_{true} = \{4, 6, 8, 10\}$.

Element-wise prior models show better adaptability with varying cluster number than IW prior model in a high-dimensional setting (Figure 2). M_7 and M_8 have comparable performances (with M_7 slightly better than M_8). This can be attributed to the fact that a Laplace distribution is equivalent to a marginalized Gaussian distribution with Exponential prior on the variance parameter. Owing to its sparsity inducing effect (Jing et al., 2024), we call M_7 as the Sparse DPMM, and consider it to be the best possible choice among the models defined in Table 1.

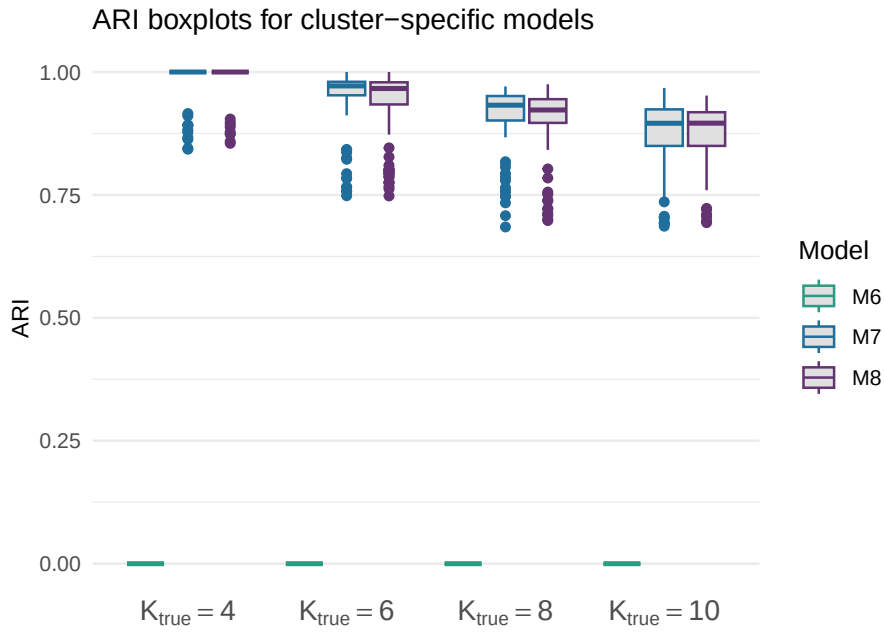

Figure 2 ARI for different cluster-specific models when K_{true} is varying

3.2. Effect of hyper-parameters of cluster distributions $\{\mu_k, \Sigma_k\}$

Real high-dimensional data suffers from the *curse of dimensionality* due to low signal-to-noise ratio and limited sample size. Sparse DPMM is equiped with hyper-parameters of $\{\mu_k, \Sigma_k\}$ that can weed out features with weaker contributions, as well as increase the effective sample size, thus reducing model variability which is otherwise an issue for *low-n-high-d* settings. Due to standard procedure of normalising such data sets (Xia, 2023), we chose $\mu_0 = \mathbf{0}$ as the prior hyper-parameter for μ_k and $a_0 = b_0$ for the diagonal elements of Σ_k^{-1} (2.4.2). Along with $a_0 = b_0$, the scaling factor k_0 for $p(\mu_k|\Sigma_k) \sim MVN(\mathbf{0}, \frac{1}{k_0}\Sigma_k^{-1})$ (2.4.2) plays a significant role in defining the posterior cluster distributions via their variational updates (7), thus effecting posterior estimation of K.

We performed the sensitivity analysis with $N = 100, d = 1000$ Negative Binomial variables clustered into $K_{true} = 3$ groups to illustrate the robustness of Sparse DPMM on non-Gaussian data with more than 2 clusters. For 1000 random initialisations, we compared the VLL values obtained for a) different $a_0 = b_0$ with fixed $k_0 = 1$ (Figure 3(a)) and b) different k_0 with fixed $a_0 = b_0 = 0.001$ (Figure 3(b)).

For Sparse DPMM, the derivatives of VLL with respect to $a_0 = b_0$ and k_0 can be expressed as

$$\frac{\delta}{\delta a_0} VLL = \frac{1}{2}\sum_k \Bigg(d\sum_n q_{nk}^{\circ}\Psi^1(a_0 + \sum_n q_{nk}^{\circ}) - \sum_i \Bigg[\frac{a_0 \sum_n q_{nk}^{\circ} + (\frac{\sum_n q_{nk}^{\circ} x_{n_i}^2}{2})^2 - \frac{\sum_n q_{nk}^{\circ} \sum_n q_{nk}^{\circ} x_{n_i}^2}{2}}{(a_0 + \frac{1}{2}\sum_n q_{nk}^{\circ} x_{n_i}^2)^2} \Bigg] \Bigg) \tag{11a}$$

$$\frac{\delta}{\delta k_0} VLL = \sum_k \Bigg[\frac{1}{(k_0 + \sum_n q_{nk}^{\circ})^2} \Bigg(d\sum_n q_{nk}^{\circ} - (2 - \sum_n q_{nk}^{\circ}) \sum_i (\sum_n q_{nk}^{\circ} x_{n_i})^2 \Bigg) \Bigg] \tag{11b}$$

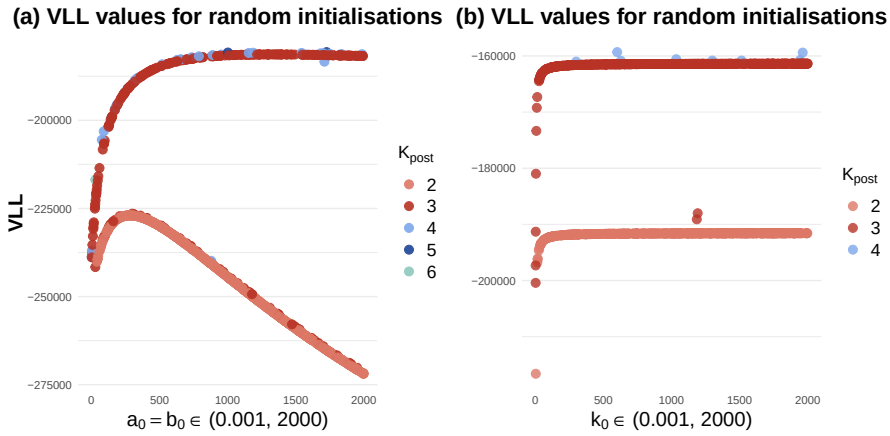

Figure 3 Sensitivity analysis of Sparse DPMM over the hyper-parameters 3(a) $a_0 = b_0$ and 3(b) k_0 for Negative Binomial simulations with $K_{true} = 3$

where $\{q^{\circ}_{nk}\}$ are a random allocation probability values and $\Psi^1(\cdot)$ is the *trigamma* function (details in Supplementary material). Functional behaviour of $\frac{\delta}{\delta a_0}VLL$ relies on $\{q^{\circ}_{nk}\}$, and VLL can either attain maxima or converge to a stabilising value with increasing $a_0 = b_0$ (Figure 3(a)). For both the conditions, we estimate a_0 for which either $\frac{\delta}{\delta a_0}VLL = 0$ or the maximum curvature point of $\frac{\delta}{\delta a_0}VLL$ that corresponds to the stabilising convergence of VLL (details in Supplementary material). In either case, the estimated a_0 represents the optimal value after which there is no further gain in the log likelihood, hence no additional prior information can be incorporated with higher a_0. Thus, for a random choice of $\{q^{\circ}_{nk}\}$, a_0 estimation inherently puts an upper bound on the degree of information required for maximising the VLL. This enables the model to threshold dimensions that have stronger contributions to effect the posterior estimation, and subsequently the VLL for a given data.

On the other hand, behaviour of $\frac{\delta}{\delta k_0}VLL$ is converging, and the value of convergence depends on $\{q^{\circ}_{nk}\}$ (Figure 3(b)). k_0 acts as the equivalent sample size of the μ_k prior (Murphy, 2007), and scales the variational hyper-parameters of μ_k in Eq: (6). So, a smaller value of $\frac{1}{k_0}$ reduces the variability of μ_k which is equivalent of having a larger data sample. As the variational updates of μ_k in Eq: (6) is a weighted average between cluster-specific samples and prior mean, a natural choice is to consider $k0 = N + 1$ where N is the data sample size.

3.3. Computational scalability

We implemented our model framework in an R-package **vimixr** with all the 8 parameterisations given in Table 1. For Sparse DPMM, the computation scales with sample size N as order $\mathcal{O}(Nlog(N))$ due to the allocation variable $\mathbf{Z}$ updates with conditional expectation on $z_n|\mathbf{z}^{-n}$. The computational dependence for dimension d is of order $\mathcal{O}(dlog(d)log(log(d))$ (details in Supplementary material). Although higher dimensions achieve convergence with less number of iterations, the rate is slightly higher due to computation of cluster-specific d-dimensional precision matrices.

For a slice-sampling based MCMC approach, we used the function *DPMGibbsN* from **NPflow** R package (Hejblum et al., 2019) and compared our Sparse DPMM for $N = 100, d = 100$ Gaussian data. Even with 1000 MCMC iterations, *DPMGibbsN* often fails to achieve convergence. Sparse DPMM, on the other hand, demonstrates an increase in computational time by a magnitude of two orders (~ 100 times) during bench-marking (details in Supplementary material).

3.4. Cancer sub-type estimation

We applied the proposed clustering framework on the gene-expression data set from leukemia sub-type study by Armstrong et al. (2002), previously used as a benchmark for comparative analysis of clustering algorithms De Souto et al. (2008). Armstrong et al. (2002) showed that leukemia with rearrangement of the MLL gene (mixed lineage leukemia gene, now renamed KMT2A) should be classified as a distinct clinical entity rather than common acute lymphoblastic leukemia. The chromosomal rearrangement of MLL creates fusion genes with various partner genes that influence lineage commitment (Krivtsov and Armstrong, 2007). For example, certain MLL rearrangements result in overexpression of **HOXA9** and **PRG1** genes, which are characteristic markers of acute myeloid leukemia (AML) rather than ALL (Armstrong et al., 2002). Recent studies have further revealed the lineage plasticity of MLL-rearranged leukemia, showing dynamic transitions between lymphoid and myeloid phenotypes (Chen et al., 2022; Janssens et al., 2024).

Working on the Affymetrix data from Armstrong et al. (2002), De Souto et al. (2008) provides a filtered set of $d = 2194$ expressed genes for clustering $N = 72$ leukemia samples into ALL-MLL-AML sub-types (Fig 4(a)). For a strong choice of prior hyper-parameters corresponding to Section 3.2, we obtained $K_{post} = 3$ with ARI score of 0.92 (Figure 4(b)).

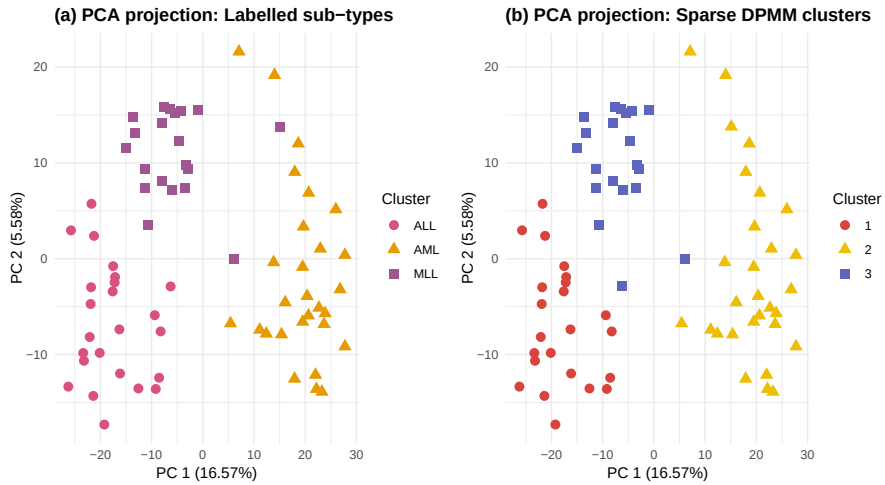

Figure 4 PCA projection on the first 2 principal components for (a) labelled Leukemia sub-types based on 2194 genes and (b) Sparse DPMM cluster estimates obtained using strong hyper-parameters

Following our empirical Bayes approach in Section 3.2, we estimated $a_0 = b_0 = 28.9076$, and along with $k_0 = N + 1 = 73$, we applied Sparse DPMM for random initial probability allocation values and chose the best performing model based on VLL. To check the sensitivity of Sparse DPMM with a weaker prior hyper-parameter, we used $a_0 = b_0 = 10$ instead of 50 and clustered the same data set. For this particular choice of $a_0 = b_0$, we obtained $K_{post} = 4$ with ARI score 0.86. (Figure 5).

Comparing with known sub-types in Figure 4(a), Sparse DPMM with weaker hyper-parameters identifies an additional 4^{th} cluster with one sample replacing a labelled AML sub-type (Figure 5). Analogous to (Armstrong et al., 2002), we plot a heat-map based on the available genes (De Souto et al., 2008) that are specifically expressed for the known sub-types to compare the lineage of the 4^{th} estimated cluster (Figure 6). The heat-map reveals a mixed expression pattern for the 4th cluster. The cluster shows elevated expression of genes characteristic of ALL sub-types (**CD22** and **CD24**), suggesting lymphoblastic features. It also displays moderate expression of genes typically overexpressed in MLL sub-types (**PROML1**, **FLT3** and **ADAM10**). In contrast, genes predominantly expressed in AML sub-types (**DF**, **CTSD** and **BSG**) are not dominantly expressed in this cluster. However,

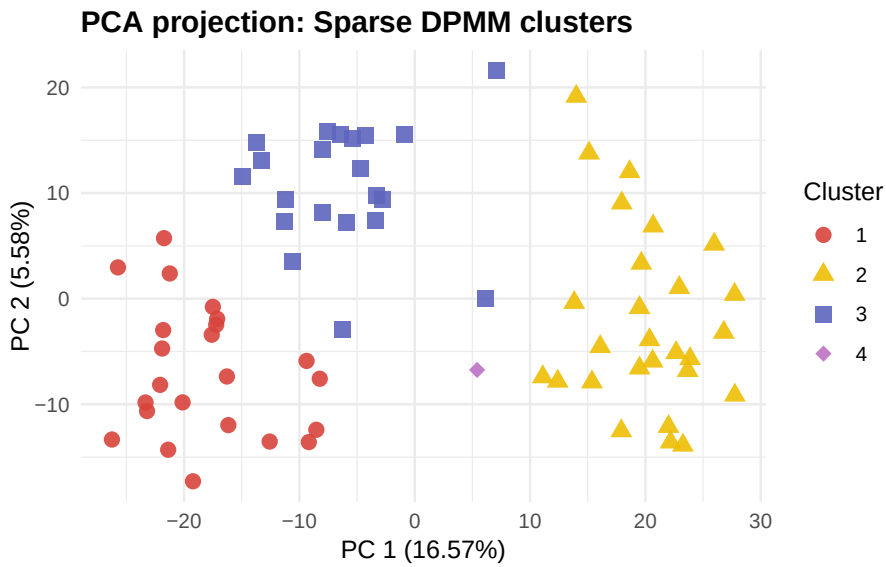

Figure 5 PCA projection on the first 2 principal components for Sparse DPMM cluster estimates obtained using weaker hyper-prarameters

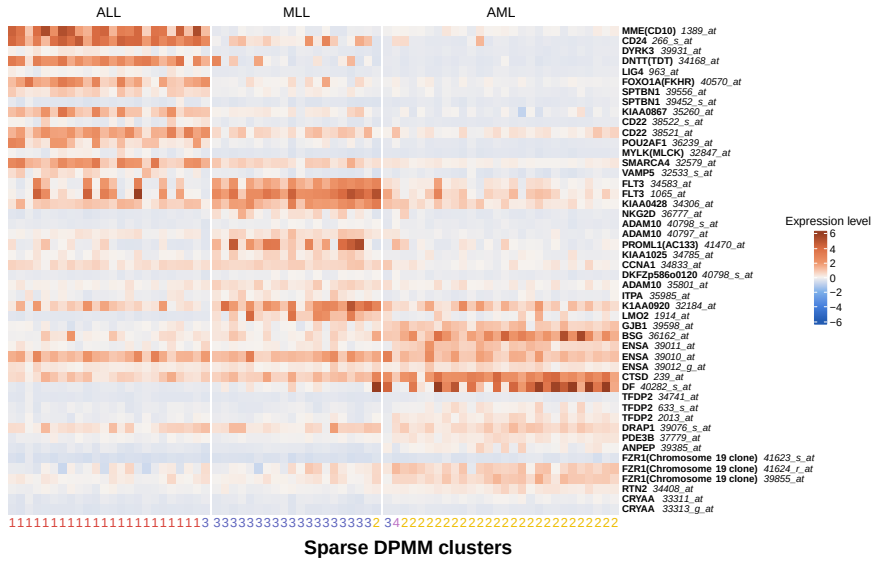

Figure 6 Heatmap showing comparison between labelled sub-types and Sparse DPMM (weaker) cluster estimates with respect to specifically expressed genes for the labelled sub-types (Armstrong et al., 2002) for all the 72 samples (genes in bold, probe-id in italics)

this expression pattern is not clear-cut, with the 4th cluster displaying characteristics intermediate between ALL and MLL rather than fitting cleanly into a single category. This ambiguous profile is consistent with the documented lineage plasticity of MLL-rearranged leukemias, which can undergo dynamic transitions between lymphoid and myeloid phenotypes during disease progression and treatment (Chen et al., 2022; Janssens et al.,

2024). The identification of this 4th cluster by Sparse DPMM thus appears to capture this biological reality of MLL heterogeneity rather than representing a technical artifact.

3.4.1. Performance benchmarking with current state of the art clustering techniques

For high-dimensional clustering analysis, there are certain popular choice of working algorithms with varying implementation techniques. Methods like DBSCAN (Ester et al., 1996) and HDBSCAN (Campello et al., 2013) are based on density of the data distributed over dimensional space, where data points in the low-density region are identified as noise/outliers and form clusters using core points. Shared nearest neighbours based method (Ertoz et al., 2002) works by grouping data points with higher similarity of shared neighbours across a k-nearest neighbour graph. Bouveyron et al. (2007) proposed a high dimensional model based clustering that implements dimensionality reduction for each cluster with parsimonious covariance structure and estimate the Gaussian mixture model parameters using EM iterations. Another class of popular methods include graph based algorithms, like Leiden (an upgrade of Louvain) (Traag et al., 2019) that constructs a k-nearest neighbour graph from the data and detects communities (clusters) by maximizing graph modularity. Classical K-means (MacQueen, 1967), on the other hand, is a partitioning based method that minimizes the within-cluster variance by iteratively assigning data points to their nearest centroids.

For performance on the Leukemia data set, we implemented the methods in R and chose the key parameters involved with these methods based on internal validation metrics specific to each method (details in Supplementary material). Based on corresponding optimal parameters, these methods are compared with Sparse DPMM implementation with empirical Bayes hyper-parameters, and we validated their performances with run-time (in seconds), estimated cluster numbers and ajusted Rand index (ARI). Due to diversity in methodology of these techniques, we avoided any internal metric for comparison (due to possible bias for particular technique(s)) and focused on external validation metrics. Figure 7 shows that Sparse DPMM is significantly slower ($\sim$ 8.091 seconds per iteration) in comparison to alternative methods. However, the performance metrics based on estimated number of clusters and ARI scores validate the (relative) superiority of Sparse DPMM.

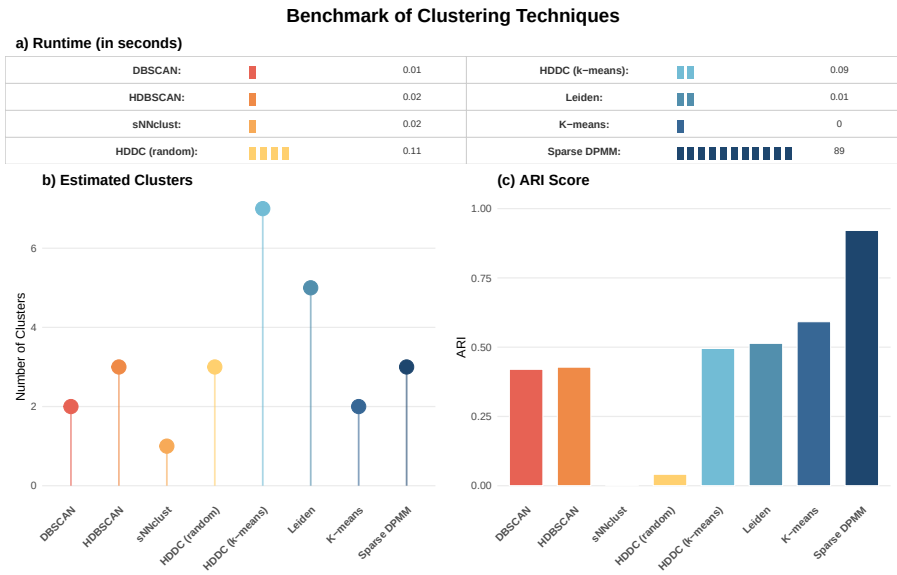

Figure 7 Performance benchmarking of popular clustering methods against Sparse DPMM; based on *a*) implementation run-time in seconds (iterations are indicated with mini-bars), *b*) estimated number of clusters and *c*) corresponding ARI scores

4. Discussion

We developed a Bayesian non-parametric mixture model using collapsed variational inference (VI) to perform unsupervised clustering. Our approach hierarchically incorporates the Dirichlet process (DP) concentration parameter α as a variable in the DP mixture model (DPMM) with Gaussian kernels. Based on the importance of adaptive estimation of α (Ascolani et al., 2023), the proposed framework enhances the consistency of collapsed VI model developed by Kurihara et al. (2007). We also illustrated the effect of prior for covariance matrix, and concluded that a cluster-specific Sparse DPMM is the best choice. Implementing our theoretical work into an R-package **vimixr**, the user is allowed to choose the model and prior choice based on parameterisations given in Table 1.

The Sparse DPMM provides significantly faster convergence than an MCMC slice sampling approach (Hejblum et al., 2019). Although the convergence speed varies with the dimension d due to cluster-specific covariance structure, the Sparse DPMM performs robustly well for a large range of samples N, dimensions d and true clusters K_{true} present in the data. Comparing with conventional clustering techniques, the iterative nature of Sparse DPMM, added to the model structure complexity, leads to a slower convergence than the conventional techniques considered. However, it provides better clustering estimates with empirical Bayes hyper-parameters using gene expression data for leukemia sub-types (Armstrong et al., 2002). Even with weaker hyper-parameter values, Sparse DPMM estimates are consistent with gene expression profiles, hence providing meaningful sub-clusters of the already known sub-types.

A natural extension of Sparse DPMM is to generalize beyond Gaussian kernels for the DP base distribution parameter G_0, which becomes eminent for real-world data. For instance, overdispersed count data is better modelled using Negative Binomial distributions (Fernandez and Vatcheva, 2022). So, incorporating Negative Binomial kernels can aid in avoiding model misspecification for clustering such data sets like gene expression data (Anders and Huber, 2010). Another aspect is to introduce probabilistic feature selection variables that could be updated simultaneously (Tadesse et al., 2005) or guide the clustering process itself (Rouanet et al., 2024). This can provide insights on the features defining the estimated clusters and prove beneficial, both computationally and inferentially, for high-dimensional data sets.

As an optimisation technique, VI algorithms heavily depend on the initial choice of latent variables (latent allocation vectors $\{z_n\}$ in our case) for a given data. We advocate random initialisation for these parameters, and the choice is based on variational log likelihood (VLL) values. For VI implementation, alternative approaches (Zhang et al., 2018) can be explored for further refinement. We have also shown that the hyper-parameters' choice for G_0 significantly effect the cluster estimation. Based on the analytical properties of VLL, empirical Bayes estimates for the hyper-parameters prove to be efficient for robust estimation. Due to their implied significance over effective dimensions and sample size, it becomes important to consider hierarchical hyper-priors for the hyper-parameter $a_0 = b_0$. However, the conjugate prior for both unknown shape and rate parameters of a Gamma distribution does not have a closed form (Miller, 1980). So, incorporating an $a_0 = b_0$ hyper-prior would add another layer of analytical complexity, along with increased computational cost due to additional model parameters.

The exchangeability of mixture components in DPMM introduces label switching between atoms as they are ordered sequentially through the stick-breaking construction (Papaspiliopoulos and Roberts, 2008; Hastie et al., 2015). Following Kurihara et al. (2007), we addressed the issue by ordering the latent allocation probability updates such that the clusters are arranged in decreasing proportions. Alternatively, Frühwirth-Schnatter (2011) shows identification methods for finite mixture model, that could be extrapolated to non-parametric mixtures.

5. Conflicts of interest

The authors declare that they have no competing interests.

6. Funding

This work is supported by funds from the Programmes et équipements prioritaires de recherche (PEPR) Santé Numérique SMATCH Task 2.N1, under the Agence nationale de la recherche (ANR), France.

7. Software and Data availability

Software in the form of R package **vimixr** with complete documentation is available on CRAN (`https://cran.r-project.org/web/packages/vimixr/index.html`). The Leukemia sub-type gene expression data is publicly available online at `https://schlieplab.org/Static/Supplements/CompCancer/Affymetrix/armstrong-2002-v2/`. The package implementation is provided in the Zenodo repository `https://doi.org/10.5281/zenodo.18405688`, along with metadata and results.

8. Supplementary Material

Supplementary material is available online at `http://biostatistics.oxfordjournals.org`.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions. This work was supported by a French government grant managed by the Agence Nationale de la Recherche (ANR) under the France 2030 program (reference SMATCH ANR-22-PESN-0003). For heavy computation, the study used Cluster Curta, a project by Mésocentre de Calcul Intensif Aquitain (MCIA), France (`https://redmine.mcia.fr/projects/cluster-curta`)

REFERENCES

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, 2(6):1152–1174.

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, 30(1):41–47.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2023). Clustering consistency with dirichlet process mixtures. *Biometrika*, 110(2):551–558.

Bernardi, M., Bianchi, D., and Bianco, N. (2024). Variational inference for large bayesian vector autoregressions. *Journal of Business & Economic Statistics*, 42(3):1066–1082.

Biernacki, C., Celeux, G., and Govaert, G. (2002). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bock, H. H. (1996). Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, 23(1):5–28.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional data clustering. *Computational statistics & data analysis*, 52(1):502–519.

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. PMLR.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Chandra, N. K., Canale, A., and Dunson, D. B. (2023). Escaping the curse of dimensionality in bayesian model-based clustering. *Journal of machine learning research*, 24(144):1–42.

Chen, C., Yu, W., Alikarami, F., Qiu, Q., Chen, C.-h., Flournoy, J., Gao, P., Uzun, Y., Fang, L., Davenport, J. W., et al. (2022). Single-cell multiomics reveals increased plasticity, resistant populations, and stem-cell–like blasts in kmt2a-rearranged leukemia. *Blood, The Journal of the American Society of Hematology*, 139(14):2198–2211.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A., and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature reviews cancer*, 8(1):37–49.

De Souto, M. C., Costa, I. G., De Araujo, D. S., Ludermir, T. B., and Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC bioinformatics*, 9:1–14.

Dinh, T., Wong, H., Lisik, D., Koren, M., Tran, D., Yu, P. S., and Torres-Sospedra, J. (2025). Data clustering: a fundamental method in data science and management. *Data Science and Management*.

Ertoz, L., Steinbach, M., and Kumar, V. (2002). A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on clustering high dimensional data and its applications at 2nd SIAM international conference on data mining*, volume 8.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

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 *kdd*, volume 96, pages 226–231.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, 1:209–230.

Fernandez, G. A. and Vatcheva, K. P. (2022). A comparison of statistical methods for modeling count data with an application to hospital length of stay. *BMC Medical Research Methodology*, 22(1):211.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769.

Frühwirth-Schnatter, S. (2011). Dealing with label switching under model uncertainty. *Mixtures: estimation and applications*, pages 213–239.

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*, 13:33–64.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Hastie, D. I., Liverani, S., and Richardson, S. (2015). Sampling from dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and computing*, 25(5):1023–1037.

Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F., and Thiébaut, R. (2019). Sequential dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. *The Annals of Applied Statistics*, 13:638–660.

Hennig, C., Meila, M., Murtagh, F., and Rocci, R. (2015). *Handbook of cluster analysis*. CRC press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association*, 96(453):161–173.

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

Janssens, D. H., Duran, M., Otto, D. J., Wu, W., Xu, Y., Kirkey, D., Mullighan, C. G., Yi, J. S., Meshinchi, S., Sarthy, J. F., et al. (2024). Mll oncoprotein levels influence leukemia lineage identities. *Nature Communications*, 15(1):9341.

Jing, W., Papathomas, M., and Liverani, S. (2024). Variance matrix priors for dirichlet process mixture models with gaussian kernels. *International Statistical Review*.

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kasa, S. R. and Rajan, V. (2023). Avoiding inferior clusterings with misspecified gaussian mixture models. *Scientific Reports*, 13(1):19164.

Kelly, S. T. (2023). *leiden: R implementation of the Leiden algorithm*. R package version 0.4.3.1.

Krivtsov, A. V. and Armstrong, S. A. (2007). Mll translocations, histone modifications and leukaemia stem-cell development. *Nature reviews cancer*, 7(11):823–833.

Kurihara, K., Welling, M., and Teh, Y. W. (2007). Collapsed variational dirichlet process mixture models. *20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*.

Li, Y., Schofield, E., and Gönen, M. (2019). A tutorial on dirichlet process mixture modeling. *Journal of mathematical psychology*, 91:128–144.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Loya, H., Kalantzis, G., Cooper, F., and Palamara, P. F. (2025). A scalable variational inference approach for increased mixed-model association power. *Nature Genetics*, 57(2):461–468.

MacQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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. *The Journal of Machine Learning Research*, 15(1):3333–3370.

Miller, R. B. (1980). Bayesian analysis of the two-parameter gamma distribution. *Technometrics*, 22(1):65–69.

Müller, P. and Mitra, R. (2013). Bayesian nonparametric inference–why and how. *Bayesian analysis (Online)*, 8(2):10–1214.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

O'Hagan, A. and Forster, J. J. (2004). *Kendall's advanced theory of statistics, volume 2B: Bayesian inference*, volume 2. Arnold.

Orbanz, P. and Teh, Y. W. (2010). Bayesian nonparametric models. *Encyclopedia of machine learning*, 1:81–89.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, pages 169–186.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572.

Rouanet, A., Johnson, R., Strauss, M., Richardson, S., Tom, B. D., White, S. R., and Kirk, P. D. (2024). Bayesian profile regression for clustering analysis involving a longitudinal response and explanatory variables. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(2):314–339.

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

Steinhaus, H. et al. (1956). Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci*, 1(804):801.

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. W. et al. (2010). Dirichlet process. *Encyclopedia of machine learning*, 1063:280–287.

Tokuda, T., Goodrich, B., Van Mechelen, I., Gelman, A., and Tuerlinckx, F. (2025). Visualizing distributions of covariance matrices. *Journal of Data Science, Statistics, and Visualisation*, 5(7).

Traag, V. A., Waltman, L., and Van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Wani, A. A. (2025). Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions. *PeerJ Computer Science*, 11:e3025.

West, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03.

Xia, Y. (2023). Statistical normalization methods in microbiome data with application to microbiome cancer research. *Gut microbes*, 15(2):2244139.

Yu, X., Yu, G., and Wang, J. (2017). Clustering cancer gene expression data by projective clustering ensemble. *PloS one*, 12(2):e0171429.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2018). Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026.

Zhang, H., Swallow, B., and Gupta, M. (2022). Bayesian hierarchical mixture models for detecting non-normal clusters applied to noisy genomic and environmental datasets. *Australian & New Zealand Journal of Statistics*, 64(2):313–337.