Consensus Monte Carlo for mixtures of categorical distributions

Application to identify patterns of multi-morbidity from EHRs

Julie Fendler

MRC Biostatistics Unit, University of Cambridge

19th June 2026

Federated Learning - definition

Federated learning is a setting for machine learning in which data are split across multiple computers (local nodes). Model inference is coordinated by a central server (global node). None of the local nodes' data may be directly shared, exposed, or transferred to other local nodes or to the global node. [1]

Federated Learning - definition

Federated learning is a setting for machine learning in which data are split across multiple computers (local nodes). Model inference is coordinated by a central server (global node). None of the local nodes' data may be directly shared, exposed, or transferred to other local nodes or to the global node. [1]

Federated learning settings are divided into three categories:

- *data centre distributed learning:* artificially splits a single large dataset across nodes for computational reasons (distributed learning);
- *cross-device federated learning:* each shard contains only one observation of the full data (ex: personal mobile devices);
- *cross-silo federated learning*: the full dataset is naturally split into shards (ex : a geographic pattern).

Cross-silo federated learning for health data

In medical applications, data may be naturally siloed within the medical centres at which they are collected.

In medical applications, data may be naturally siloed within the medical centres at which they are collected.

Goals:

- To identify **clusters of patients** with similar characteristics;
- To **avoid individual data sharing**.

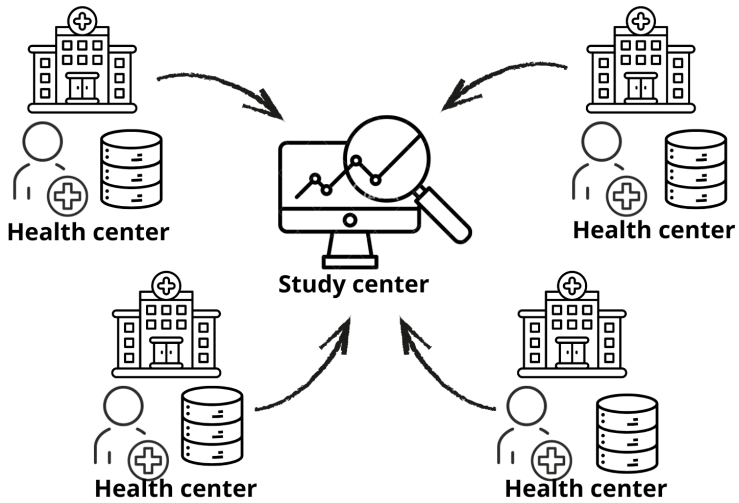


Figure 1: Diagram of a multi-centric study

In medical applications, data may be naturally siloed within the medical centres at which they are collected.

Goals:

- To identify **clusters of patients** with similar characteristics;
- To **avoid individual data sharing**.

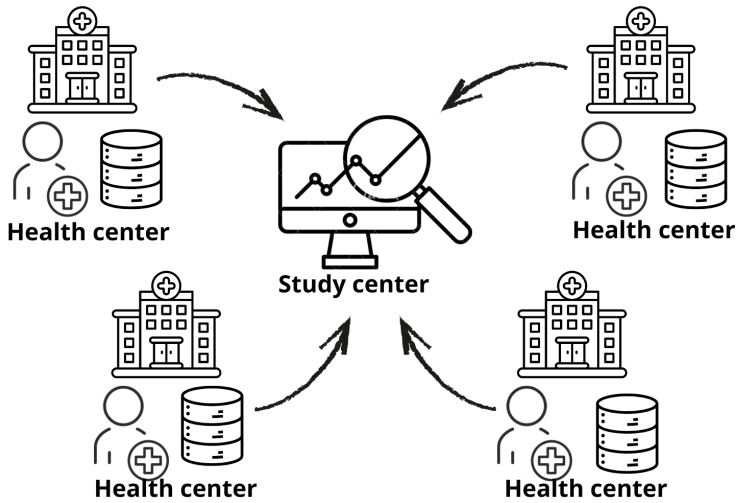


Figure 1: Diagram of a multi-centric study

Only few studies focus on the development of clustering algorithms for cross-silo federated learning [2].

The observed data : X_s^i for individual i in each shard s.

The observed data : X_s^i for individual i in each shard s.

We assume a **clustering structure with C clusters** :

$$\forall i, \forall s, \forall c, X_s^i | C_s^i = c \sim f(X_s^i; \phi^c)$$

with C_s^i the cluster label of the individual i in shard s.

The observed data : X_s^i for individual i in each shard s.

We assume a **clustering structure with C clusters** :

$$\forall i, \forall s, \forall c, X_s^i | C_s^i = c \sim f(X_s^i; \phi^c)$$

with C_s^i the cluster label of the individual i in shard s.

The likelihood of a **mixture of** C^+ **distributions** is defined by the following equation :

$$\ell(.|\theta) : x \mapsto \sum_{c=1}^{C^+} \pi^c f(x; \phi^c).$$

with $C^+ \geq C$ and $\forall c \in [\![1, C^+]\!], \pi_c \geq 0, \sum_{c=1}^{C^+} \pi_c = 1$.

Let us define $\dot{\boldsymbol{X}}_s$ the observed data on the shard s, $s \in \{1, ..., S\}$. We write the likelihood associated with the data on shard s as $\ell(\dot{\boldsymbol{X}}_s|\theta)$.

Let us define $\dot{\boldsymbol{X}}_s$ the observed data on the shard s, $s \in \{1, ..., S\}$. We write the likelihood associated with the data on shard s as $\ell(\dot{\boldsymbol{X}}_s|\theta)$.

Assuming the data are independent and identically distributed, the **global posterior density** $p(\theta|\dot{\boldsymbol{X}}_{1:S})$ can be written

$$\begin{aligned} p(\theta|\dot{\boldsymbol{X}}_{1:S}) &\propto p(\theta) \prod_{s=1}^{S} \ell(\dot{\boldsymbol{X}}_s|\theta) \\ &= \prod_{s=1}^{S} \ell(\dot{\boldsymbol{X}}_s|\theta) p(\theta)^{1/S} \\ &\propto \prod_{s=1}^{S} p_s(\theta|\dot{\boldsymbol{X}}_s) \end{aligned}$$

with $p(\theta)$ the prior density on θ, $p_s(\theta|\dot{\boldsymbol{X}}_s)$ the **local posterior density** on the shard s and $\dot{\boldsymbol{X}}_{1:S}$ the full data.

A Consensus Monte Carlo approach is performed in two steps :

- **Apply step:** Bayesian sampling-based method to sample from each of the local posterior distributions.
- **Aggregate step:** Combine the local posterior distributions to recover the global posterior distribution by assuming :

$$p(\theta|\dot{\boldsymbol{X}}_{1:S}) \approx \prod_{s=1}^{S} p_s(\theta|\dot{\boldsymbol{X}}_s)\delta_{\theta=F(\theta_1,\dots\theta_S)}$$

with F a deterministic **aggregation function**.

Aggregate step: Variational Consensus Monte Carlo

The **Variational Consensus Monte Carlo** (VCMC) approach as proposed by Rabinovich *et al.* (2015) [3] formulates the choice of the aggregation function F as a variational inference problem.

$$\min D_{KL}(q_F || p(\cdot | \dot{\boldsymbol{X}}_{1:S})) \text{ subject to } q_F \in \mathcal{Q}_F$$

with :

$$q_F : \theta \mapsto \int_{\Theta^S} \prod_{s=1}^{S} p_s(\theta_s | \dot{\boldsymbol{X}}_s) \delta_{\theta = F(\theta_1, \ldots, \theta_S)} d\theta_{1:S}$$

Aggregate step: Variational Consensus Monte Carlo

The **Variational Consensus Monte Carlo** (VCMC) approach as proposed by Rabinovich *et al.* (2015) [3] formulates the choice of the aggregation function F as a variational inference problem.

$$\min D_{KL}(q_F || p(\cdot|\dot{\boldsymbol{X}}_{1:S})) \text{ subject to } q_F \in \mathcal{Q}_F$$

with :

$$q_F : \theta \mapsto \int_{\Theta^S} \prod_{s=1}^{S} p_s(\theta_s|\dot{\boldsymbol{X}}_s)\delta_{\theta=F(\theta_1,\ldots,\theta_S)} d\theta_{1:S}$$

$\Rightarrow$ Optimisation of the **relaxed variational objective** (blockwise concave under mild conditions).

The **Variational Consensus Monte Carlo** (VCMC) approach as proposed by Rabinovich *et al.* (2015) [3] formulates the choice of the aggregation function F as a variational inference problem.

$$\min D_{KL}(q_F || p(\cdot|\dot{\boldsymbol{X}}_{1:S})) \text{ subject to } q_F \in \mathcal{Q}_F$$

with :

$$q_F : \theta \mapsto \int_{\Theta^S} \prod_{s=1}^{S} p_s(\theta_s|\dot{\boldsymbol{X}}_s)\delta_{\theta=F(\theta_1,...,\theta_S)} d\theta_{1:S}$$

$\Rightarrow$ Optimisation of the **relaxed variational objective** (blockwise concave under mild conditions).

Aggregation function: $F(\theta_1, ..., \theta_S) = \left(\sum_{s=1}^{S} W_s^c \theta_s^c \right)_{c=1}^{C}$
with $\{W_s^c\}_{s\in[\![1,S]\!],c\in[\![1,C]\!]} \in [0,1]^{dim(\Phi)+1}$ a set of weights (**aggregation weights**);

VCMC for mixture models: the matching problem

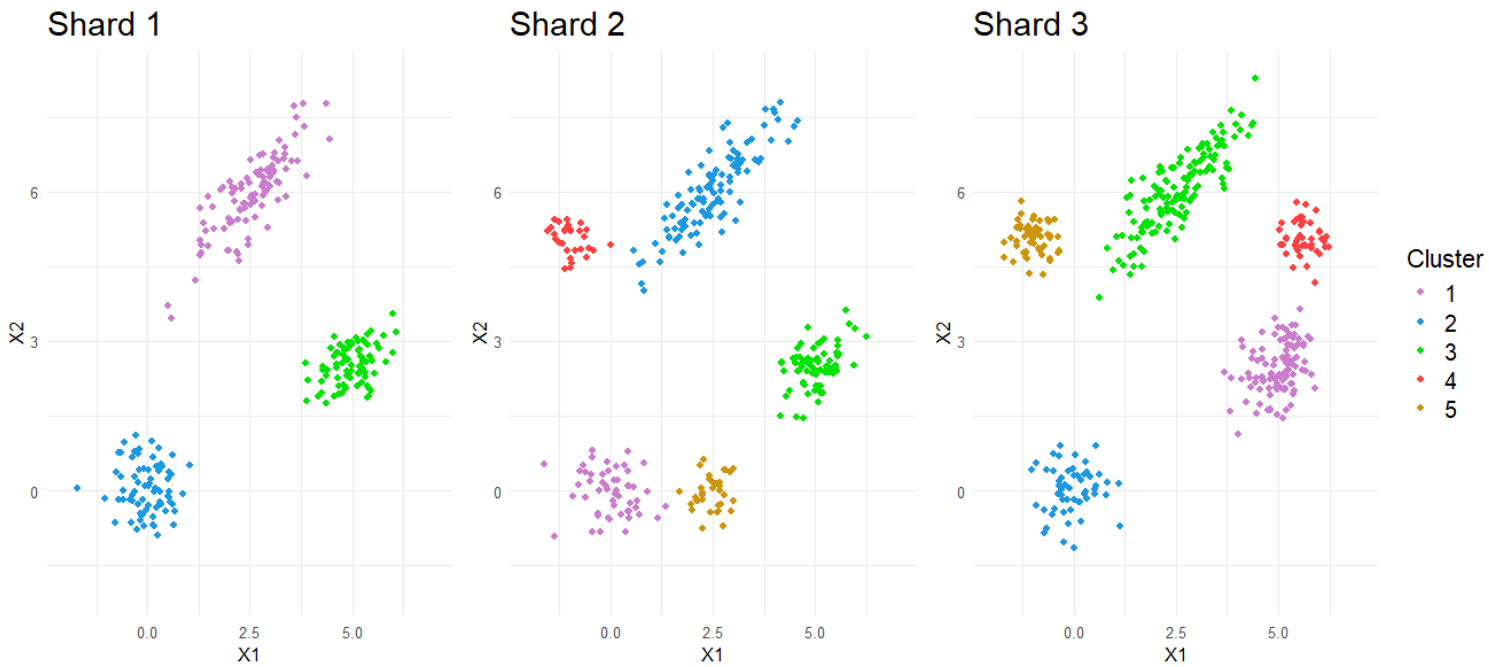
VCMC infers mixture of Gaussian with known number of clusters, mixture weights and variance of the cluster-specific distributions.

- Simplifies the problem of cluster matching between the shards to an alignment problem
- but, often unrealistic in practice.

VCMC for mixture models: the matching problem

VCMC infers mixture of Gaussian with known number of clusters, mixture weights and variance of the cluster-specific distributions.

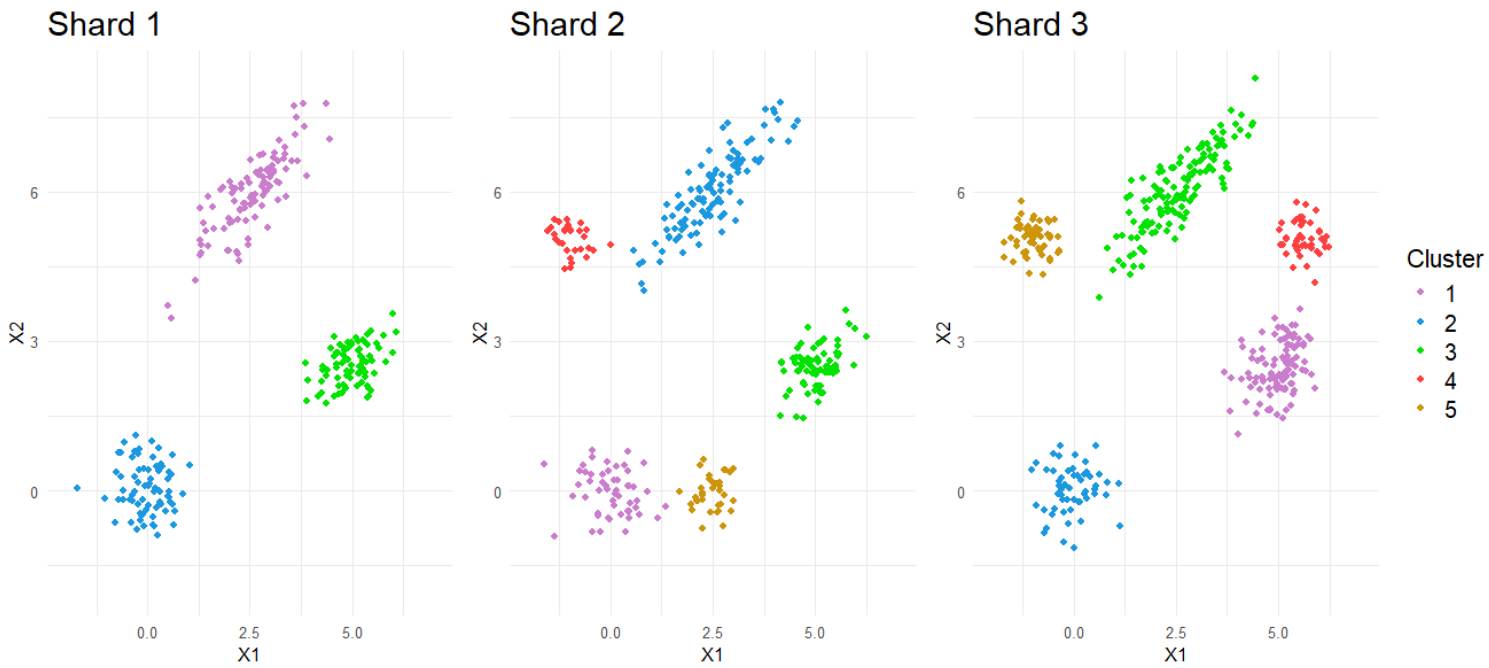
- Simplifies the problem of cluster matching between the shards to an alignment problem
- but, often unrealistic in practice.



VCMC for mixture models: the matching problem

VCMC infers mixture of Gaussian with known number of clusters, mixture weights and variance of the cluster-specific distributions.

- Simplifies the problem of cluster matching between the shards to an alignment problem
- but, often unrealistic in practice.



⇒ We propose two methods for cluster matching that relax these assumptions.

Assumption: There is at least one shard that contains all the clusters and this shard is known.

The following alignment problem is solved using the Hungarian algorithm:

$$\arg \min_{(y_{c_1,c_2})} \sum_{c_1=1}^{\hat{C}_1} \sum_{c_2=1}^{\hat{C}_2} y_{c_1,c_2} d(\phi^1_{c_1}, \phi^2_{c_2})$$

$$\text{s.t } \forall (c_1, c_2), y_{c_1,c_2} \in \{0,1\}, \sum_{c_1=1}^{\hat{C}_1} y_{c_1,c_2} = 1, \sum_{c_2=1}^{\hat{C}_2} y_{c_1,c_2} = 1.$$

where d is a distance between two elements of Φ and $\hat{C}^s$ the number of clusters estimated in shards s.

Cluster matching: Minimum Divergence (1/2)

Let us define $\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'}) \in \mathcal{M}_{n_s \times C^+}$ with $C^+ \in \mathbb{N}^*$ such that

$$\forall i \in [\![1, n_s]\!], \forall j \in [\![1, C^+]\!], \left(\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'})\right)_{ij} = \frac{\mathbb{P}(C_s^i = j | X_s^i, \phi^{s'})}{n_s}$$

Cluster matching: Minimum Divergence (1/2)

Let us define $\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'}) \in \mathcal{M}_{n_s \times C^+}$ with $C^+ \in \mathbb{N}^*$ such that

$$\forall i \in [\![1, n_s]\!], \forall j \in [\![1, C^+]\!], \left(\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'})\right)_{ij} = \frac{\mathbb{P}(C_s^i = j | X_s^i, \phi^{s'})}{n_s}$$

Two partitions are matched by choosing $\hat{a}(.)$ that solves

$$\hat{a}(.) = \arg\min_{a(.)} \mathbb{E}_{p(\phi^1, \phi^2 | \dot{\boldsymbol{x}})} \left[D_{KL} \left(\boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^2_{a(.)}) || \boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^1) \right) \right]$$

where $a(.)$ is a permutation of $[\![1, C^+]\!]$.

Cluster matching: Minimum Divergence (1/2)

Let us define $\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'}) \in \mathcal{M}_{n_s \times C^+}$ with $C^+ \in \mathbb{N}^*$ such that

$$\forall i \in [\![1, n_s]\!], \forall j \in [\![1, C^+]\!], \left(\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'})\right)_{ij} = \frac{\mathbb{P}(C_s^i = j | X_s^i, \phi^{s'})}{n_s}$$

Two partitions are matched by choosing $\hat{a}(.)$ that solves

$$\hat{a}(.) = \arg\min_{a(.)} \mathbb{E}_{p(\phi^1, \phi^2 | \dot{\boldsymbol{x}})} \left[D_{KL} \left(\boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^2_{a(.)}) || \boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^1) \right) \right]$$

where $a(.)$ is a permutation of $[\![1, C^+]\!]$.

In practice, there exists $\sum_{c_1=1}^{\hat{C}^1} \sum_{c_2=1}^{min(c_1, \hat{C}^2)} \binom{\hat{C}^1}{c_1} \binom{\hat{C}^2}{c_2}$ different combinations.

Cluster matching: Minimum Divergence (1/2)

Let us define $\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'}) \in \mathcal{M}_{n_s \times C^+}$ with $C^+ \in \mathbb{N}^*$ such that

$$\forall i \in [\![1, n_s]\!], \forall j \in [\![1, C^+]\!], \left(\boldsymbol{P}(\dot{\boldsymbol{X}}_s; \phi^{s'})\right)_{ij} = \frac{\mathbb{P}(C_s^i = j | X_s^i, \phi^{s'})}{n_s}$$

Two partitions are matched by choosing $\hat{a}(.)$ that solves

$$\hat{a}(.) = \arg\min_{a(.)} \mathbb{E}_{p(\phi^1, \phi^2 | \dot{\boldsymbol{x}})} \left[D_{KL} \left(\boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^2_{a(.)}) || \boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^1) \right) \right]$$

where $a(.)$ is a permutation of $[\![1, C^+]\!]$.

In practice, there exists $\sum_{c_1=1}^{\hat{C}^1} \sum_{c_2=1}^{min(c_1, \hat{C}^2)} \binom{\hat{C}^1}{c_1} \binom{\hat{C}^2}{c_2}$ different combinations.

We propose a workflow to simplify the problem to an alignment problem.

Cluster matching: Minimum Divergence (2/2)

We consider a mixture model of size $C^m = \max(\hat{C}^1, \hat{C}^2) + 1$.

Cluster matching: Minimum Divergence (2/2)

We consider a mixture model of size $C^m = \max(\hat{C}^1, \hat{C}^2) + 1$.
The problem is an alignment problem with final cost equal to

$$\sum_{c=1}^{C^m} \mathbb{E}_{p(\phi^1,\phi^2|\dot{\boldsymbol{X}})} \left[D_{KL}\left(\boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^2_{\hat{a}(c)}) || \boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^1) \right) \right]$$

Cluster matching: Minimum Divergence (2/2)

We consider a mixture model of size $C^m = \max(\hat{C}^1, \hat{C}^2) + 1$.
The problem is an alignment problem with final cost equal to

$$\sum_{c=1}^{C^m} \mathbb{E}_{p(\phi^1,\phi^2|\dot{\boldsymbol{X}})} \left[D_{KL}\left(\boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^2_{\hat{a}(c)}) || \boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^1) \right) \right]$$

There exists c s.t. $\phi^2_{\hat{a}(c)}$ and ϕ^1_c follow the prior distribution.

Cluster matching: Minimum Divergence (2/2)

We consider a mixture model of size $C^m = \max(\hat{C}^1, \hat{C}^2) + 1$.
The problem is an alignment problem with final cost equal to

$$\sum_{c=1}^{C^m} \mathbb{E}_{p(\phi^1,\phi^2|\dot{\boldsymbol{x}})} \left[D_{KL}\left(\boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^2_{\hat{a}(c)}) || \boldsymbol{P}(\dot{\boldsymbol{X}}_2; \phi^1) \right) \right]$$

There exists c s.t. $\phi^2_{\hat{a}(c)}$ and ϕ^1_c follow the prior distribution.

Cost of matching two clusters at random:

$$\mathbb{E}_{p(\phi^1,\phi^2|\boldsymbol{X})} \left[D_{KL}\left(\boldsymbol{P}(\boldsymbol{X}_2; \phi^2_{\hat{a}(c)}) || \boldsymbol{P}(\boldsymbol{X}_2; \phi^1_{\epsilon(c)}) \right) \right]$$

.

Cluster matching: Minimum Divergence (2/2)

We consider a mixture model of size $C^m = \max(\hat{C}^1, \hat{C}^2) + 1$.
The problem is an alignment problem with final cost equal to

$$\sum_{c=1}^{C^m} \mathbb{E}_{p(\phi^1,\phi^2|\dot{\boldsymbol{x}})}\left[D_{KL}\left(\boldsymbol{P}(\dot{\boldsymbol{X}}_2;\phi^2_{\hat{a}(c)})||\boldsymbol{P}(\dot{\boldsymbol{X}}_2;\phi^1)\right)\right]$$

There exists c s.t. $\phi^2_{\hat{a}(c)}$ and ϕ^1_c follow the prior distribution.

Cost of matching two clusters at random:

$$\mathbb{E}_{p(\phi^1,\phi^2|\boldsymbol{X})}\left[D_{KL}\left(\boldsymbol{P}(\boldsymbol{X}_2;\phi^2_{\hat{a}(c)})||\boldsymbol{P}(\boldsymbol{X}_2;\phi^1_{\epsilon(c)})\right)\right]$$

We "unmatch" all pairs of clusters indexed by $(c', a(c'))$ s.t.:

$$\mathbb{E}_{p(\phi^1,\phi^2|\boldsymbol{X})}\left[D_{KL}\left(\boldsymbol{P}(\boldsymbol{X}_2;\phi^2_{\hat{a}(c')})||\boldsymbol{P}(\boldsymbol{X}_2;\phi^1_{\epsilon(c')})\right)\right] \geq \mathbb{E}_{p(\phi^1,\phi^2|\boldsymbol{X})}\left[D_{KL}\left(\boldsymbol{P}(\boldsymbol{X}_2;\phi^2_{\hat{a}(c)})||\boldsymbol{P}(\boldsymbol{X}_2;\phi^1_{\epsilon(c)})\right)\right].$$

Cluster matching: Ball matching

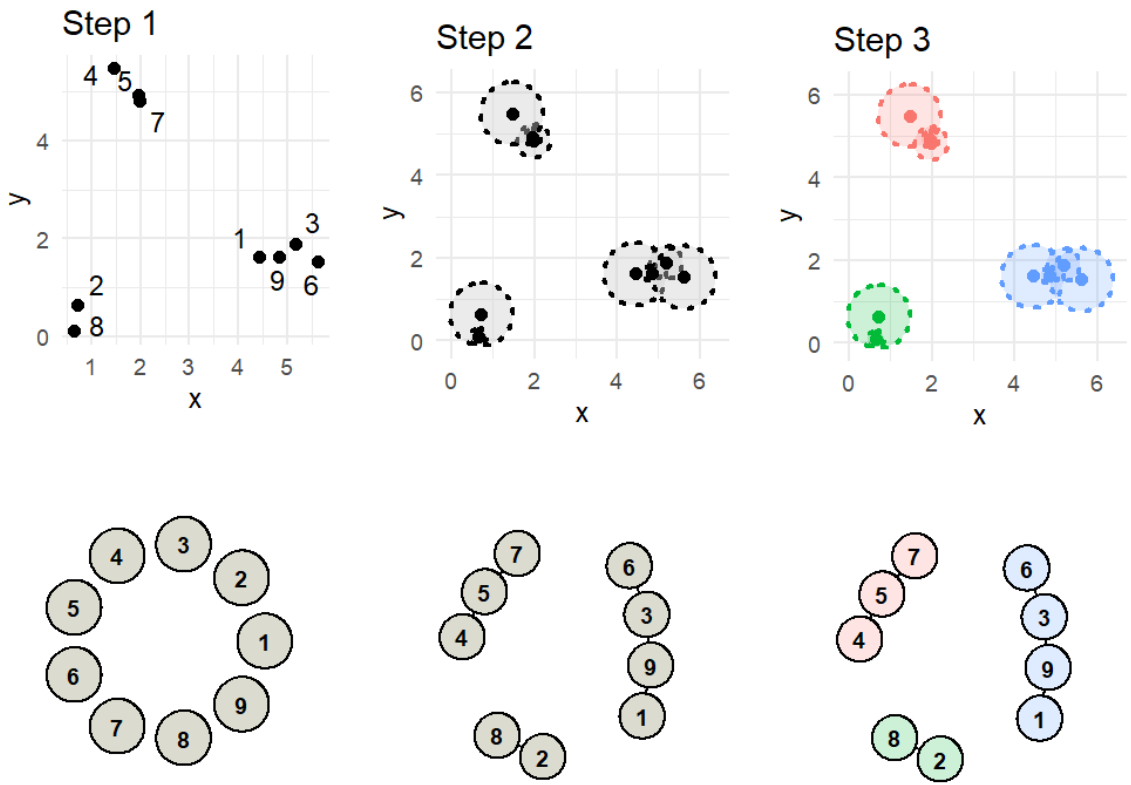


Figure 2: Diagram of the Ball matching procedure.

- **Step 1:** identification of the clusters coordinates;
- **Step 2:** identification of the close clusters;
- **Step 3:** identification of the groups of close clusters (connected components).

Performance of the algorithm: description of the simulations (1/2)

- We perform a simulation study to:
  - compare the different matching procedures;
  - assess the performance of our algorithm;
  - compare VCMC with other frameworks.

Performance of the algorithm: description of the simulations (1/2)

- We perform a simulation study to:
  - compare the different matching procedures;
  - assess the performance of our algorithm;
  - compare VCMC with other frameworks.
- Questions studied :
  - Can we recover the clustering structure of the data ?
  - Can we infer the model parameters ?

Performance of the algorithm: description of the simulations (1/2)

- We perform a simulation study to:
  - compare the different matching procedures;
  - assess the performance of our algorithm;
  - compare VCMC with other frameworks.
- Questions studied :
  - Can we recover the clustering structure of the data ?
  - Can we infer the model parameters ?
- Simulated data:
  - Each simulation scenario is composed of 100 datasets.
  - Each dataset contains:
    - 4000 data points distributed across 4 shards;
    - 10 **categorical variables** with cluster-specific distributions.

Performance of the algorithm: description of the simulations (1/2)

- We perform a simulation study to:
  - compare the different matching procedures;
  - assess the performance of our algorithm;
  - compare VCMC with other frameworks.
- Questions studied :
  - Can we recover the clustering structure of the data ?
  - Can we infer the model parameters ?
- Simulated data:
  - Each simulation scenario is composed of 100 datasets.
  - Each dataset contains:
    - 4000 data points distributed across 4 shards;
    - 10 **categorical variables** with cluster-specific distributions.
- Inference model:
  - Over-fitted mixture of categorical distributions;
  - Prior distribution on the mixture weights: symmetric Dirichlet distribution $\mathcal{D}ir(0.5)$;
  - Prior distribution on the parameters of the categorical distributions: symmetric Dirichlet distribution $\mathcal{D}ir(0.5)$.

Performance of the algorithm: description of the simulations (2/2)

Distribution of the clusters across the shards :

- Homogeneous ;
- Heterogeneous nested;
- Heterogeneous non-nested;

Separability of the clusters:

- Easily separable;
- Poorly separable.

Distribution of the clusters across the shards :

- Homogeneous ;

Shard \ Cluster	1	2	3	4	5
1	1/5	1/5	1/5	1/5	1/5
2	1/5	1/5	1/5	1/5	1/5
3	1/5	1/5	1/5	1/5	1/5
4	1/5	1/5	1/5	1/5	1/5
Global	0.20	0.20	0.20	0.20	0.20

Table 1: Distribution of the clusters across the homogeneous shards and in the total datasets (global)

- Heterogeneous nested;
- Heterogeneous non-nested;

Separability of the clusters:

- Easily separable;
- Poorly separable.

Distribution of the clusters across the shards :

- Homogeneous ;
- Heterogeneous nested;

Shard \ Cluster	1	2	3	4	5	6
1	1/6	1/6	1/6	1/6	1/6	1/6
2	1/5	1/5	1/5	1/5	1/5	0
3	1/3	1/3	1/3	0	0	0
4	1/2	1/2	0	0	0	0
Global	0.30	0.30	≈ 0.18	≈ 0.09	≈ 0.09	≈ 0.04

Table 1: Distribution of the clusters across the heterogeneous shards and in the total datasets (global)

- Heterogeneous non-nested;

Separability of the clusters:

- Easily separable;
- Poorly separable.

Distribution of the clusters across the shards :

- Homogeneous ;
- Heterogeneous nested;
- Heterogeneous non-nested;

Shard \ Cluster	1	2	3	4	5	6
1	1/5	1/5	1/5	1/5	1/5	0
2	0	1/5	1/5	1/5	1/5	1/5
3	1/3	1/3	1/3	0	0	0
4	1/2	1/2	0	0	0	0
Global	≈ 0.26	≈ 0.31	≈ 0.18	0.10	0.10	0.05

Table 1: Distribution of the clusters across the heterogeneous shards and in the total datasets (global)

Separability of the clusters:

- Easily separable;
- Poorly separable.

Distribution of the clusters across the shards :

- Homogeneous ;
- Heterogeneous nested;
- Heterogeneous non-nested;

Separability of the clusters:

- Easily separable;
- Poorly separable.

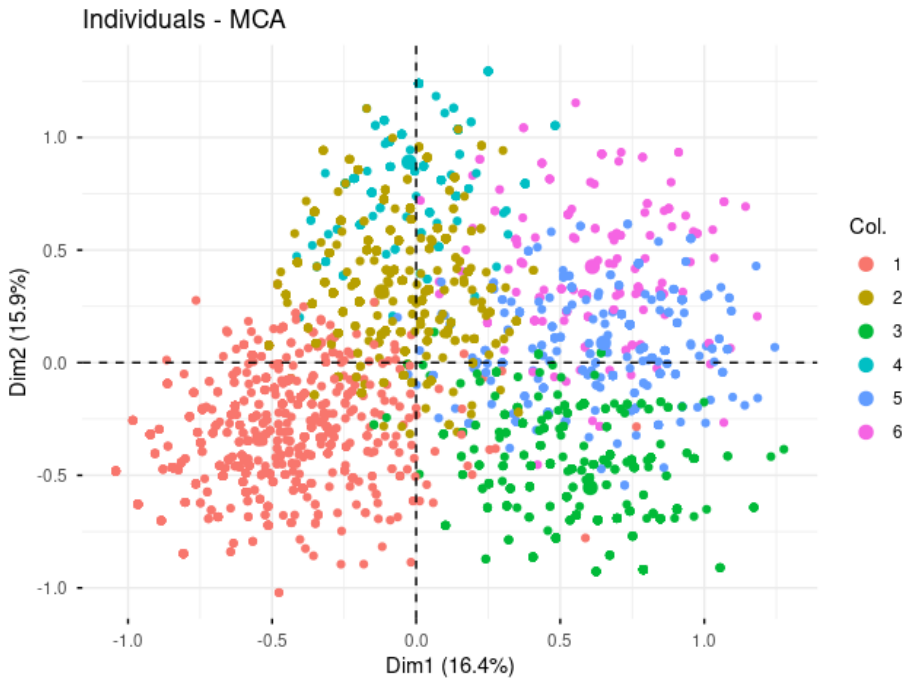


Figure 3: MCA coordinates on the first and second axes of the data in the case of easy separability of the clusters

Distribution of the clusters across the shards :

- Homogeneous ;
- Heterogeneous nested;
- Heterogeneous non-nested;

Separability of the clusters:

- Easily separable;
- Poorly separable.

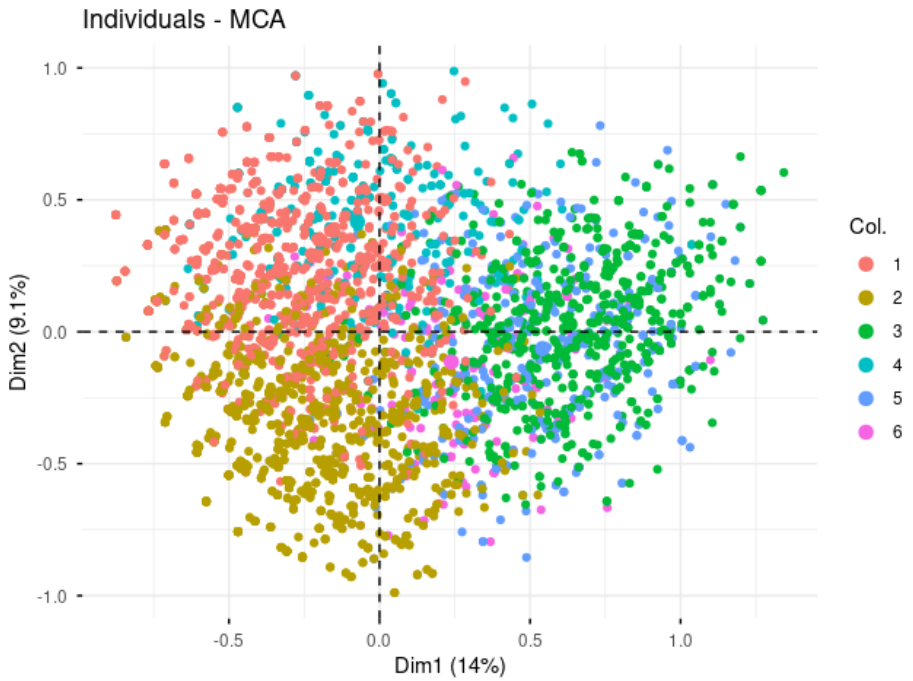


Figure 3: MCA coordinates on the first and second axes of the data in the case of poor separability of the clusters

Matching comparison

Type of shards	Cluster separability	Matching method	Correct matching recovered	mean ARI
Homogeneous	easy	Hungarian	61%	0.88
		Minimum divergence	61%	0.96
		Ball	61%	0.96
	poor	Hungarian	69%	0.57
		Minimum divergence	69%	0.62
		Ball	71%	0.62
Heterogeneous nested	easy	Hungarian	63%	0.97
		Minimum divergence	55%	0.96
		Ball	58%	0.97
	poor	Hungarian	71%	0.73
		Minimum divergence	71%	0.72
		Ball	73%	0.73
Heterogeneous non-nested	easy	Hungarian	0%	0.90
		minimum divergence	58%	0.97
		Ball	58%	0.98
	poor	Hungarian	0%	0.69
		Minimum divergence	74%	0.76
		Ball	76%	0.77

Table 1: Percentage of simulated data for which the correct matching is recovered and mean adjusted Rand index (ARI) between the estimated partition and the true partition on the full data when the partition in the local shards is estimated for 100 simulated datasets in different data simulation scenarios.

We assess the performance of our method and compare it to:

- **Full MCMC:** an MCMC algorithm on the full dataset;
- **Fed K-modes:** a federated version of the K-modes algorithm [4]
- **SNOB** [5] & **SIGN** [6]: a federated MCMC for semi-parametric Bayesian mixture models;
- **FedMerDel** [7]: a federated variational inference algorithm for over-fitted mixture models.

Methods comparison: heterogeneous non-nested shards with poor separability

	Fed K-modes	FedMerDel	SNOB & SIGN	VCMC	Full MCMC
Mean running time (sd)	8.79 (2.15)	2.17 (0.30)	2615 (126)	4390 (2353)	4488 (123)
Mean number of clusters (sd)	6 (0)	5.85 (0.43)	5.52 (0.72)	7.09 (1.33)	6.07 (0.38)
Mean ARI (sd)	0.31 (0.10)	0.70 (0.12)	0.72 (0.09)	0.74 (0.06)	0.69 (0.07)
Mean VI (sd)	2.16 (0.29)	1.04 (0.27)	0.97 (0.19)	0.95 (0.19)	1.20 (0.20)

Table 2: Mean running time in seconds (s), estimated number of clusters, adjusted Rand index (ARI) and variation of information (VI) and standard deviations (sd) for 100 simulated datasets.

Methods comparison: heterogeneous non-nested shards with poor separability

	Fed K-modes	FedMerDel	SNOB & SIGN	VCMC	Full MCMC
Mean running time (sd)	8.79 (2.15)	2.17 (0.30)	2615 (126)	4390 (2353)	4488 (123)
Mean number of clusters (sd)	6 (0)	5.85 (0.43)	5.52 (0.72)	7.09 (1.33)	6.07 (0.38)
Mean ARI (sd)	0.31 (0.10)	0.70 (0.12)	0.72 (0.09)	0.74 (0.06)	0.69 (0.07)
Mean VI (sd)	2.16 (0.29)	1.04 (0.27)	0.97 (0.19)	0.95 (0.19)	1.20 (0.20)

Table 2: Mean running time in seconds (s), estimated number of clusters, adjusted Rand index (ARI) and variation of information (VI) and standard deviations (sd) for 100 simulated datasets.

	$p^q_{k,c} < 0.0001$		$p^q_{k,c} \in [0.0001, 0.1[$		$p^q_{k,c} \in [0.1, 0.5[$		$p^q_{k,c} \in [0.5, 0.9[$		$p^q_{k,c} \in [0.9, 0.9999[$		$p^q_{k,c} \geq 0.9999$	
Algorithm	RB	IC95	RB	IC95	RB	IC95	RB	IC95	RB	IC95	RB	IC95
Fed K-modes	-1e4	na	-9.244	na	-0.011	na	0.086	na	0.1714	na	na	na
FedMerDel	-1e2	0.00	-1.058	0.05	-0.041	0.07	0.027	0.09	0.021	0.08	na	na
SNOB & SIGN	-2e2	0.00	-0.333	0.30	0.320	0.25	0.340	0.23	0.347	0.20	na	na
VCMC	-1e3	0.00	-3.043	0.34	-0.058	0.75	0.041	0.61	0.049	0.33	na	na
Full	-2e3	0.00	-3.325	0.23	-0.078	0.63	0.053	0.45	0.056	0.24	na	na

Table 3: Empirical relative bias (RB) and 95% coverage interval (IC95) on the proportion of individuals with modality k on covariate q within cluster c, $p^q_{k,c}$, estimated on 100 simulated datasets within six ranges of values of $p^q_{k,c}$. (na: not applicable)

Description of the database

We consider a subset of The Health Improvement Network (THIN) database:

- **Electronic Health Records** (EHRs);
- N = 289,821 individuals over 80 years old in the United-Kingdom;
- 94 conditions studied: binary variables indicating the occurrence of the condition.

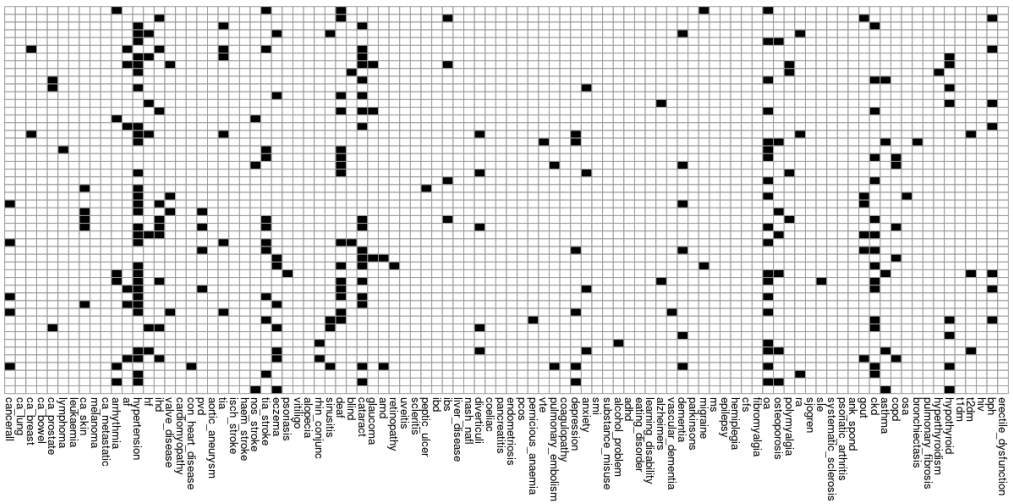


Figure 4: Overview of the THIN database for 50 individuals (synthetic data).

Description of the data (2/2)

Mortality (%)	48.1
Sex (%)	
Male	35.3
Female	64.7
Age at entry(in years)	
Mean (sd)	85.4 (4.4)
Min - Max	80 - 112
Age at death (in years)	
Mean (sd)	89.4 (4.7)
Min - Max	80 - 110
Number of co-occurring diseases	
Mean (sd)	6.2 (3.1)
Min - Max	2 - 33

Table 4: Description of the studied subset of the THIN data (N = 289,821)

Region of residency (%)	
East Midlands	2.4
East of England	5.7
London	10.4
North East	1.8
North West	9.2
Northern Ireland	3.9
Scotland	14.9
South Central	10.4
South East Coast	11.0
South West	8.7
Wales	12.0
West Midlands	8.0
Yorkshire & Humber	1.6

Table 5: Region of residency of the studied subset of the THIN data

The data shards are defined according to the region of residency.

Identifying patterns of muti-morbidity in the THIN database

Figure 5: Scaled difference between the proportion of individuals presenting each disease in each cluster and the proportions in the full data.

Summary

- We extended the Variational Consensus Monte Carlo framework to infer over-fitted mixture models in a Federated Learning setting.
- When the shards match the data collection pattern, the federated learning framework outperformed the inference algorithm on the full data.
- Unlike other state-of-the-arts methods, **no prior-posterior conjugacy** is required to perform the inference in a fully federated setting.
- We developed **three algorithmic strategies** that adapt to different federated learning settings.

Variational Consensus Monte Carlo for Bayesian Mixture

Limits and Future work

- **Minimum divergence matching:** can not match clusters within shards at the global level.
- **Ball matching:** based on convergence results that assume a uniform prior distribution on the model parameters.
- **Projected gradient descent algorithm:** very slow to converge for some datasets.

Variational Consensus Monte Carlo for Bayesian Mixture

[1] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. DOliveira, H. Eichner, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, H. Qi, D. Ramage, R. Raskar, M. Raykova, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 12, pp. 1–210, 2021.

[2] B. Pfitzner, N. Steckhan, and B. Arnrich, "Federated learning in a medical context: A systematic literature review," *ACM Trans. Internet Technol.*, vol. 21, June 2021.

[3] M. Rabinovich, E. Angelino, and M. I. Jordan, "Variational consensus monte carlo," in *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, (Cambridge, MA, USA), p. 12071215, MIT Press, 2015.

[4] S. Garst and M. Reinders, "Federated k-means clustering," in *Pattern Recognition* (A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal, eds.), (Cham), pp. 107–122, Springer Nature Switzerland, 2025.

[5] D. A. Zuanetti, P. Müller, Y. Zhu, S. Yang, and Y. Ji, "Bayesian nonparametric clustering for large data sets," *Stat. Comput.*, vol. 29, pp. pp. 203–215, 2019.

[6] Y. Ni, P. Müller, M. Diesendruck, S. Williamson, Y. Zhu, and Y. Ji, "Scalable bayesian nonparametric clustering and classification," *Journal of Computational and Graphical Statistics*, vol. 29, no. 1, pp. pp. 53–65, 2020.

[7] J. Rao, F. L. Crowe, T. Marshall, S. Richardson, and P. D. W. Kirk, "Federated variational inference for bayesian mixture models," 2025.

Method development

Paul D.W. Kirk
MRC Biostatistics Unit
University of Cambridge

Sylvia Richardson
MRC Biostatistics Unit
University of Cambridge

Data collection & Data Management

Francesca L. Crowe
Institute of Applied Health Research
University of Birmingham

Tom Marshall
Institute of Applied Health Research
University of Birmingham

Method development

Paul D.W. Kirk
MRC Biostatistics Unit
University of Cambridge

Sylvia Richardson
MRC Biostatistics Unit
University of Cambridge

Data collection & Data Management

Francesca L. Crowe
Institute of Applied Health Research
University of Birmingham

Tom Marshall
Institute of Applied Health Research
University of Birmingham

Thank you for listening!