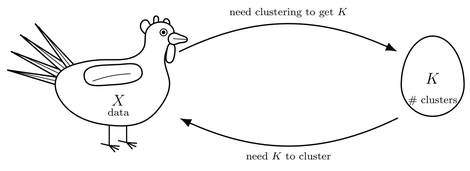
ISBA World Meeting 2026 — Nagoya, Japan

Collapsed Variational Inference for Dirichlet Process Mixtures with unknown Concentration parameter and Covariance: Scaling High Dimensional Clustering

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

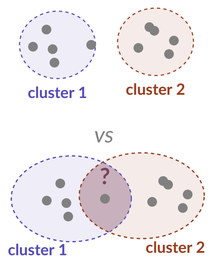
[1]Université de Bordeaux, INSERM, INRIA, BPH U1219, SISTM, 33000 Bordeaux, France [2]Vaccine Research Institute, 94010 Créteil, France [3]CHU Bordeaux, Laboratoire d'Hématologie, 33000 Bordeaux, France [4]CHU Bordeaux, Service d'Information Médicale, 33000 Bordeaux, France

Introduction



Standard clustering: X and K need each other

- **Dirichlet process mixture models** cluster $X_{N\times d}$ without pre-specifying K
  - governed by Dirichlet process $DP(\alpha, G_0)$ (*Antoniak (1974)*)
- Posterior sampling via MCMC: slow and difficult to converge !
  - ✔ Alternative: fast and deterministic **Variational Inference** (*Blei et al. (2006)*)
- Mean-field VI assumes independence in otherwise dependent parameters
  - **Collapsed VI** integrates them out (*Kurihara et al. (2007)*) but **looses conjugacy** for α prior
- No existing CVI implementation with adaptive α
  - required for **consistent K estimation** (*Ascolani et al. (2023)*)
- G_0 **covariance** choice affects the estimated K (*Hejblum et al. (2019)*) - requires **data-driven prior**



Importance of cluster covariance (Hejblum et al. (2019))

G_0 Prior choice: Covariance parameterisations

Assumption	Structure	Prior
Fixed	$M_1 : \sigma^{-1}\mathbf{I}_{D\times D}$	—
	$M_2 : \Sigma$	—
Unknown (global)	$M_3 : \sigma^{-1}\mathbf{I}_{D\times D}$	$\sigma \sim \Gamma$
	$M_4 : \Sigma$	$\Sigma \sim IW$
	$M_5 : \Sigma^{-1} = LL^\top$	$L \sim N,\ L_{ii}^2 \sim \Gamma$
Unknown (cluster)	$M_6 : \Sigma_k$	$\Sigma_k \sim IW$
	$M_7 : \Sigma_k^{-1}$ (sparse)	Γ diag, Laplace off-diag
	$M_8 : \Sigma_k^{-1}$	Γ diag, Normal off-diag

- Implementation of adaptive Γ prior for α: **Euler-Maclaurin** and **Taylor series** expansions
- **8** different parameterisation for $\boldsymbol{\Sigma_k}$ (or $\boldsymbol{\Sigma_k^{-1}}$)
- Sparsity-inducing M_7: **Sparse DPMM** - optimal model choice (*Jing et al. (2024)*)

Generative algorithm: Sparse DPMM

$$\alpha \sim \Gamma(\alpha|s_1, s_2) \quad \text{DP concentration}$$

$$z_n \sim p_0(\alpha) \quad \text{Kurihara et al. (2007)}$$

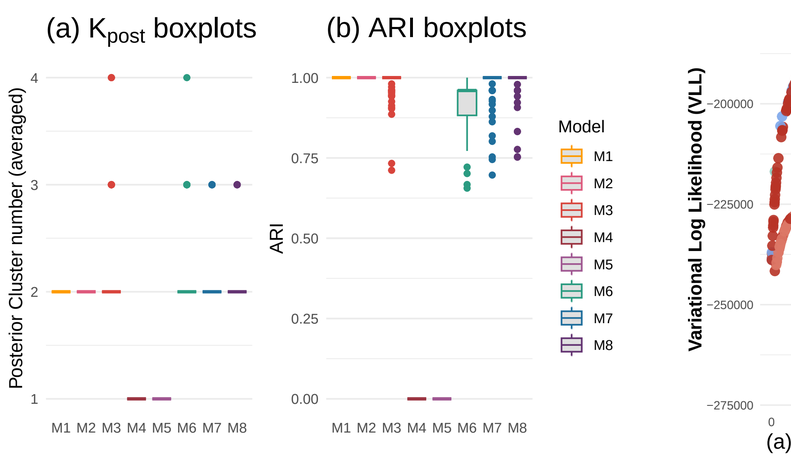
$$\{\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k^{-1}}\} \sim \mathcal{N}(\mu_k|\boldsymbol{\mu_0}, \tfrac{1}{\boldsymbol{\kappa_0}}\Sigma_k) \prod_{i=1}^{D} \Gamma(\Sigma_{k_{ii}}^{-1}|\boldsymbol{a_0}, \boldsymbol{b_0}) \prod_{i<j}^{D} Lap(\Sigma_{k_{ij}}^{-1}|0, c_0)\mathbb{1}_{[\Sigma_k^{-1}\in M^+]} \quad \text{DP base distribution}$$

$$\text{Observations:} \quad x_n|z_n, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k^{-1}} \sim \mathcal{N}(x_n|\boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n}^{-1})$$

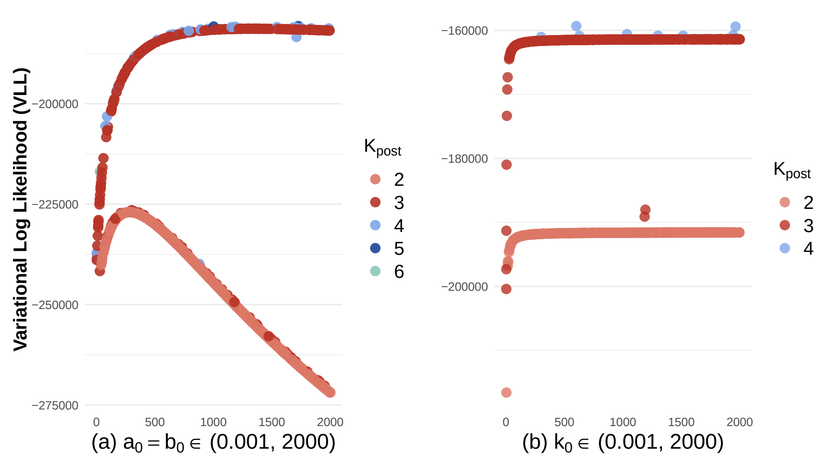
- **Data-driven** hyper-parameter choice: sensitivity analysis of $\boldsymbol{a_0}$, $\boldsymbol{b_0}$ and $\boldsymbol{\kappa_0}$
- **Collapsed VI** inference: **1000×** faster than slice sampler **DPMGibbsN** (*Hejblum et al. (2019)*)
  - scales at $\mathcal{O}(N)$ and $\mathcal{O}(d^2)$
- **Variational Log Likelihood** validation: **BIC** not suitable for BNP models (*Liu and Yang (2011)*)

! **numerical** values of the **hyper-parameters** affect **VLL**

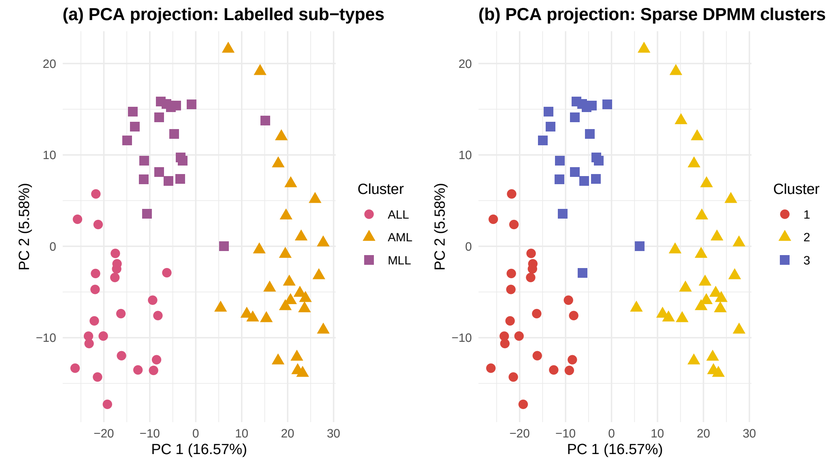
Results



Model comparison with $MVN_{100\times 2}$ & $K_{true} = 2$

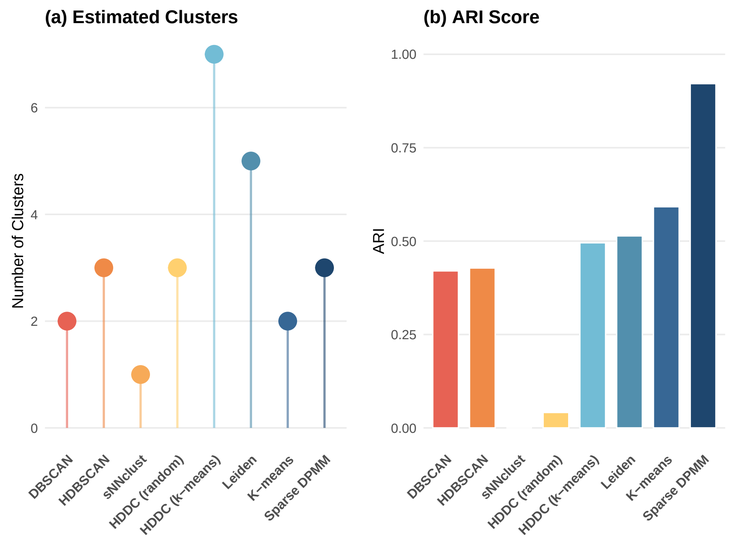


Data-driven hyper-parameter selection

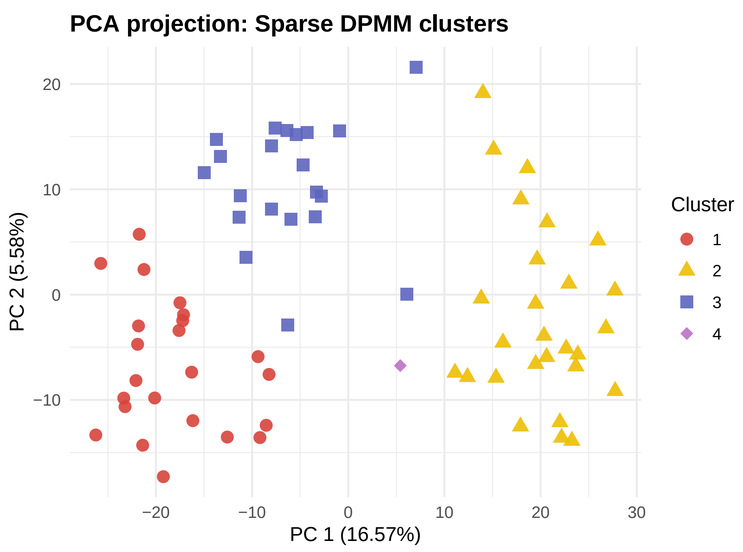


Leukemia cluster recovery: Empirical Bayes hyper-parameter

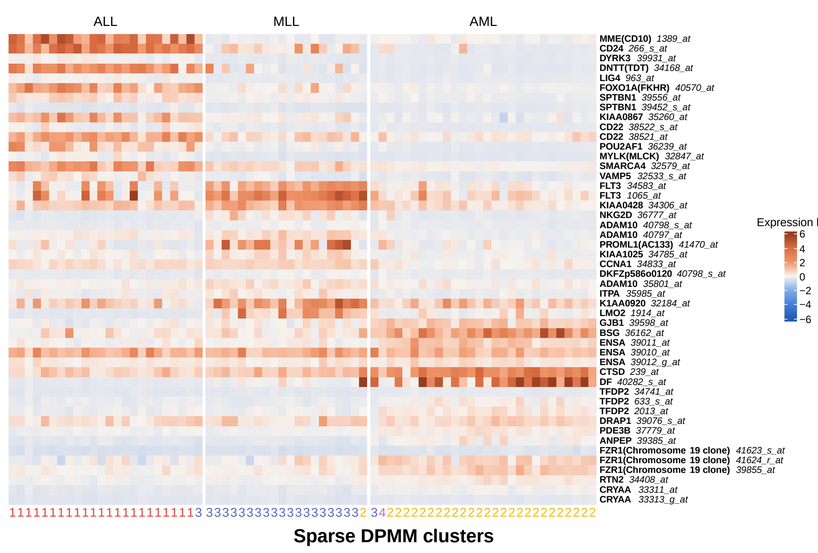
- **Sparse DPMM** - more robust for **varying** $\boldsymbol{K_{true}} \in \{4, 6, 8, 10\}$ across **cluster-specific models**
- Hyper-parameter sensitivity with **Negative Binomial** simulation - $\boldsymbol{a_0 = b_0}$ as **empirical Bayes** estimate while maximising **VLL**; $\boldsymbol{\kappa_0 = N + 1}$ as **effective prior sample** (*Murphy (2007)*)
  - **Mean pairwise** $\mathcal{KL}$ **divergence** of estimated clusters: an **internal validation metric** unaffected by numerical hyper-parameter values ⇒ **degree of overlap** between estimated cluster distributions
- Recovers 3 Leukemia - **ALL**, **MLL**, **AML** - subtypes from microarray data of $N = 72$ patients and $D = 2194$ genes (*Armstrong et al. (2002); De Souto et al. (2008)*)
  - ARI score **0.92** with empirical Bayes hyper-parameters $a_0 = b_0 = 28.91$ (**lowest** mean pairwise $\mathcal{KL}$ score ⇒ **highest** cluster distribution overlap)



Sparse DPMM vs standard techniques



Leukemia subcluster for weaker hyper-parameter



MLL plasticity of the 4^{th} cluster

- Performs **significantly better** than standard clustering techniques like **Leiden**, **HDBSCAN**, **HDDC**, **K-means**, etc. on the Leukemia data
- Weak hyper-parameters $\boldsymbol{a_0 = b_0 = 10}$ retrieve additional 4^{th} cluster with ARI score **0.86** (**highest** mean pairwise $\mathcal{KL}$ score ⇒ **lowest** cluster distribution overlap)
- **Marker genes** validation: Heat-map (*Armstrong et al. (2002)*) shows shared lineage of the outlier sample across **MLL** and **ALL**; whereas **PCA** plot shows **AML** proximity
  - reflects **plasticity** of **MLL** subtype based on fusion genes (*Krivtsov et al. (2007)*) ⇒ importance of **covariance hyper-parameter tuning** as it can reveal **underlying biological complexity**

Sparse DPMM infers $\boldsymbol{K}$ while adaptively updating $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma_k^{-1}}$

R package **vimixr** for clustering **high dimensional** transcriptomic data in **minutes**

This research was conducted in the **SISTM** team, Bordeaux Population Health research centre, INSERM, Inria, University of Bordeaux, France. Funded by the PIA 3 through the Agence Nationale de la Recherche (Ref: 17-EURE-0019) and by the PEPR Santé Numérique project **SMATCH** (Ref: 22-PESN-0003).

Pre-print: arXiv

Pre-print: arxiv.org/abs/2601.21106 or scan the **QR code** on the left. **vimixr:** github.com/annesh07/vimixr or scan the **QR code** on the right.

vimixr R-package