International Biometric Conference IBC

Seoul, 2026

université de BORDEAUX | Inria | Inserm | BORDEAUX POPULATION HEALTH Research Center - U1219 | SISTM | VRI VACCINE RESEARCH INSTITUTE | CHU BDX CENTRE HOSPITALIER UNIVERSITAIRE BORDEAUX

Adaptive DPMM via Collapsed VI for High-Dimensional Transcriptomic Clustering

Annesh Pal, Aguirre Mimoun, Rodolphe Thiébaut, Boris Hejblum

Neither fish nor fowl

Leukemia subtypes can be defined by **gene expression**

- **lineage** impacts **treatment**

Armstrong et al., 2002 [1]

ALL
lymphoid

MLL
mixed lineage

AML
myeloid

Recent biological continuum

MLL plasticity [2]



ALL MLL AML

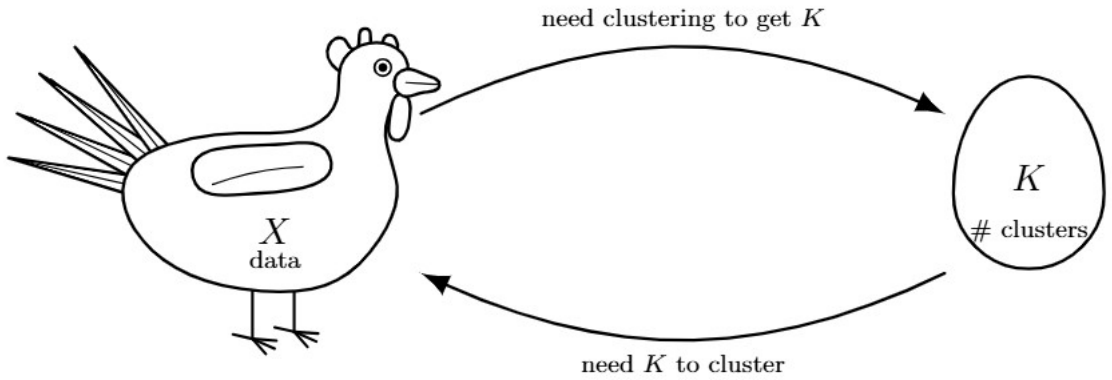
Plasticity between subtypes

- More fine-grained [3] : $K \neq 3$ as previously perceived !
- A **clustering problem** with **unknown** cluster number K

[1] Armstrong et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1):41–47
[2] Chen, C. et al. (2022). Single-cell multiomics reveals increased plasticity, resistant populations, and stem-cell like blasts in kmt2a-rearranged leukemia. Blood, The Journal of theAmerican Society of Hematology, 139(14):2198–2211
[3] Park, Silvia, et al. (2023). Prognostic value of European LeukemiaNet 2022 criteria and genomic clusters using machine learning in older adults with acute myeloid leukemia. *Haematologica*, 109.4:1095.

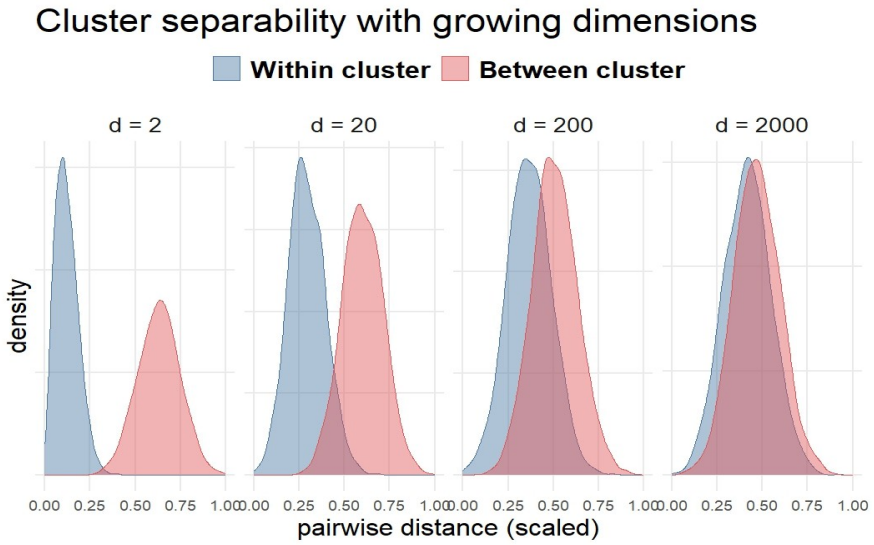
Challenges: Clustering Transcriptomics data

- Pre-specify ***K***



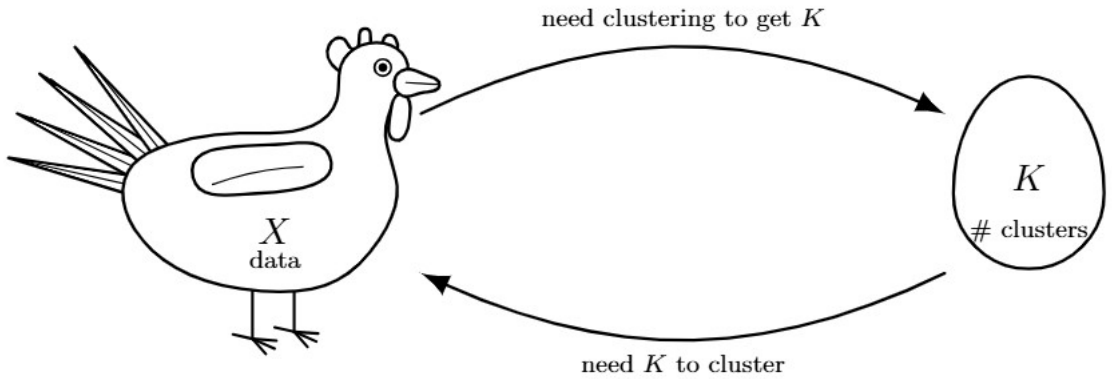
Standard clustering: X and K need each other

- High Dimension



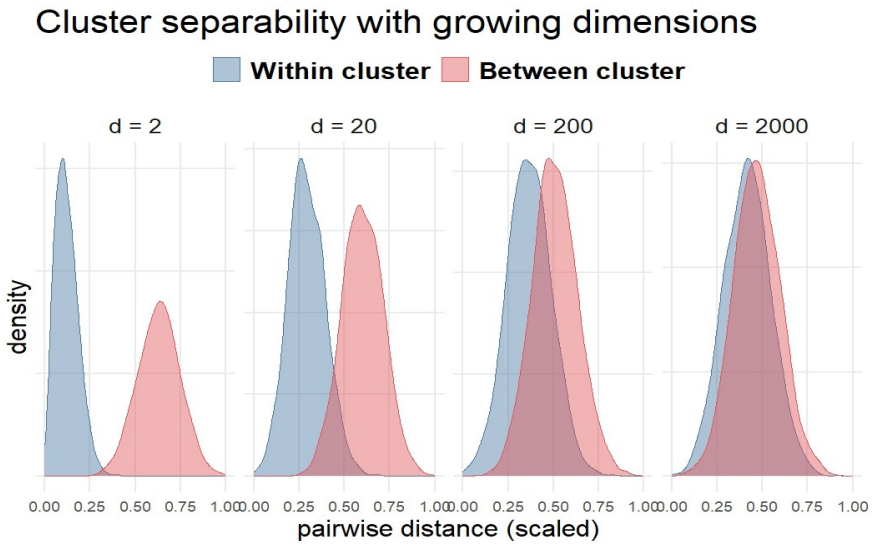
Challenges: Clustering Transcriptomics data

- Pre-specify ***K***



Standard clustering: X and K need each other

- High Dimension

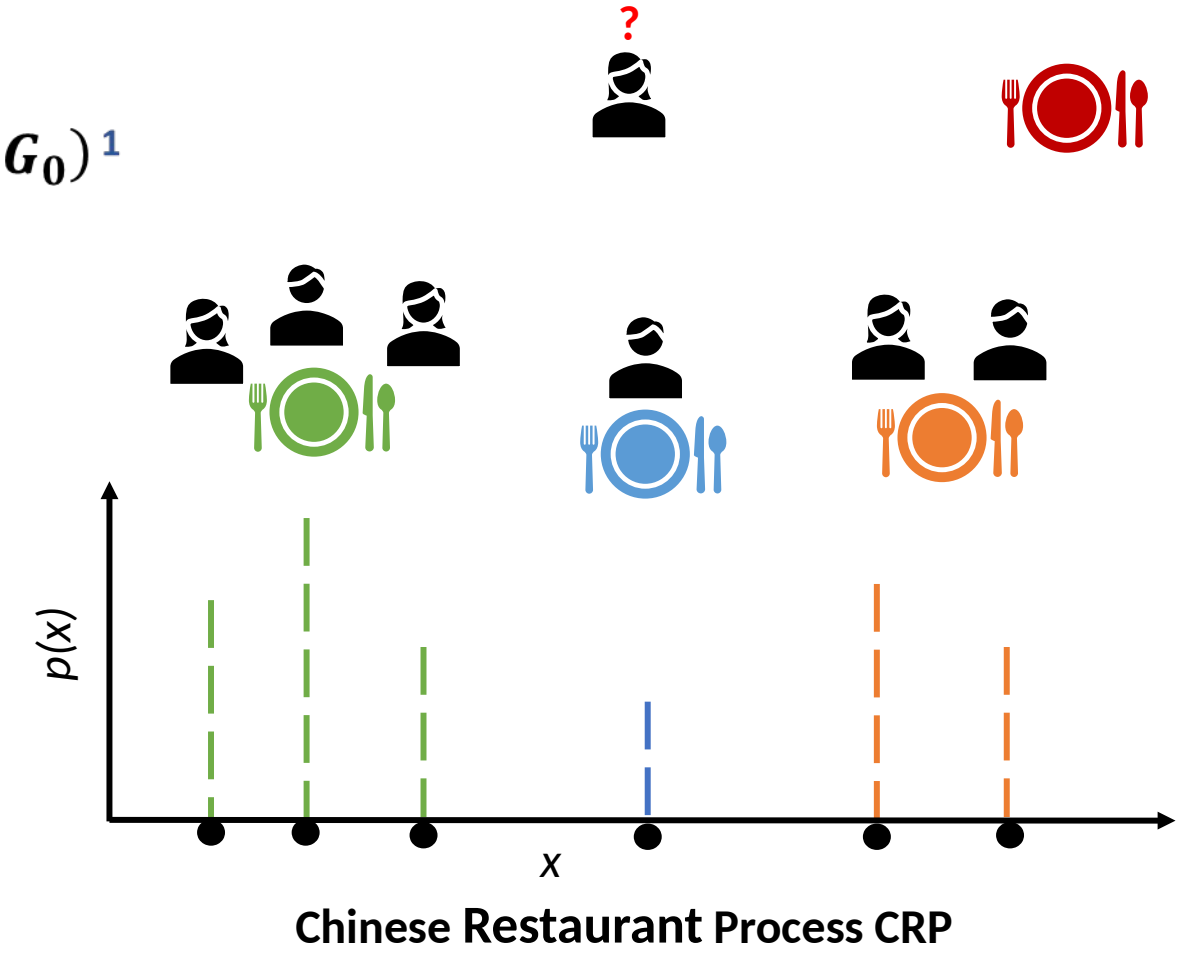


✓ **Dirichlet Process Mixture Models** [1]

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

DPMM: a Chinese Restaurant representation

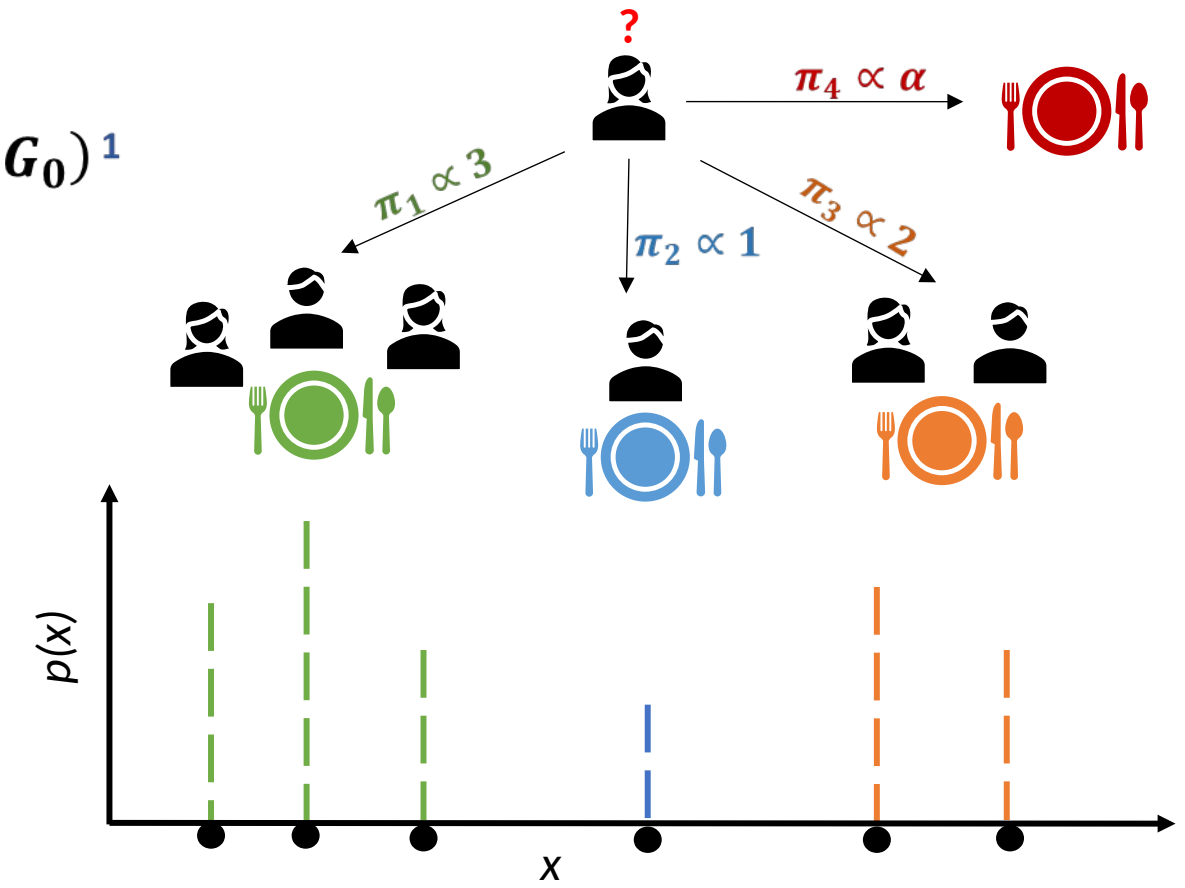
- Clustering governed by **Dirichlet process** $DP(\alpha, G_0)$ [1]



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

DPMM: a Chinese Restaurant representation

- Clustering governed by **Dirichlet process** $DP(\alpha, G_0)$ [1]

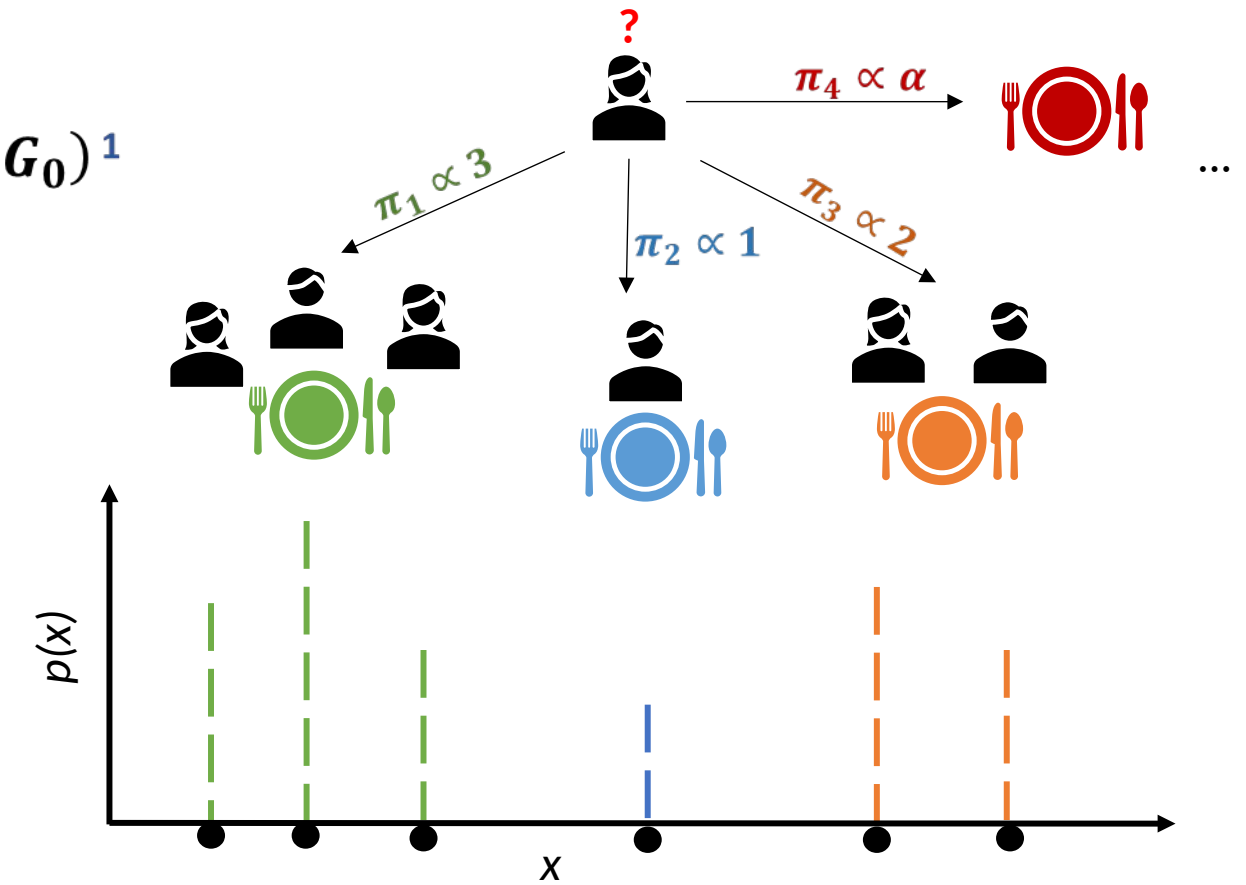


Chinese Restaurant Process CRP

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

DPMM: a Chinese Restaurant representation

- Clustering governed by **Dirichlet process** $DP(\alpha, G_0)$ [1]
- Posterior K directly proportional to α [2]
  - $E[K|N] = \sum_{i=1}^{N} \frac{\alpha}{\alpha+i-1} \sim \alpha \log(1 + \frac{N}{\alpha})$
  - K estimate consistency [3] : update α using $\Gamma(a, b)$ prior *(hierarchical Bayesian paradigm)*



Chinese Restaurant Process CRP

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

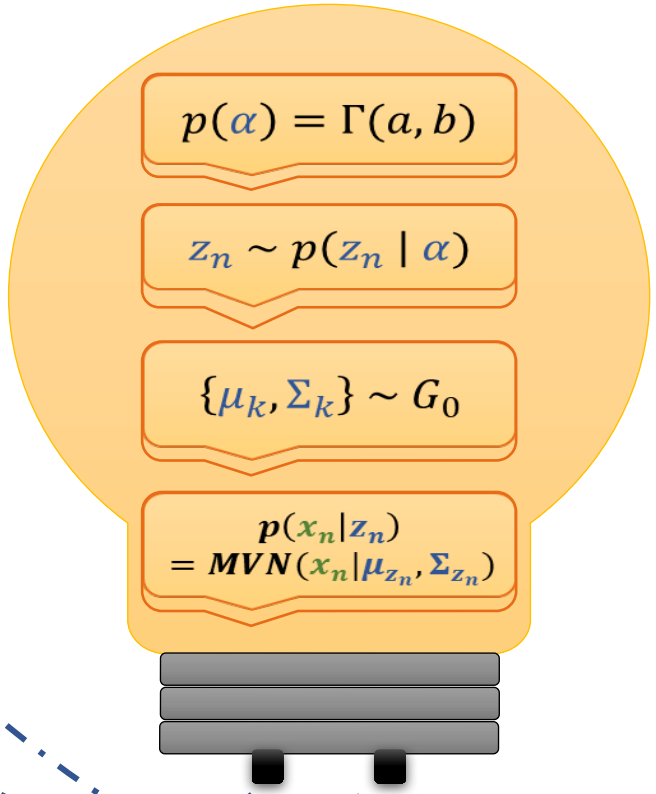
[2] Teh, Y. W. (2010). Dirichlet process. *Encyclopedia of machine learning*, *1063*, 280-287.

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

A new collapsed VI Gaussian mixture model

Assume $\{\boldsymbol{X_n}\}_{n=1}^{N} \in \mathbb{R}^d$ random variables: $p(\boldsymbol{X_n} = \boldsymbol{x_n}) = \prod_k MVN(x_n|\{\mu_k, \Sigma_k\})^{\mathbb{1}_{[z_n=k]}}$

- **DP concentration α**
- **Latent allocation vector z_n**
- **Cluster distribution parameter $\{\mu_k, \Sigma_k\}$**
- **Observed Data $X_n = x_n$**



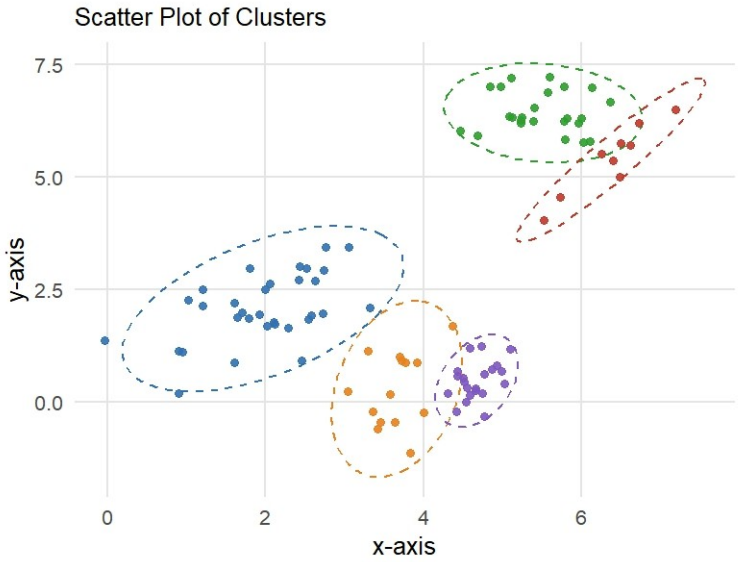
$q^*(\alpha)$: Posterior for overall **#cluster** and effecting $q(z_n)$

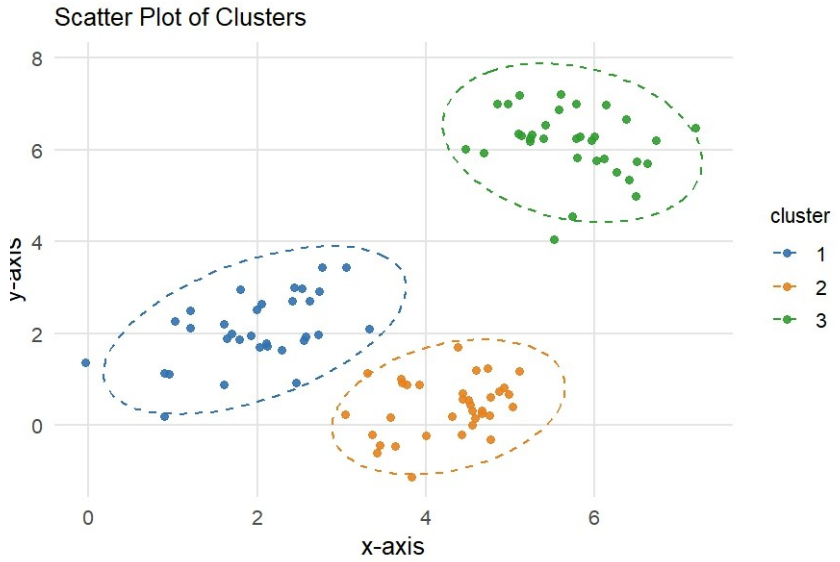
$q^*(\{\mu_k, \Sigma_k\})$: Posterior distributional parameters, defining **cluster distributions**

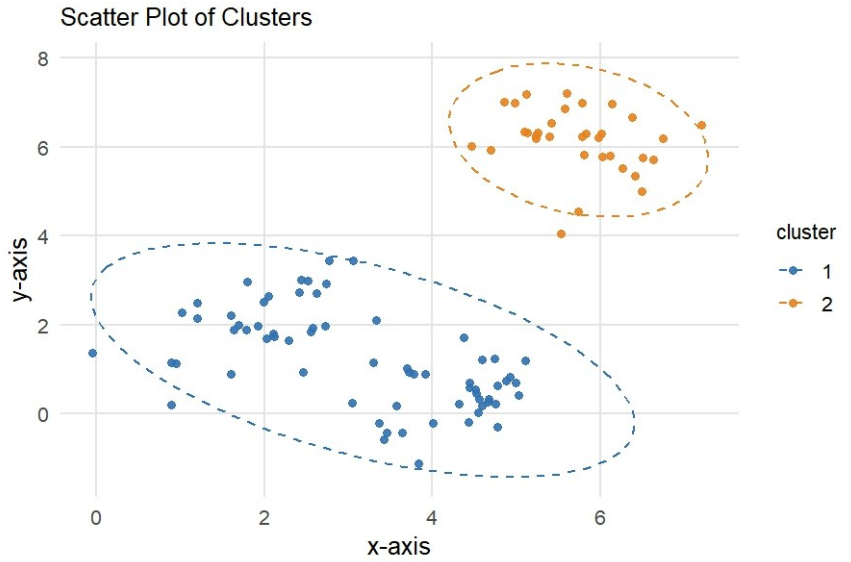
$q^*(z_n)$: Posterior latent allocation vectors, thus estimating **cluster proportions**

Priors matter: **data-driven variance** prior

Effect of G_0 **variance hyper-parameter (hp)** in estimating K [1] – especially for **high d**







Too small hp => **overclustering**

Just right!

Too large hp => **underclustering**

[1] 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

- $q^*(\alpha)$: Non-conjugacy due to $p(z_n|\alpha)$ [1] – **Taylor's series** & **Euler-Maclaurin** approximations
- **8** different parameterisations for G_0 **covariance** - **vimixr** R package

	Assumption	Structure	Prior	
	Fixed	M1: $\Sigma = \frac{1}{\sigma} I_{d\times d}$		
		M2: Σ – full matrix		
	Global	M3: $\Sigma = \frac{1}{\sigma} I_{d\times d}$	$\sigma \sim \Gamma(g_1, g_2)$	
		M4: Σ – full matrix	$\Sigma^{-1} \sim \mathrm{IW}(\nu_0, V_0)$	
		M5: $\Sigma^{-1} = LL^t$	$L_{ii}^2 \sim \Gamma(a_0, b_0)$,	$L_{ij} \sim N(\mu_0, \sigma_0)$
	Cluster-specific	M6: Σ_k – full matrix.	$\Sigma_k^{-1} \sim IW(\nu_0, V_0)$	
		M7: $\boldsymbol{\Sigma_k}$ – Element wise	$\boldsymbol{\Sigma_{k_{ii}}^{-1} \sim \Gamma(a_0, b_0)}$,	$\boldsymbol{\Sigma_{k_{ij}}^{-1} \sim Lap(0, c_0)}$
		M8: Σ_k – Element-wise	$\Sigma_{k_{ii}}^{-1} \sim \Gamma(a_0, b_0)$,	$\Sigma_{k_{ij}}^{-1} \sim N(c_0, 10^{-6})$

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

Model Comparison: Simulation check

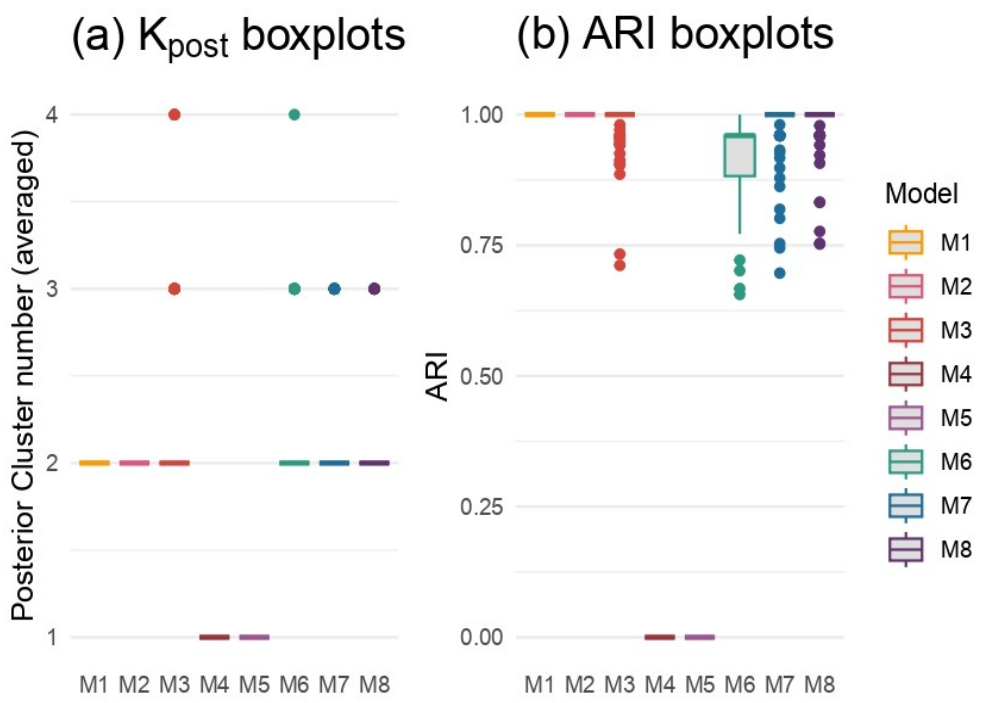


Fig: 100×2 Gaussian samples simulated with $\boldsymbol{K_{true} = 2}$

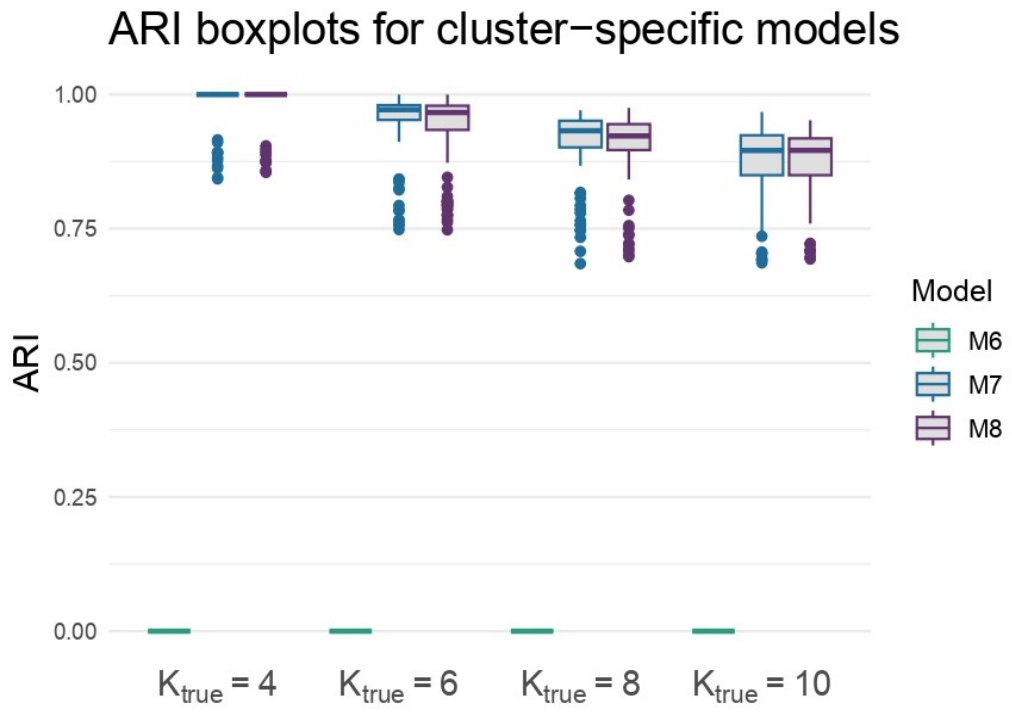


Fig: 100×100 Gaussian samples simulated with **different** $\boldsymbol{K_{true}}$

M7: **Sparse DPMM** performs *1000x* faster than **slice sampler** [1] – scales at $\boldsymbol{O(N)}$ and $\boldsymbol{O(d^2)}$

[1] 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

Model Comparison: Simulation check

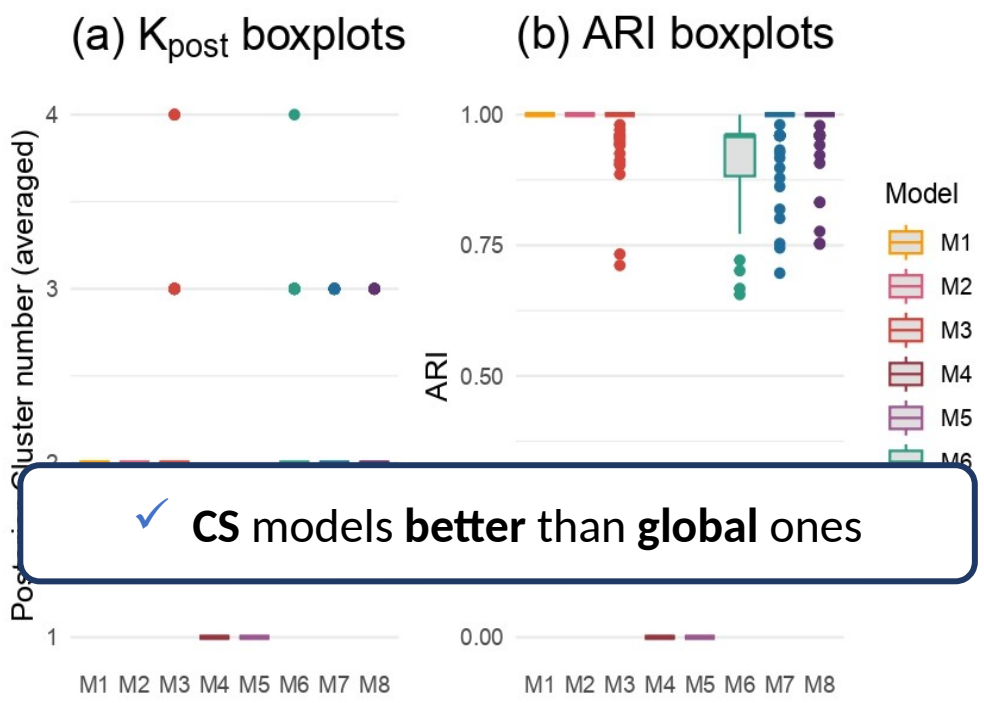


Fig: 100×2 Gaussian samples simulated with $\boldsymbol{K_{true} = 2}$

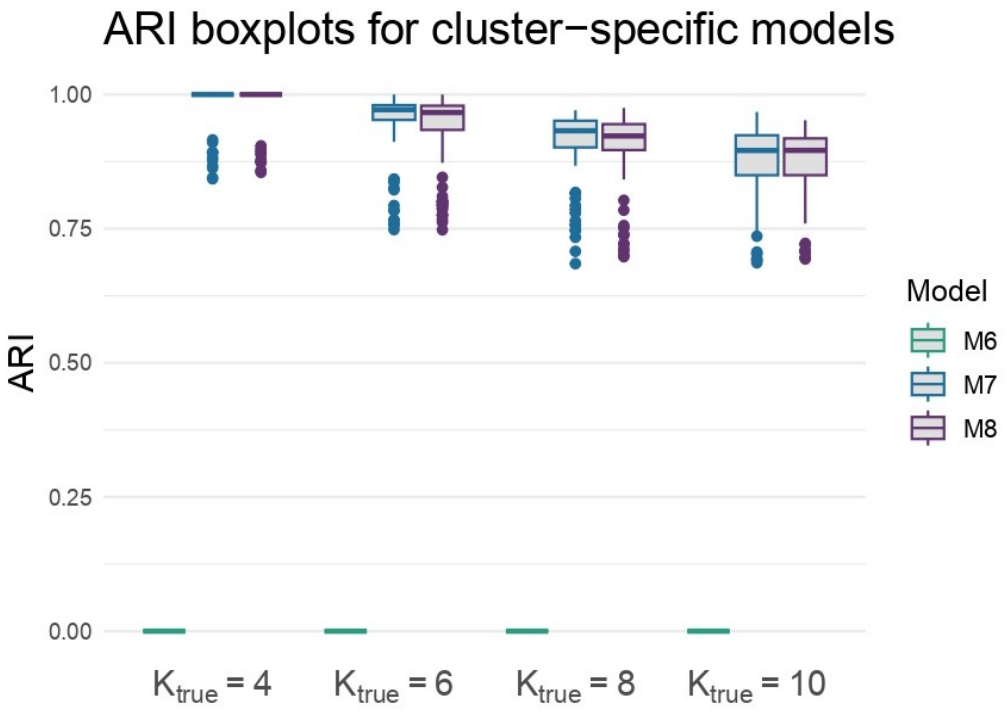


Fig: 100×100 Gaussian samples simulated with **different** $\boldsymbol{K_{true}}$

M7: **Sparse DPMM** performs *1000x* faster than **slice sampler** [1] – scales at $\boldsymbol{O(N)}$ and $\boldsymbol{O(d^2)}$

[1] 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

Model Comparison: Simulation check

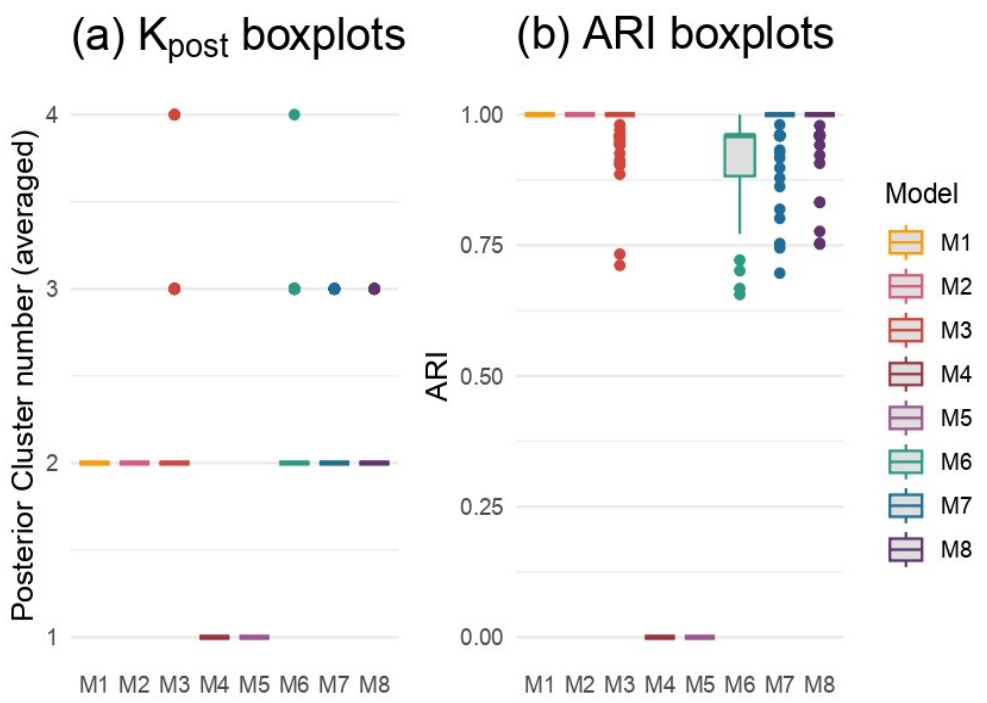


Fig: 100×2 Gaussian samples simulated with $\boldsymbol{K_{true} = 2}$

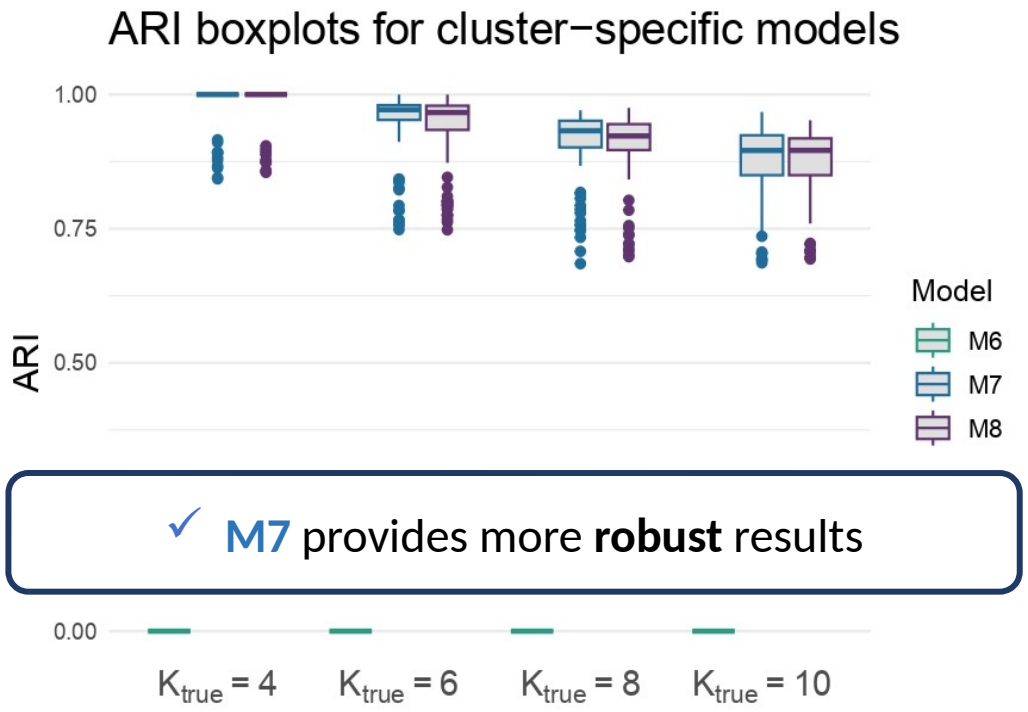


Fig: 100×100 Gaussian samples simulated with **different** $\boldsymbol{K_{true}}$

M7: **Sparse DPMM** performs *1000x* faster than **slice sampler** [1] – scales at $\boldsymbol{O(N)}$ and $\boldsymbol{O(d^2)}$

[1] 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

Model Comparison: Simulation check

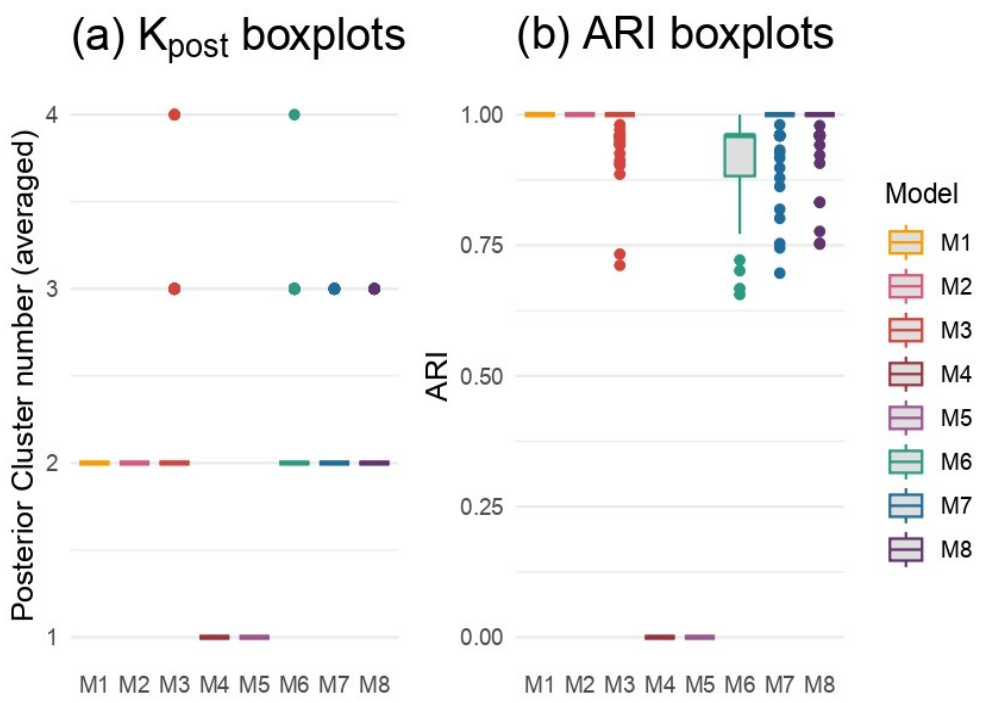


Fig: 100×2 Gaussian samples simulated with $\boldsymbol{K_{true} = 2}$

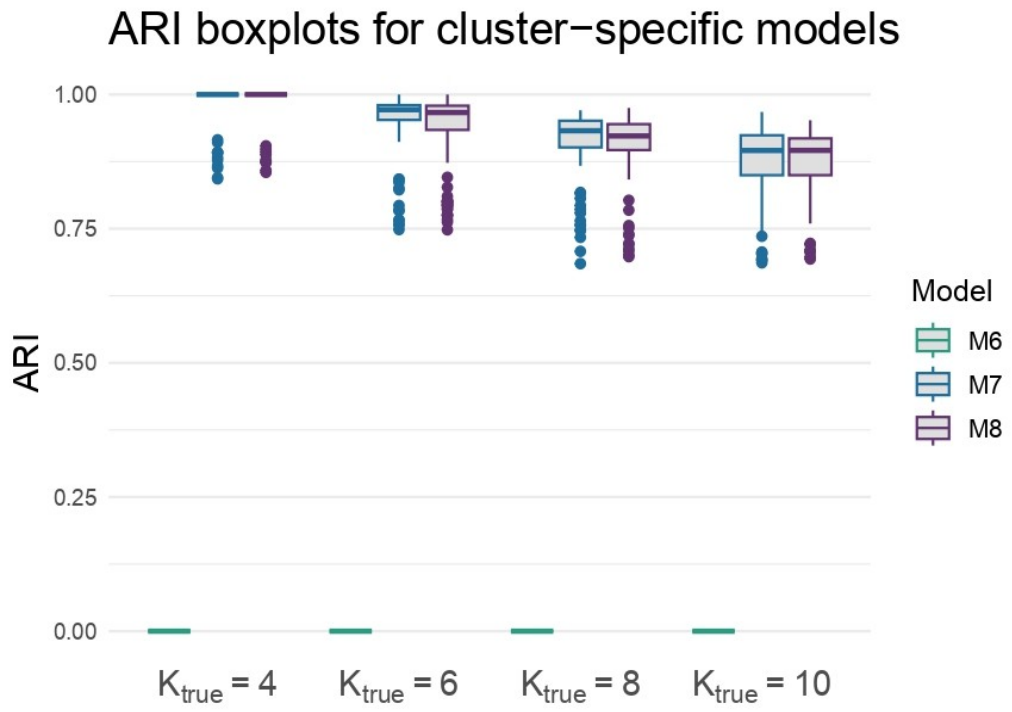


Fig: 100×100 Gaussian samples simulated with **different** $\boldsymbol{K_{true}}$

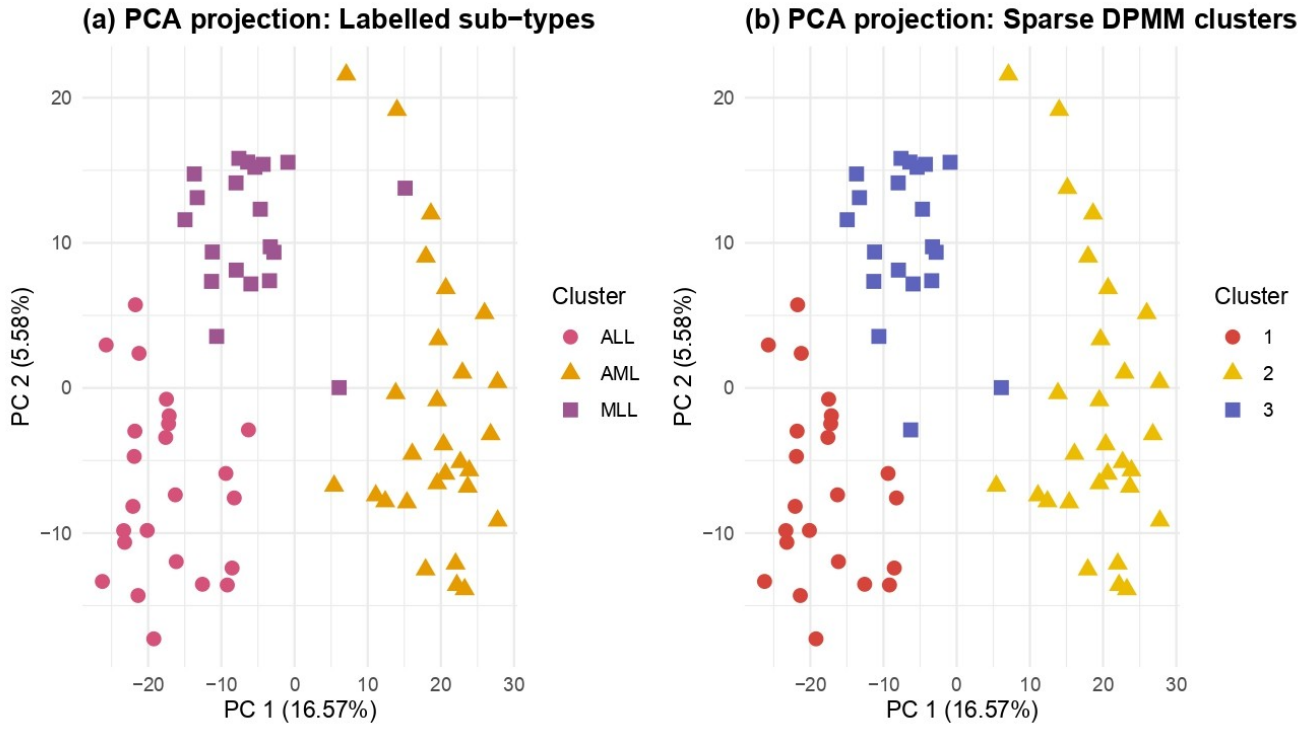
M7: **Sparse DPMM** performs *1000x* faster than **slice sampler** [1] – scales at $\boldsymbol{O(N)}$ and $\boldsymbol{O(d^2)}$

[1] 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

Leukemia Subtypes

- Study on **Leukemia** patients [1]
  - **MLL**: a distinct subtype
  - **marker genes** [1] validation
- **ALL-AML-MLL** data [2] with $d = 2194$ genes for $N = 72$ samples

✓ **Sparse DPMM** (with **Empirical Bayes** hyper-priors): $K_{post} = \mathbf{3}$ with ARI score $\mathbf{0.92}$



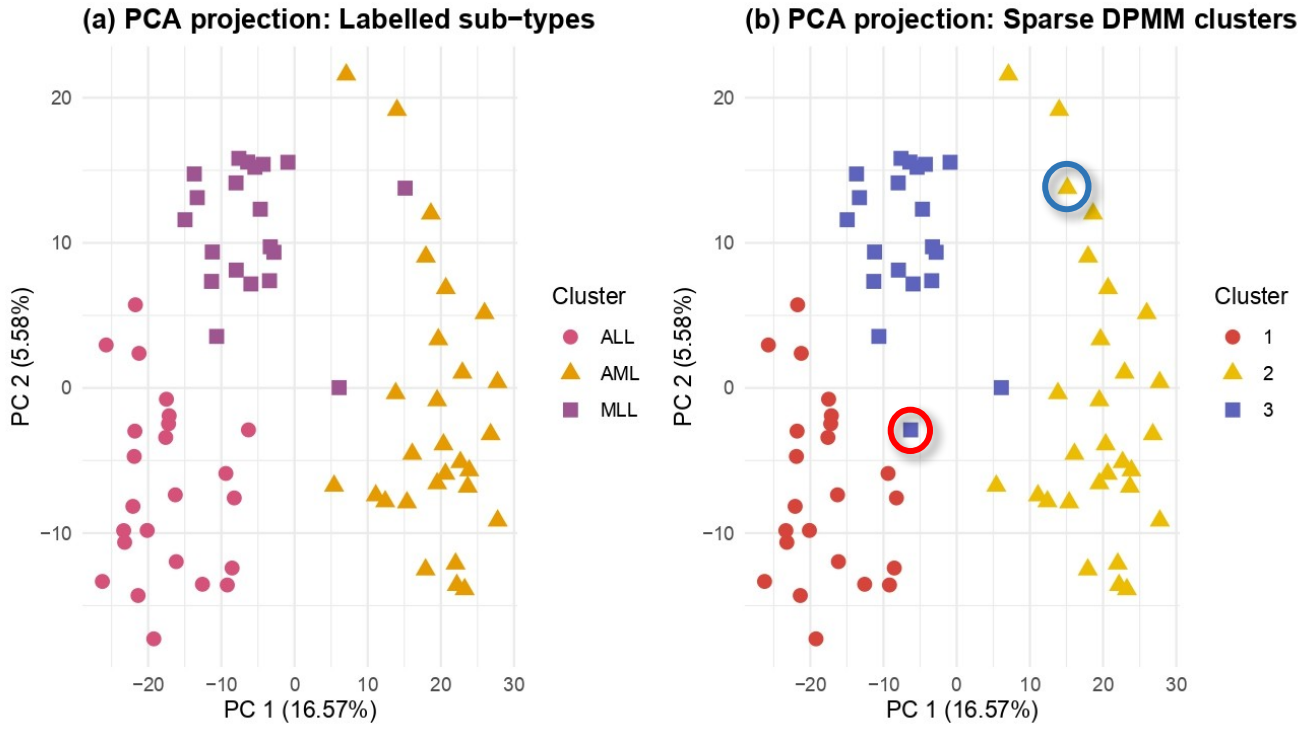
[1] Armstrong et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1):41–47

[2] 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

Leukemia Subtypes

- Study on **Leukemia** patients [1]
  - **MLL**: a distinct subtype
  - **marker genes** [1] validation
- **ALL-AML-MLL** data [2] with $d = 2194$ genes for $N = 72$ samples

✓ **Sparse DPMM** (with **Empirical Bayes** hyper-priors): $K_{post} = \mathbf{3}$ with ARI score $\mathbf{0.92}$



[1] Armstrong et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1):41–47

[2] 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

Leukemia Subtypes - Benchmarking

Popular clustering algorithms:

- **Density**-based: *DBSCAN*, *HDBSCAN*
- **Graph**-based: *sNNclust*, *Leiden*
- **Model**-based: *HDDC*
- **Geometry**-based: *k-means*



Leukemia Subtypes - Benchmarking

Popular clustering algorithms:

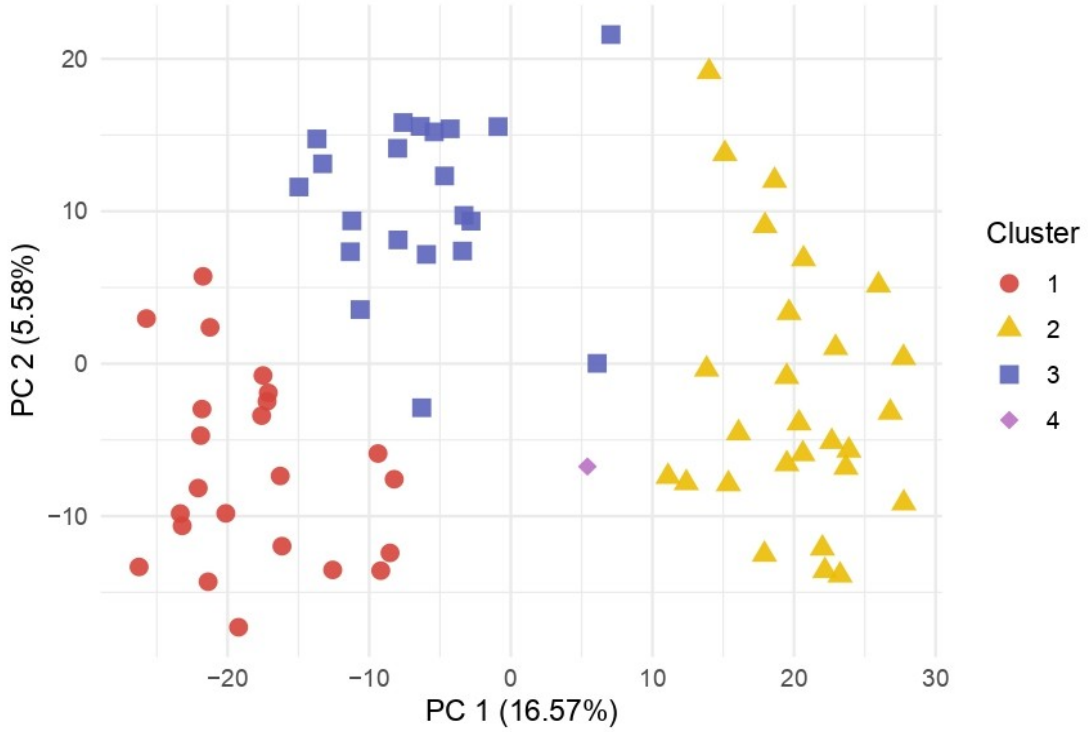
- **Density**-based: *DBSCAN*, *HDBSCAN*
- **Graph**-based: *sNNclust*, *Leiden*
- **Model**-based: *HDDC*
- **Geometry**-based: *k-means*

Sparse DPMM **outperforms** in terms of estimated K & **ARI**

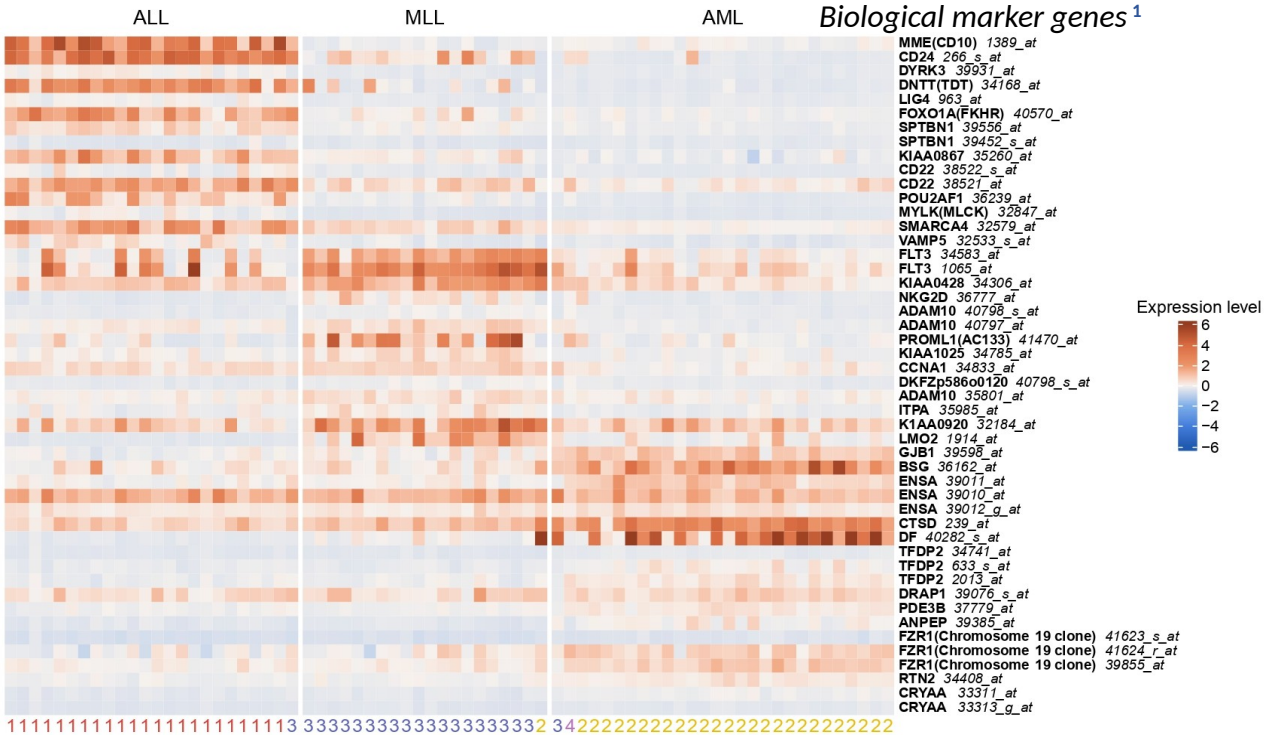


Leukemia Subtypes – 4^{th} cluster

PCA Projection: Estimated Sparse DPMM clusters



Biological marker genes [1]



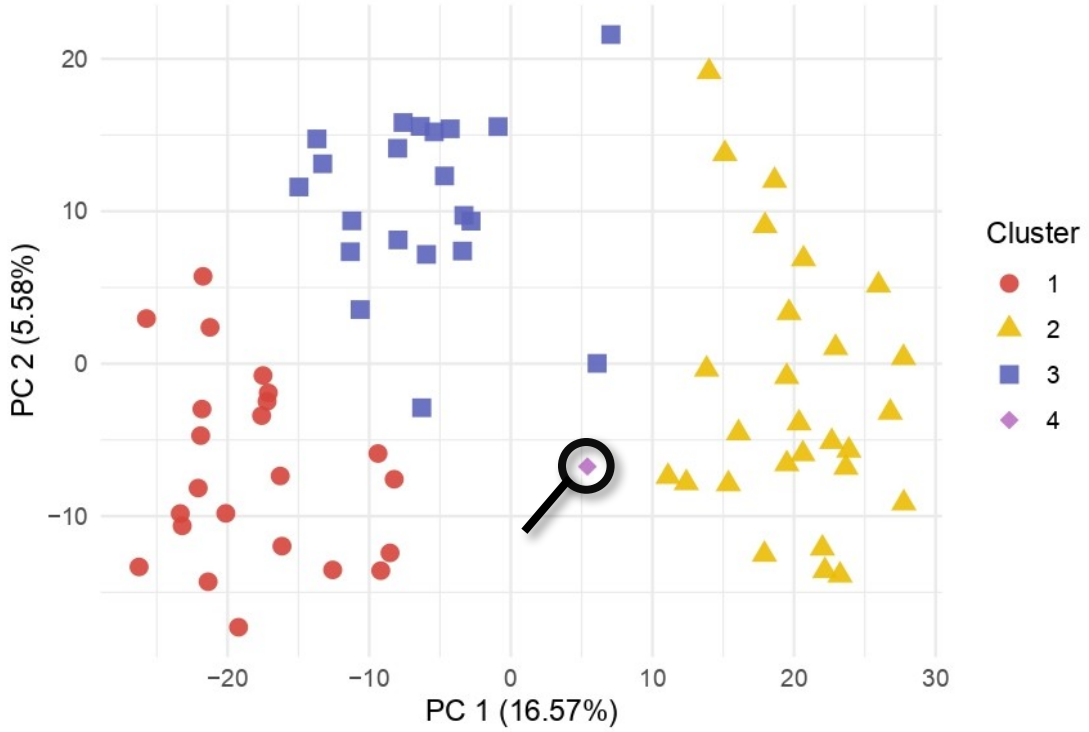
Estimated Sparse DPMM clusters

- ✓ Sparse DPMM (using **weak** hyper-priors) retrieves one **sample** as 4^{th} **cluster (ARI = 0.83)**
- Prominent expression of *PROML1*, KIAA0428, *FLT3* (**MLL marker**) as well as *CD22*, *CD24* (**ALL marker**) genes [1]
- **Numerical proximity** towards **AML** (from PCA plot) vs **Biological proximity** towards **ALL** (marker genes heatmap)

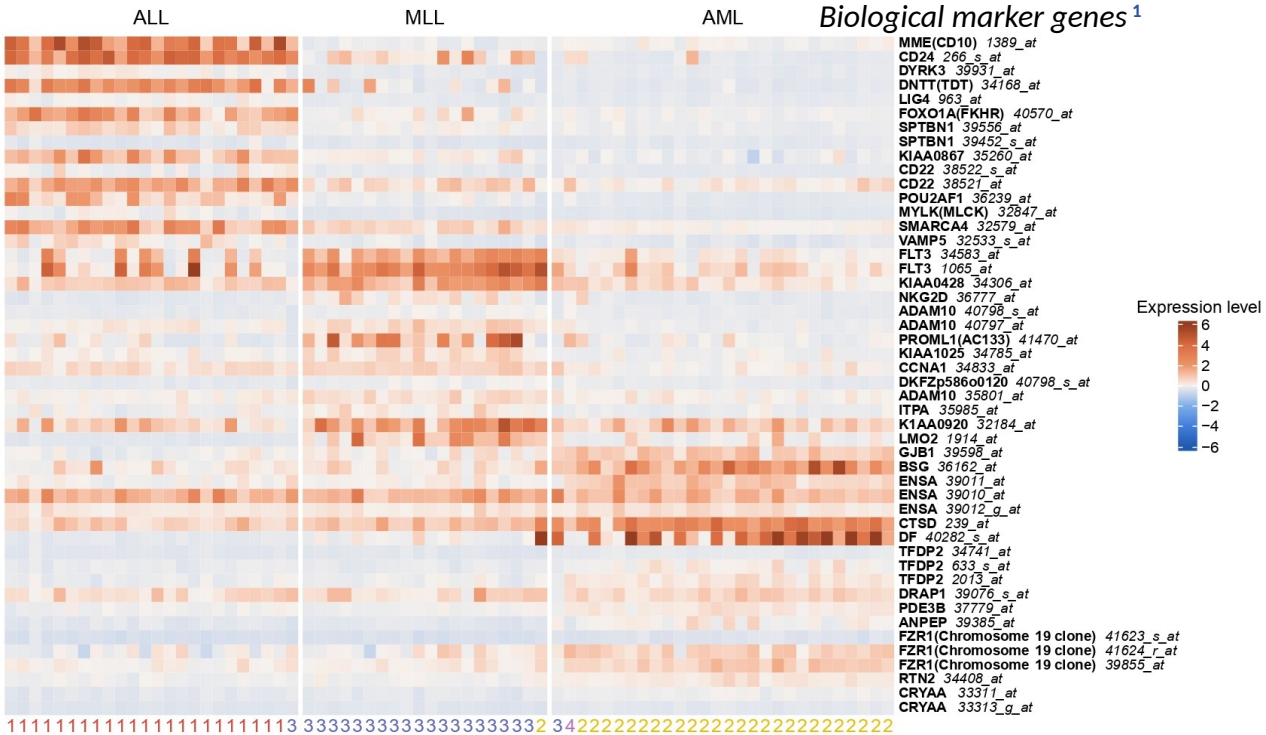
[1] Armstrong et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1):41–47

Leukemia Subtypes – 4^{th} cluster

PCA Projection: Estimated Sparse DPMM clusters



Biological marker genes [1]

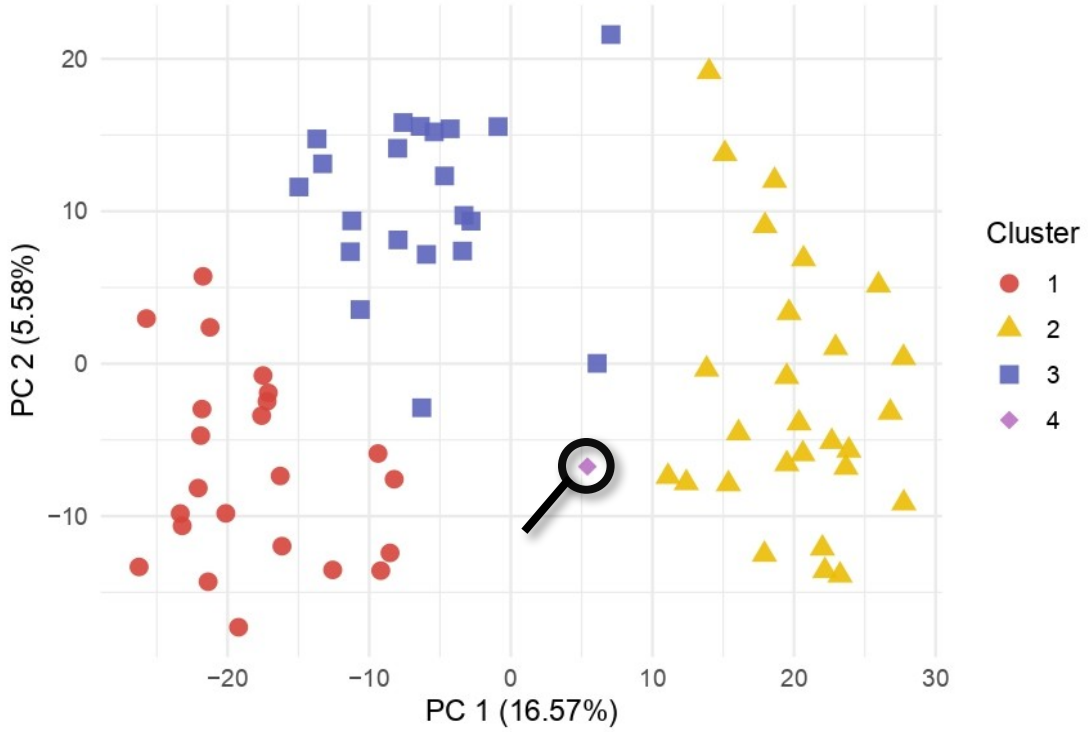


- ✓ Sparse DPMM (using **weak** hyper-priors) retrieves one **sample** as 4^{th} **cluster (ARI = 0.83)**
- Prominent expression of *PROML1*, KIAA0428, *FLT3* (**MLL marker**) as well as *CD22*, *CD24* (**ALL marker**) genes [1]
- **Numerical proximity** towards **AML** (from PCA plot) vs **Biological proximity** towards **ALL** (marker genes heatmap)

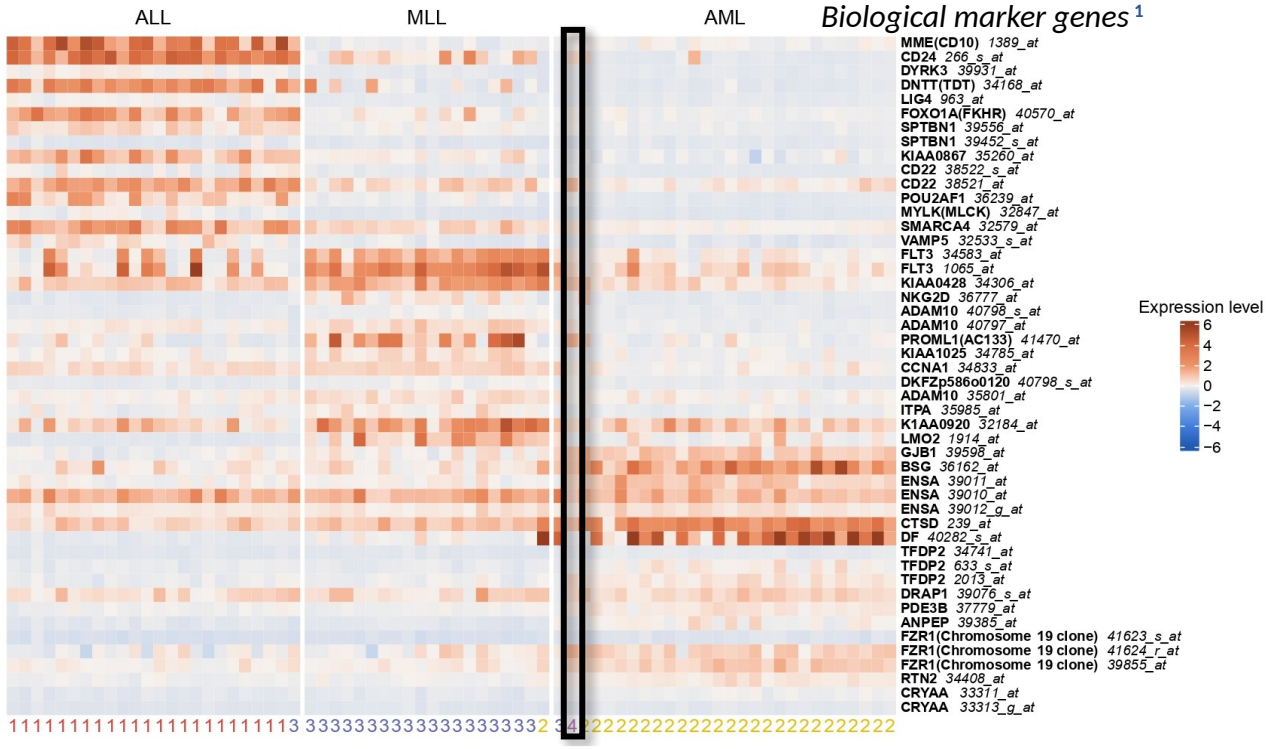
[1] Armstrong et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1):41–47

Leukemia Subtypes – 4^{th} cluster

PCA Projection: Estimated Sparse DPMM clusters



Biological marker genes [1]



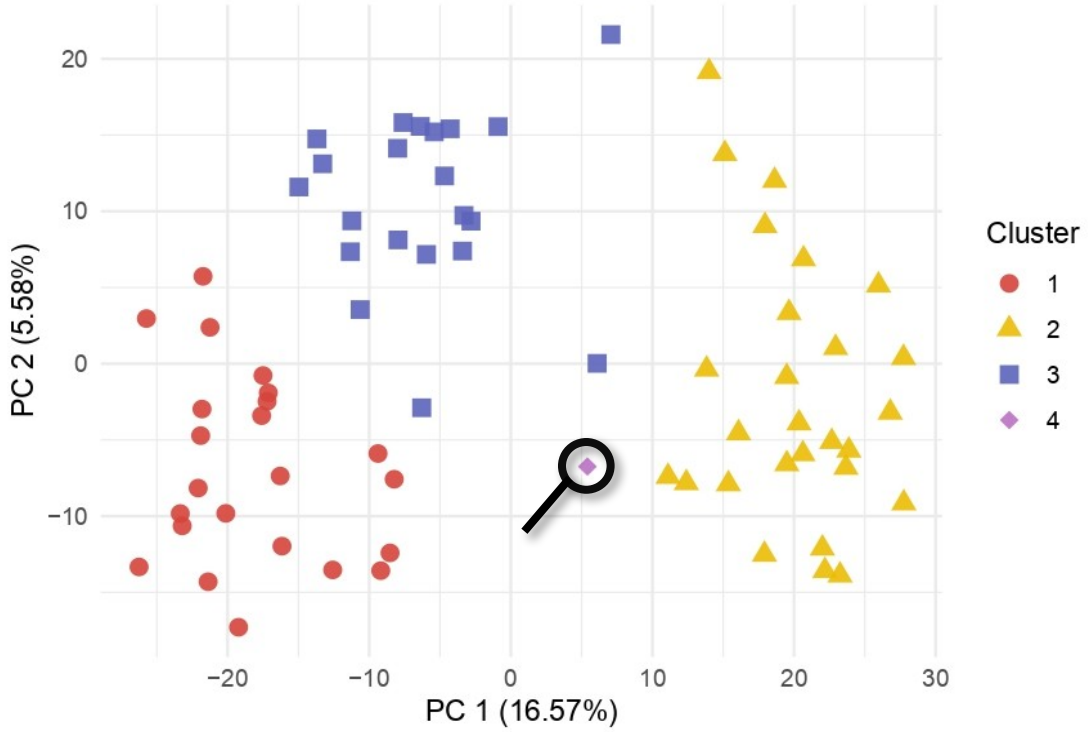
Estimated Sparse DPMM clusters

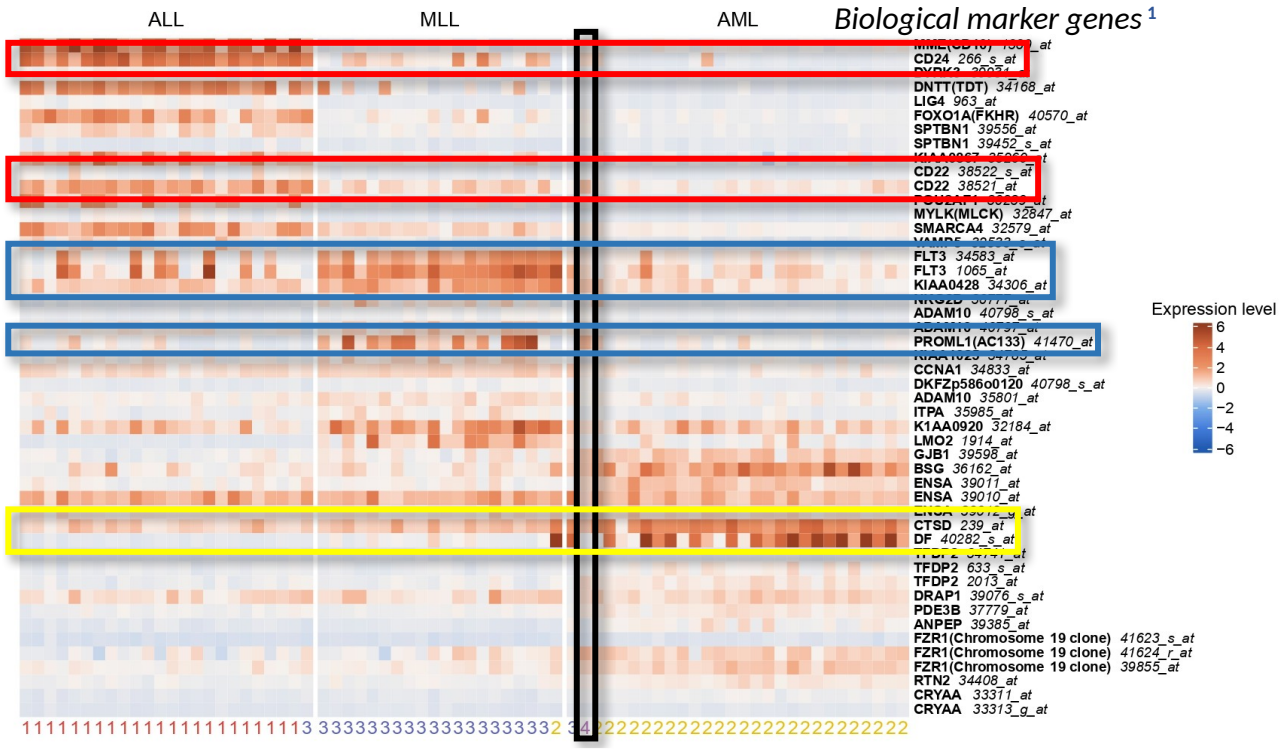
- ✓ Sparse DPMM (using **weak** hyper-priors) retrieves one **sample** as 4^{th} **cluster (ARI = 0.83)**
- Prominent expression of *PROML1*, KIAA0428, *FLT3* (**MLL marker**) as well as *CD22*, *CD24* (**ALL marker**) genes [1]
- **Numerical proximity** towards **AML** (from PCA plot) vs **Biological proximity** towards **ALL** (marker genes heatmap)

[1] Armstrong et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1):41–47

Leukemia Subtypes – 4^{th} cluster

PCA Projection: Estimated Sparse DPMM clusters





- ✓ Sparse DPMM (using **weak** hyper-priors) retrieves one **sample** as 4^{th} **cluster (ARI = 0.83)**
- Prominent expression of *PROML1*, KIAA0428, *FLT3* (**MLL marker**) as well as *CD22*, *CD24* (**ALL marker**) genes [1]
- **Numerical proximity** towards **AML** (from PCA plot) vs **Biological proximity** towards **ALL** (marker genes heatmap)

[1] Armstrong et al. (2002). Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 30(1):41–47

Conclusion

Model

A novel **collapsed VI** for **DPMM** with

- ✓ A weakly-informative $\boldsymbol{\Gamma(a, b)}$ **prior** for $\boldsymbol{\alpha}$
- ✓ Different **parameterisations** of $\boldsymbol{G_0}$ **variance**

Application: *More than meets the eye*

- ❖ **MLL** display **lineage plasticity** – coherent behaviour with recent studies
- ❖ $\boldsymbol{K = 3}$ is **not enough** for clustering
  - more **data-driven approach** to reveal hidden complexity
- ❖ Necessary as **treatment** depends on the lymphoid-myeloid **lineage**

Thank You !

- *This research was supported by the **Université de Bordeaux**'s **Digital Public Health Graduate School**, financed by the PIA 3 through the **Agence Nationale de la Recherche** (Ref: 17-EURE-0019), by the **Programme et Equipement Prioritaire de Recheche Santé Numérique** (PEPR SN) project **SMATCH** (Ref: 22-PESN-0003).*
- *Computational resource was provided by the computing facilities **Mésocentre de Calcul Intensif Aquitain** (MCIA) of the **Université de Bordeaux** and the **Université de Pau et des Pays de l'Adour**.*

smatch

FRANCE 2030

Digital Public Health Graduate Program

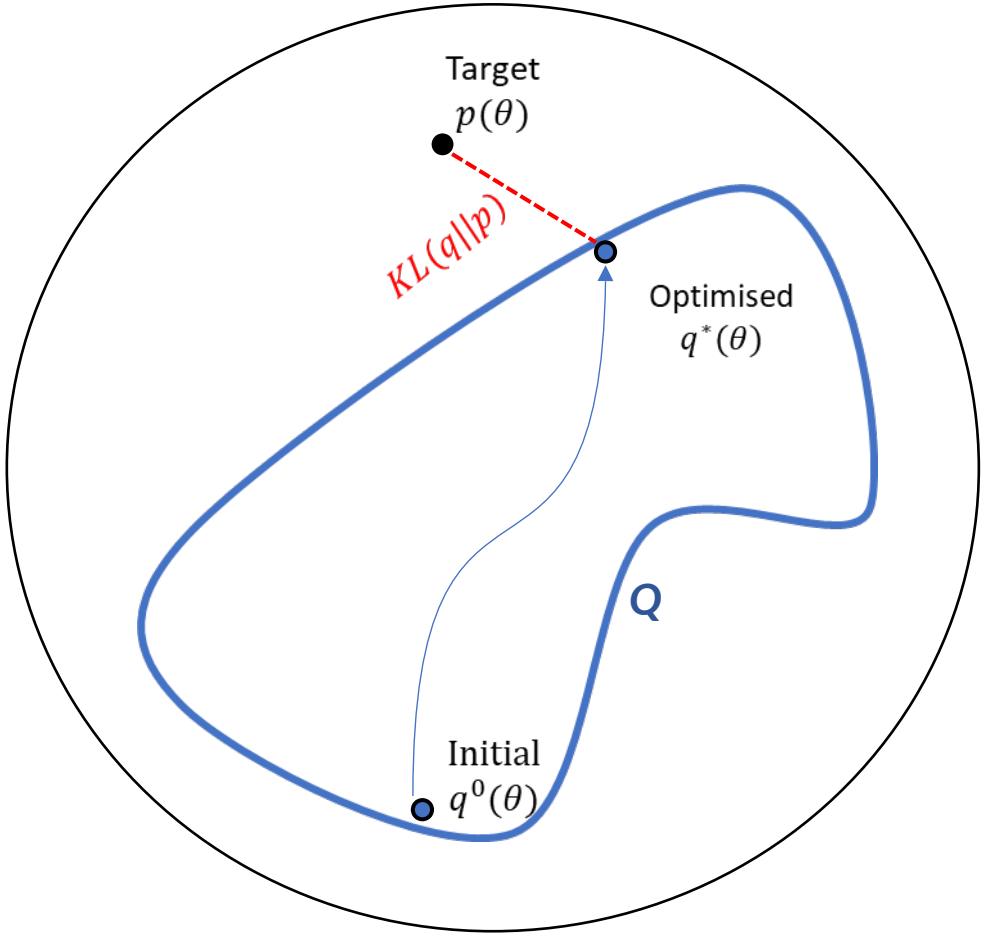
université de BORDEAUX

Supplementary: Variational Inference

- Model complexity leads to analytically intractable posterior $\boldsymbol{p(\theta)}$ – approximation methods
- **MCMC** as a popular choice

  ! Slow and difficult convergence – especially for **high dimensions**
- An alternative – **Variational Inference** (VI) [1]
- Mean-field VI [1] – independence between variational parameters

  ! Too strong for **DPMM**

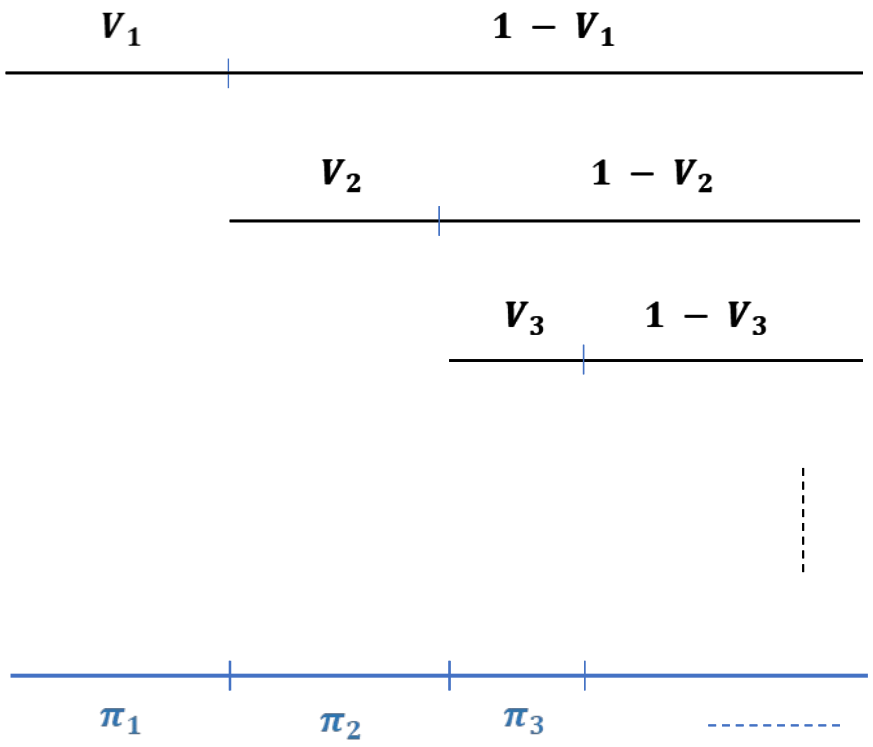


Supplementary: Collapsed VI

- **Stick-breaking** representation of DP [1]
- Mean-field VI – independence between variational parameters

  ! Too strong for **DPMM** – especially between π_k's and $z_n \sim Cat(\{\pi_k\})$

- **Collapsed VI**: integrates π_k's out
  - $p(z_n|\alpha) = \prod_k \frac{\Gamma(\mathbb{1}_{[z_n=k]}+1)\Gamma(\mathbb{1}_{[z_n>k]}+\alpha)}{\Gamma(\mathbb{1}_{[z_n\geq k]}+1+\alpha)}$ where $\mathbb{1}_{[.]}$ is the indicator function [2]



✓ Update of α depends on **#non-zero clusters**, rather than **#total variational cluster**

Supplementary: Scalability of Sparse DPMM

- Speed of convergence depends both on sample N and dimension d

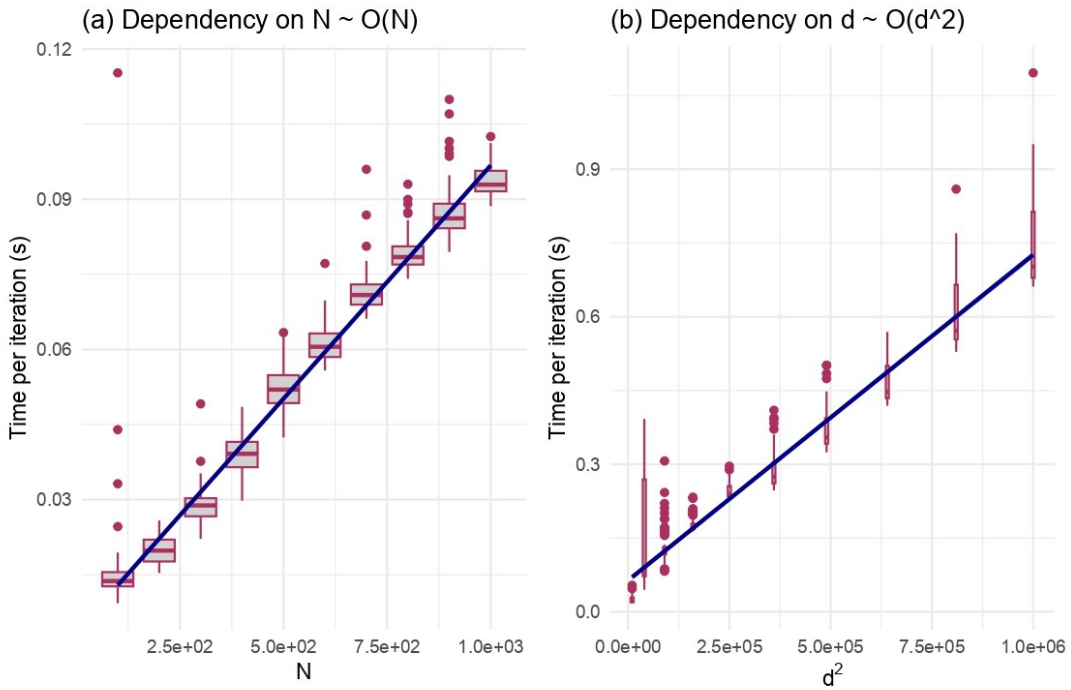


Fig: Gaussian samples simulated with $K_{true} = 2$ (time in seconds)

- Comparison with a slice sampler [1]
  - **NPflow** R package: 1000 MCMC iterations

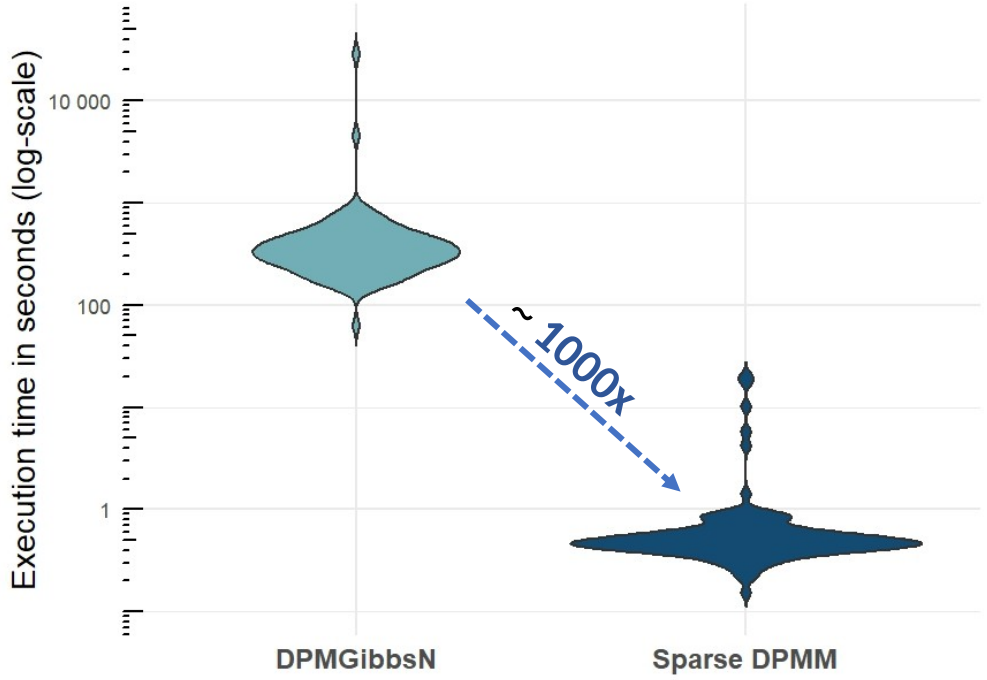


Fig: 100×100 Gaussian samples with $K_{true} = 2$ (comparison in logarithmic scale)

Supplementary: Sparse DPMM hyper-priors

Internal Validation Metric:

- mean pairwise **Kullbach-Liebler** (*KL*) **divergence** between two **cluster distributions**
- $KL_{mean} = \frac{1}{K_{post}} \Sigma_{i \neq j \in K_{post}} \left[\frac{KL(\theta_i, \theta_j) + KL(\theta_j, \theta_i)}{2}\right]$

$a_0 = b_0$	K_{post}	KL_{mean}	ELBO	VLL	Silhouette Coeff.
0.01	1	--	2.8×10^9	-6.6×10^5	--
0.1	1	--	2.8×10^9	-6.6×10^5	--
1	14	104.25	1.3×10^9	-3.2×10^5	-0.014
10	4	116.27 (*highest*)	2.2×10^9	-3.6×10^5	0.067
28.9 (*Emp. Bayes*)	3	53.4	2.2×10^9	-3.4×10^5	0.081
50	3	43.27 (*lowest*)	2.2×10^9	-3.4×10^5	0.081
100	2	69.94	2.6×10^9	-4.2×10^5	0.126
500	2	59.74	2.6×10^9	-4.1×10^5	0.126
1000	2	58.83	2.6×10^9	-4.1×10^5	0.126
10000	2	58.16	2.6×10^9	-4.1×10^5	0.126

Supplementary: Choosing K for s.o.t.a techniques

- **Diversity** in the considered approaches: ran every method from $K = \mathbf{2}$ to $K = \mathbf{10}$
  - used **method-specific internal validation metrices** to select **optimal** $K = \boldsymbol{K_{opt}}$ for input
- Compared the ARI scores for $\boldsymbol{K = 3}$: **Sparse DPMM** still performs **better** (ARI score **0.92**)

Clustering methods	Internal validation metric	Optimal K_{opt}	ARI (for K_{opt})	ARI (for $K = 3$)
DBSCAN	Silhouette coeff.	10	0.42	0.02
HDBSCAN	Silhouette coeff.	5	0.428	0.645
SNNCLUST	Silhouette coeff.	9	0	0.194
HDDC (random init.)	BIC	3	0.041	0.279
HDDC (k-means init.)	BIC	7	0.495	0.88
Leiden	Graph modularity	2	0.514	0.545
K-Means	Silhouette coeff.	2	0.592	0.442