Distance-based probabilistic clustering for functional data

Group members: Giulia Caruso, Alessio Facincani, Giulia Romani, Pietro Spina, Matteo Vescovi

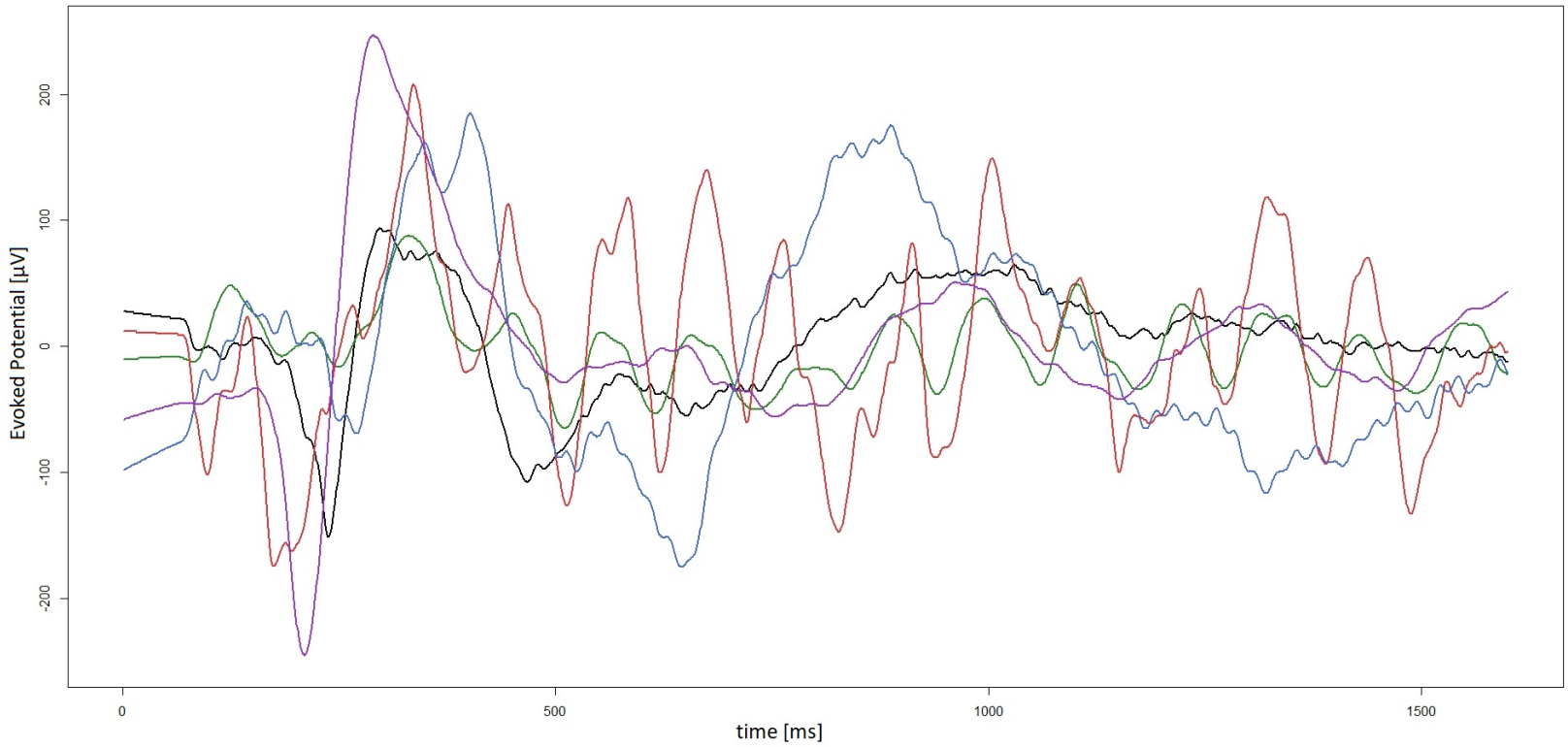
Tutors: Mario Beraha, Riccardo Corradin

POLITECNICO
MILANO 1863

Framework of the project

26 **multivariate functional observations**, each component is a function observed at 1600 time points.

RESEARCH GOAL: **Cluster** observations in different groups, considering only one functional component.



Model

Model

- *Gibbs posterior*:

$$\pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \propto \pi(\boldsymbol{c}) \exp\{-\lambda \ell(\boldsymbol{c}; \boldsymbol{X})\}$$

- *Loss function* under a GB-PPM (Generalized Bayes Product Partition Model) with *Mahalanobis distance* $M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k}) \geq 0$:

$$\ell(\boldsymbol{c}; \boldsymbol{X}) = \sum_{k=1}^{K} \sum_{i \in C_K} \mathcal{D}(\boldsymbol{x_i}; \boldsymbol{X_k}) = \sum_{k=1}^{K} \sum_{i \in C_K} M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k})$$

- *Generalized Bayes posterior under a GB-PPM*:

$$\pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \propto \pi(\boldsymbol{c}) \prod_{k=1}^{K} exp\{-\lambda \sum_{i \in C_K} M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k})\}$$

Mahalanobis distance in a functional context

Given a stochastic process $X(t) \in \mathbb{L}^2[0,1]$, $t \in [0,1]$:

- *Covariance function* K *and operator* $\mathcal{K}$:

$$K(s,t) = Cov(X(s), X(t)) \qquad \mathcal{K}f(t) = \int_0^1 K(t,s)f(s)ds$$

- α-*Mahalanobis distance* with *smoothing parameter* $\alpha > 0$: $(\forall x, y \in \mathbb{L}^2[0,1])$

$$M_\alpha(x,y)^2 = \|x_\alpha - y_\alpha\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)^2} \langle x - y, e_j \rangle^2$$

- α-*approximation* of a function x $\in \mathbb{L}^2[0,1]$:

$$x_\alpha = \underset{f \in \mathcal{H}(K)}{\arg\min} \|x - f\|^2 + \alpha \|f\|_K^2 = \sum_{j=1}^{\infty} \frac{\lambda_j}{(\lambda_j + \alpha)} \langle x, e_j \rangle e_j$$

Maximum a posteriori estimation

MAP with uniform prior

Target of inference: optimal and unknown partition c_{opt}

- Uniform prior (Stirling number of the second kind):

$$\pi(\boldsymbol{c}) = \frac{1}{\mathcal{S}(n, K)}$$

- Posterior:

$$\pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \propto \prod_{k=1}^{K} exp\{-\lambda \sum_{i \in C_K} M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k})\}$$

- Posterior maximization:

$$\begin{aligned} c_{opt} &= \underset{c \,:\, |C|=K}{\arg\max}\, \pi(\boldsymbol{c}|\lambda, \boldsymbol{X}) \\ &= \underset{c \,:\, |C|=K}{\arg\min}\, \ell(\boldsymbol{c}; \boldsymbol{X}) \\ &= \underset{c \,:\, |C|=K}{\arg\min} \sum_{k=1}^{K} \sum_{i \in C_K} M_\alpha(\boldsymbol{x_i}; \boldsymbol{X_k}) \end{aligned}$$

MAP with Pitman-Yor EPPF prior

Target of inference: optimal and unknown partition c_{opt}

- PY-EPPF prior with parameters σ, θ:

$$\pi_{PY}(C, \sigma, \theta) = \frac{\prod_{j=1}^{K-1}(\theta + j\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{K}(1-\sigma)_{n_j-1}$$

$\theta > -\sigma$, $\sigma \in [0,1)$ $\quad (b)_n = \Gamma(b+n)/\Gamma(b)$ $\quad \sum_{j=1}^{K} n_j = n$

- Posterior:

$$\pi(c|\lambda, \sigma, \theta, \boldsymbol{X}) \propto \pi_{PY}(c, \sigma, \theta) \cdot \exp\{-\lambda \ell(\boldsymbol{c}; \boldsymbol{X})\}$$

$$\propto \frac{1}{(\theta + K\sigma)} \cdot \prod_{j=1}^{K}(\theta + j\sigma) \cdot \Gamma(n_j - \sigma) \cdot \exp\{-\lambda \ell(\boldsymbol{c}; \boldsymbol{X})\}\}$$

- Posterior maximization:

$$c_{opt} = \underset{c \,:\, \sum_j n_j = n}{\arg\max}\ \pi(\boldsymbol{c}|\lambda, \sigma, \theta, \boldsymbol{X})$$

Uncertainty quantification: Gibbs sampling

Full conditionals

Idea: Cyclically re-allocate c_i by sampling from their full-conditionals.
Let $\boldsymbol{c}_{-i} = (c_1, ..., c_{i-1}, c_{i+1}, ..., c_n)$ and $\{C_{1,-i}, ..., C_{K,-i}\}$

- With uniform prior:

$$\mathbb{P}(c_i = k | \boldsymbol{c}_{-i}, -) \propto exp\left[-\lambda\left[\sum_{j \in C_k} M_\alpha(\boldsymbol{x}_j; \boldsymbol{X_k}) - \sum_{j \in C_{k,-i}} M_\alpha(\boldsymbol{x}_j; \boldsymbol{X_{k,-i}})\right]\right]$$
$$= \frac{\rho(C_k; \lambda, X_k)}{\rho(C_{k,-i}; \lambda, X_{k,-i})}$$

- With PY-EPPF prior:

$$\mathbb{P}(c_i = k | \boldsymbol{c}_{-i}, -) \propto \begin{cases} (n_k - \sigma) \cdot \frac{\rho(C_k;\lambda,X_k)}{\rho(C_{k,-i};\lambda,X_{k,-i})}, & k = 1, ..., K \\ \frac{(\theta + k\sigma)}{\theta + n}, & k = K + 1 \end{cases}$$

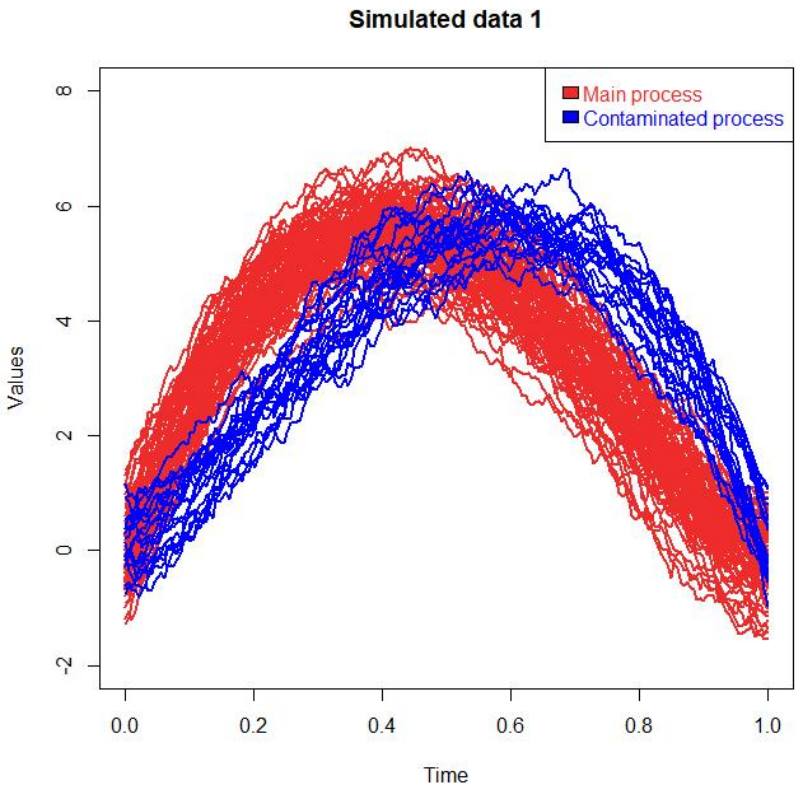
Simulated data

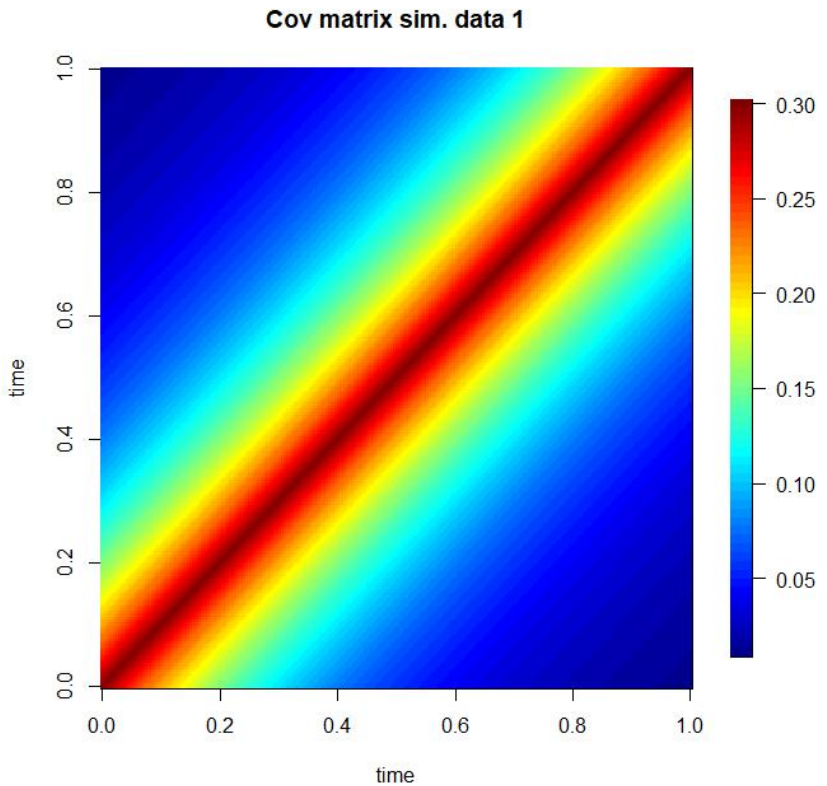
Simulated data 1

Data clustered into two groups:

- Main process: $X(t) = 30t(1-t)^{3/2} + \epsilon(t)$
- Contaminated process: $X(t) = 30t^{3/2}(1-t) + \epsilon(t) \quad t \in [0,1]$

$\epsilon(t) \sim \text{GP}(0, \mathcal{C})$ where $\mathcal{C}(s,t) = 0.3 \cdot exp(-|s-t|/0.3)$

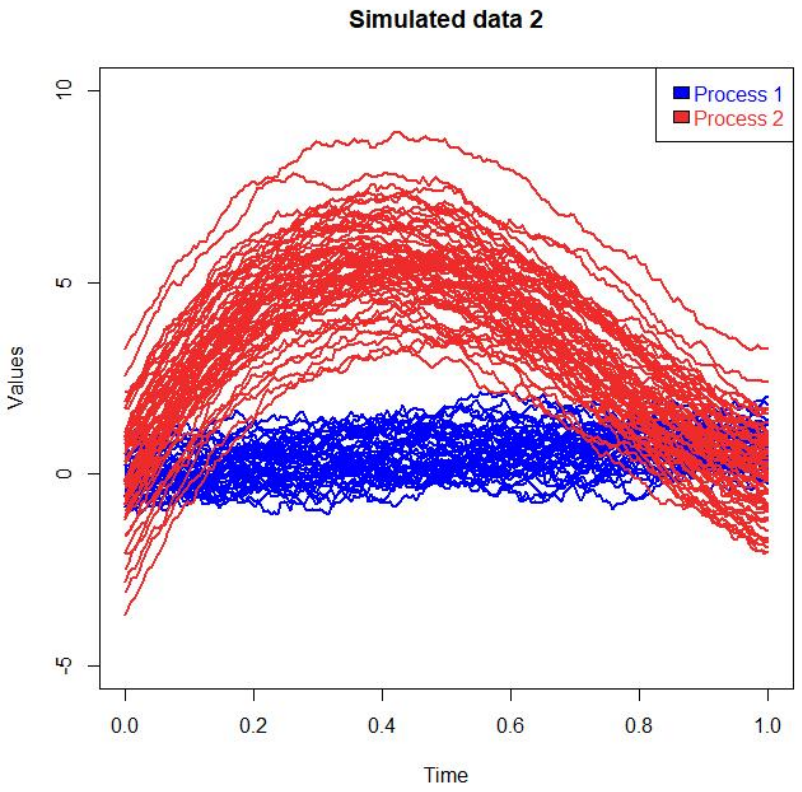


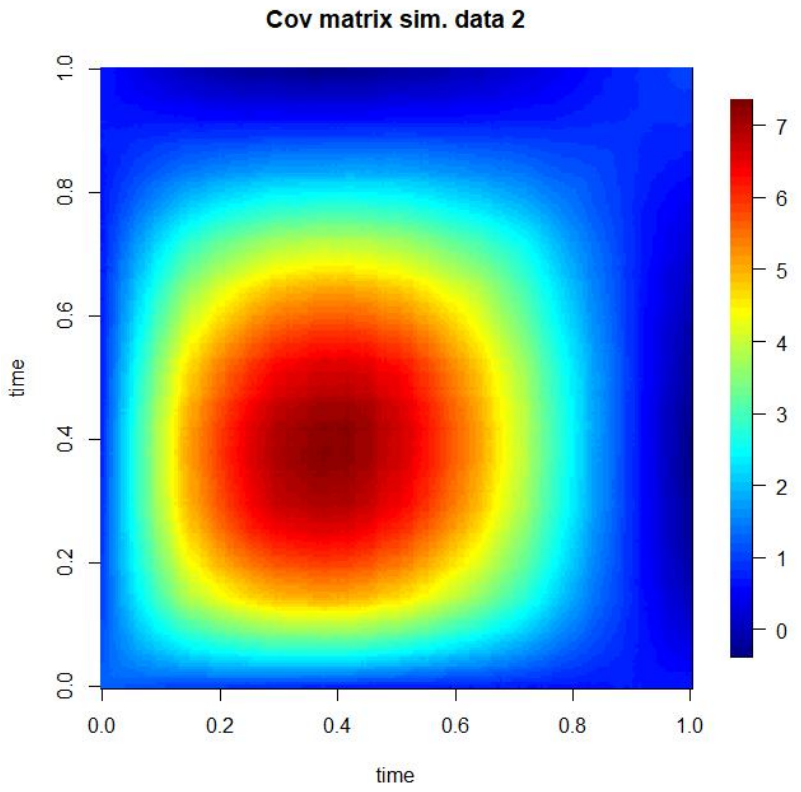


Simulated data 2

Data clustered into two groups:

- First process: $X(t) = sin(t) + \epsilon_1(t) \quad \epsilon_1(t) \sim \text{GP}\ (0, \mathcal{C}_1)$
  $\mathcal{C}_1(s,t) = 0.3 \cdot exp(-|s-t|/0.3)$
- Second process: $X(t) = 30t(1-t)^{3/2} + \epsilon_2(t) \quad \epsilon_2(t) \sim \text{GP}\ (0, \mathcal{C}_2)$
  $\mathcal{C}_2(s,t) = 1.5 \cdot exp(-|s-t|/3) \qquad (t \in [0,1])$



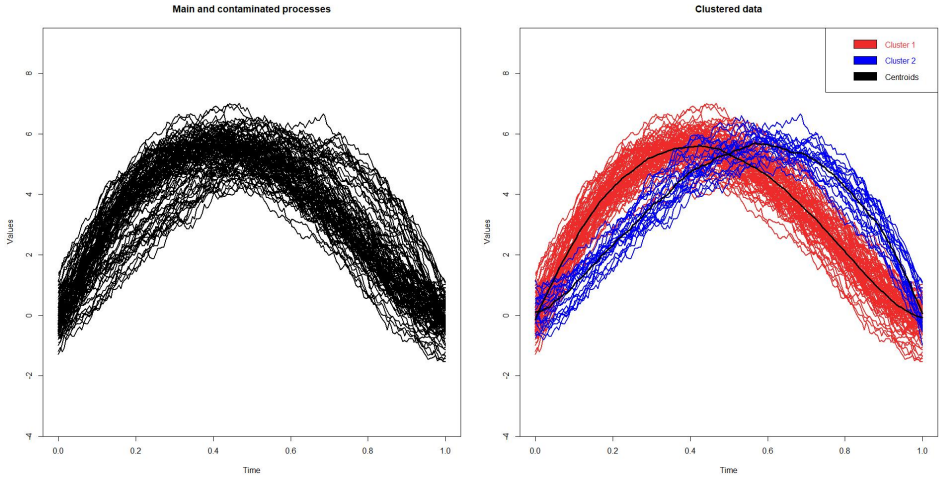


Clustering Algorithms

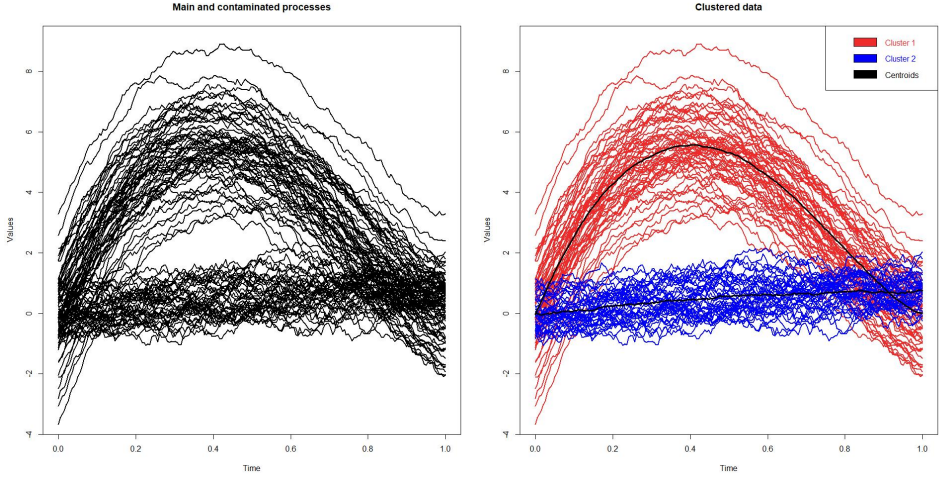
In the last presentation

Mahalanobis distance clustering algorithm with:

- Uniform prior and fixed covariance

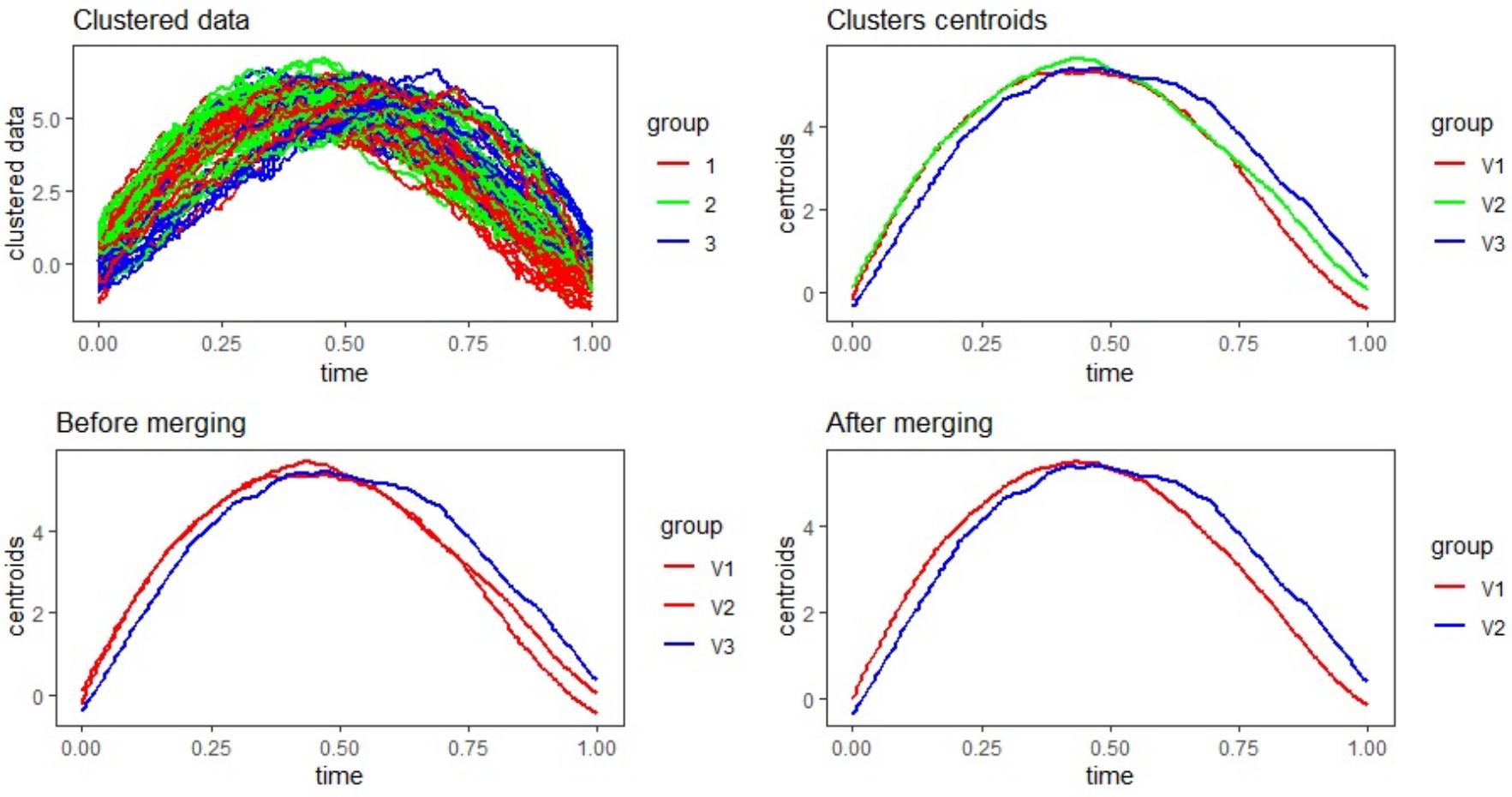


- Uniform prior and covariance updating



In the last presentation

Number of clusters K has to be given a priori $\rightarrow$ Union of similar clusters



MAP with PY-EPPF prior

Algorithm 1: Mahalanobis distance clustering with PY-EPPF prior and fixed covariance

Input : K, α, σ, θ, λ, cov, X
Output: Optimal partition (c, $\{m_k\}_k$, ℓ_1, $post_1$)

Compute eigenvalues and eigenvectors of the covariance matrix;
Sample random observations as initial centroids $m_1, \ldots, m_K$;
Initialize $\ell_1(c, X)$ and $post_1(c, \ell_1, X)$, $\ell_2(c, X)$ and $post_2(c, \ell_2, X)$;
while $post_2(c, \ell_2, X) > post_1(c, \ell_1, X)$ **do**
    $\ell_1(c, X) = \ell_2(c, X)$;
    $post_1(c, \ell_1, X) = post_2(c, \ell_2, X)$;
    for i=1,...,n **do**
        Set the cluster indicator c_i equal to k, so that $M_\alpha(\mathbf{x}_i, \mathbf{m}_k)$ is minimum
    end
    for k=1,...,K **do**
        Check that there are no empty clusters: otherwise, sample an observation and assign it to the empty k;
        Set m_k as the functional mean of the observations belonging to cluster k;
    end
    Update $\ell_2(c, X)$ and $post_2(c, \ell_2, X)$
end

MAP with PY-EPPF prior

Problem 1: one execution of Algorithm (1) $\not\Rightarrow c_{opt}$, due to initialization stochasticity

Problem 2: prior has support on the partitions of n observations with K not fixed

Solution:

Algorithm 2: Mahalanobis distance clustering with PY-EPPF and fixed covariance (overall)

for $K=1,\dots,\bar{K}$ **do**
    Run Algorithm (1) n_{simul} times with K clusters;
    Return the partition $\hat{c}_K$ with the highest posterior $\widehat{post}_K$;
end
Return (c_{opt}, $post_{opt}$, K_{opt}) with $post_{opt}$ highest among $\widehat{post}_K$

Gibbs sampling

Algorithm 3: Gibbs sampling for $\boldsymbol{c}_i$

Input : N_{iter}, $N_{burn-in}$, c_0, X, λ, α, K
Output: Matrix $C = (c_{ij})_{ij}$, clust. allocation of obs j at iteration i

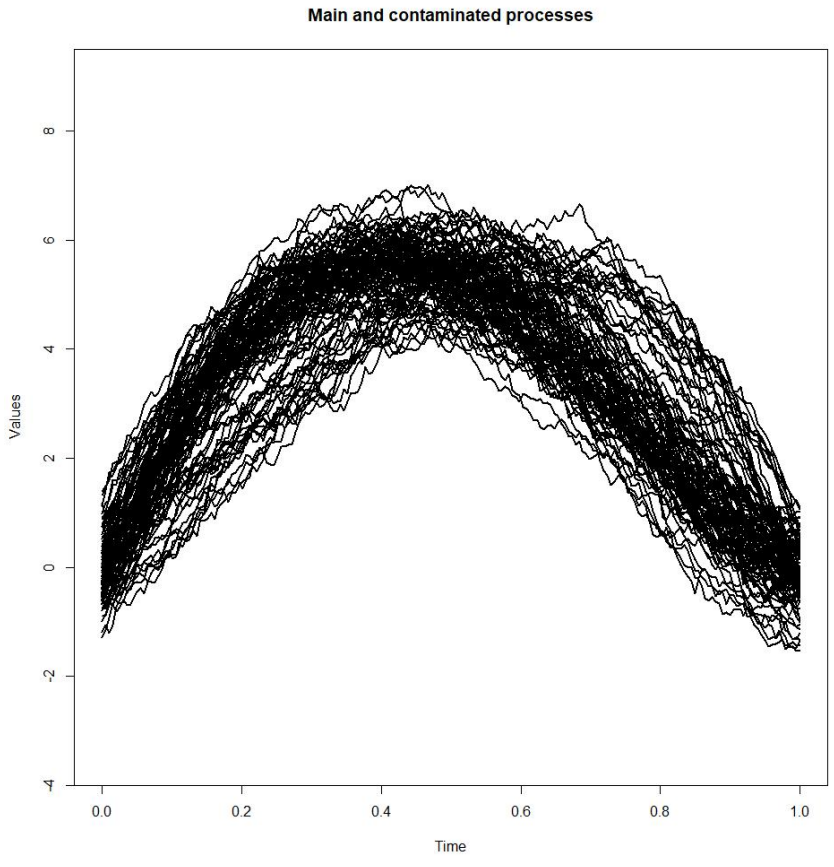
Initialize $t = 1$ and $p = 0$;
while $p < N_{iter}$ **do**
  for *i=1,...,n* **do**
    Sample c_i^t from its full-conditional, i.e. the allocation for obs i at iteration t
  end
  if $t > N_{burn-in}$ **then**
    $p = t - N_{burn-in}$
    Insert $\boldsymbol{c}_t$ at p-th row of C
  end
  t = t+1
end

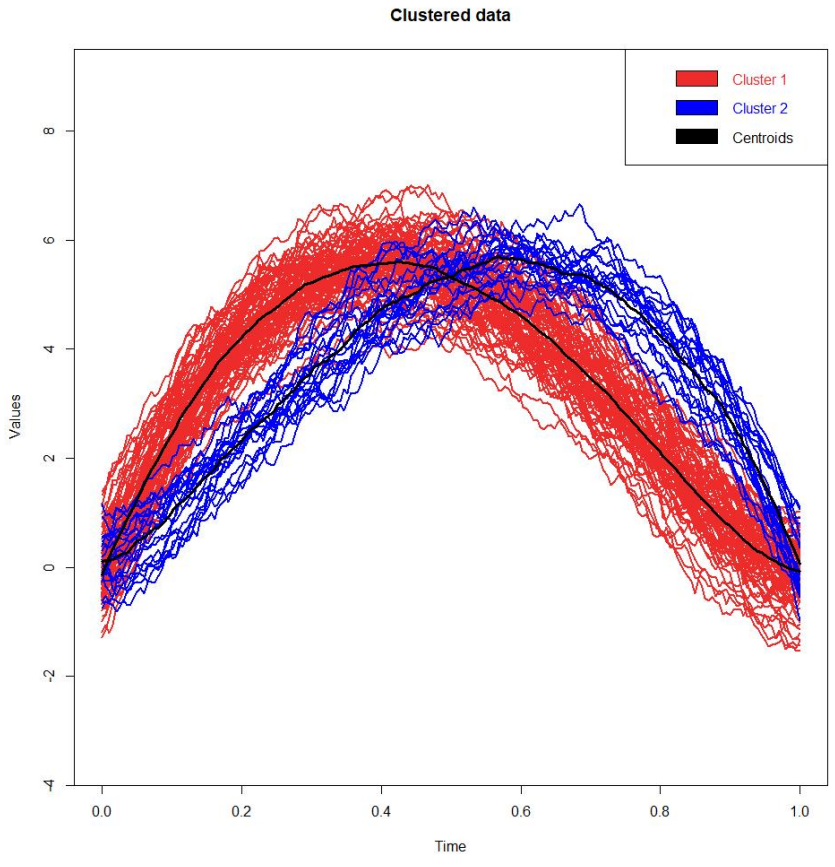
How to interpret the result?

* c_{opt} minimizes the posterior expected Binder's loss function
* Posterior similarity matrix: probability that two units are in the same cluster.

MAP: Application on simulated data 1

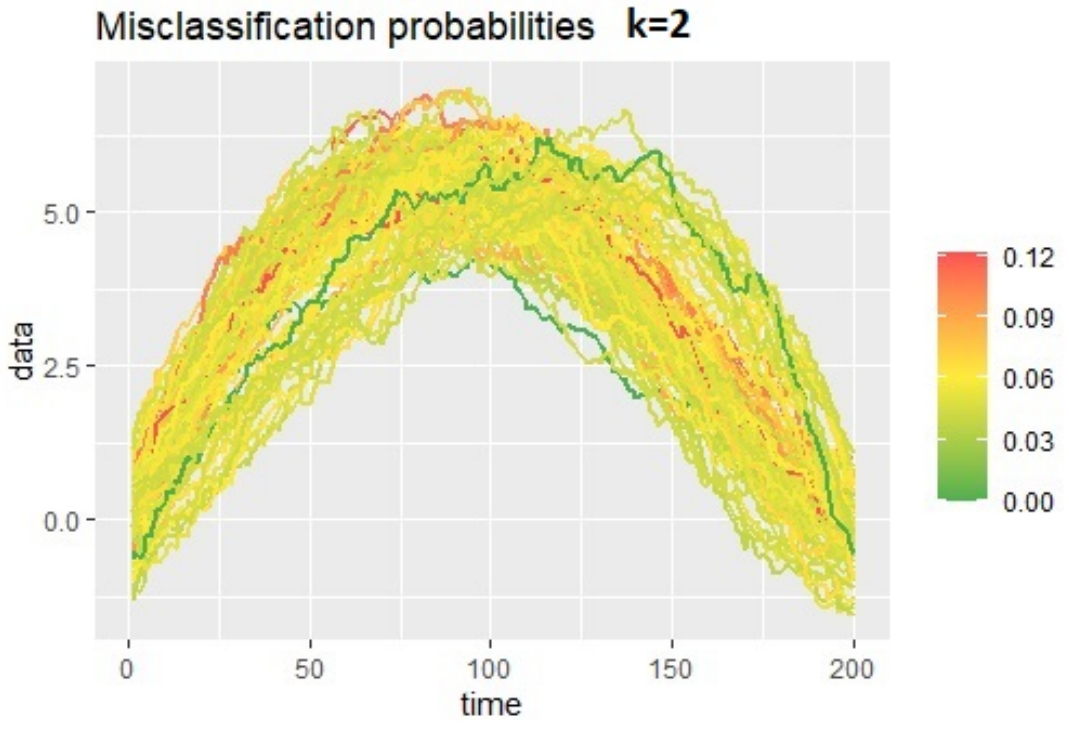
Algorithm (2) with $n_{simul} = 100$ and $\lambda = 0.75$, $\sigma = 0.50$, $\theta = -0.4$



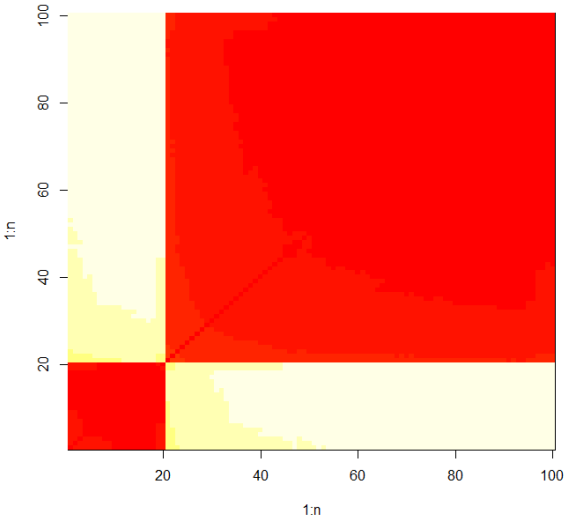


Parameters' calibration: different combinations of $(\lambda, \sigma, \theta)$ lead to c_{opt}

Gibbs: Application on simulated data 1



(a) Misclassification probability

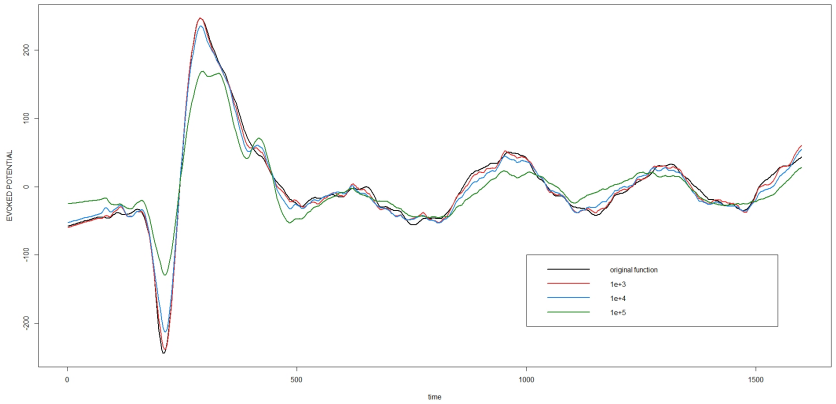


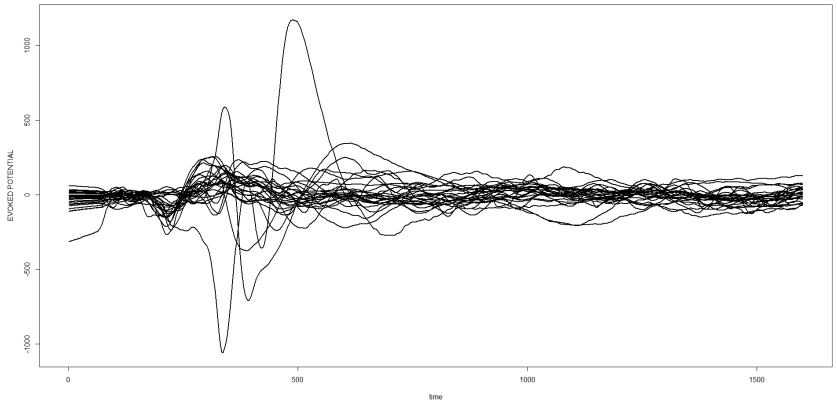
(b) Similarity matrix

Application on Clinical data

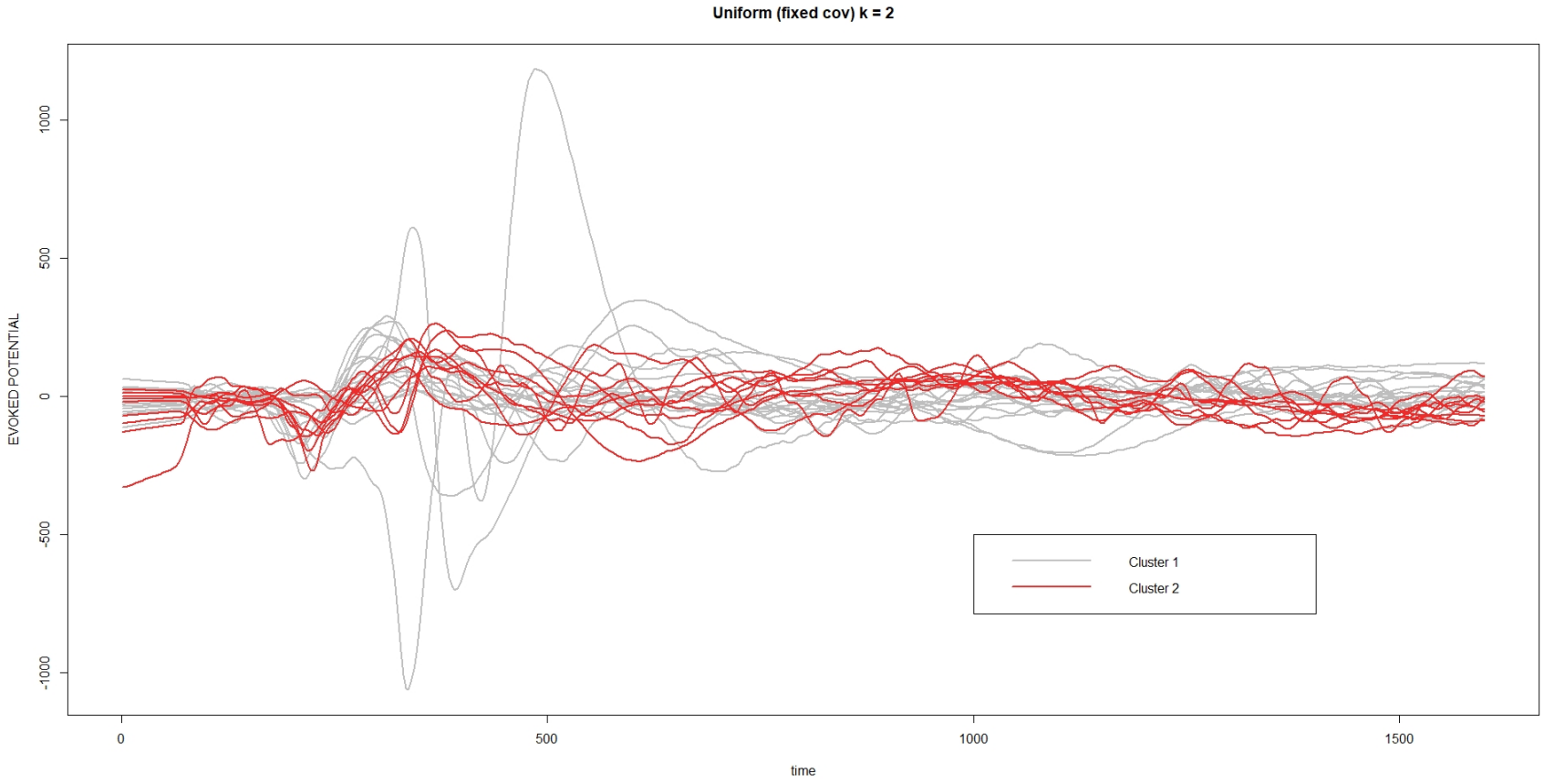
Clinical data

- Somatosensory Evoked Potential (SEP)
- $n = 26$ multivariate functional observations
- 4 electrodes (Short Latency SEP)
- 1600 data points (keeping 1 out of 3)
- Best value of $\alpha = 10^4$

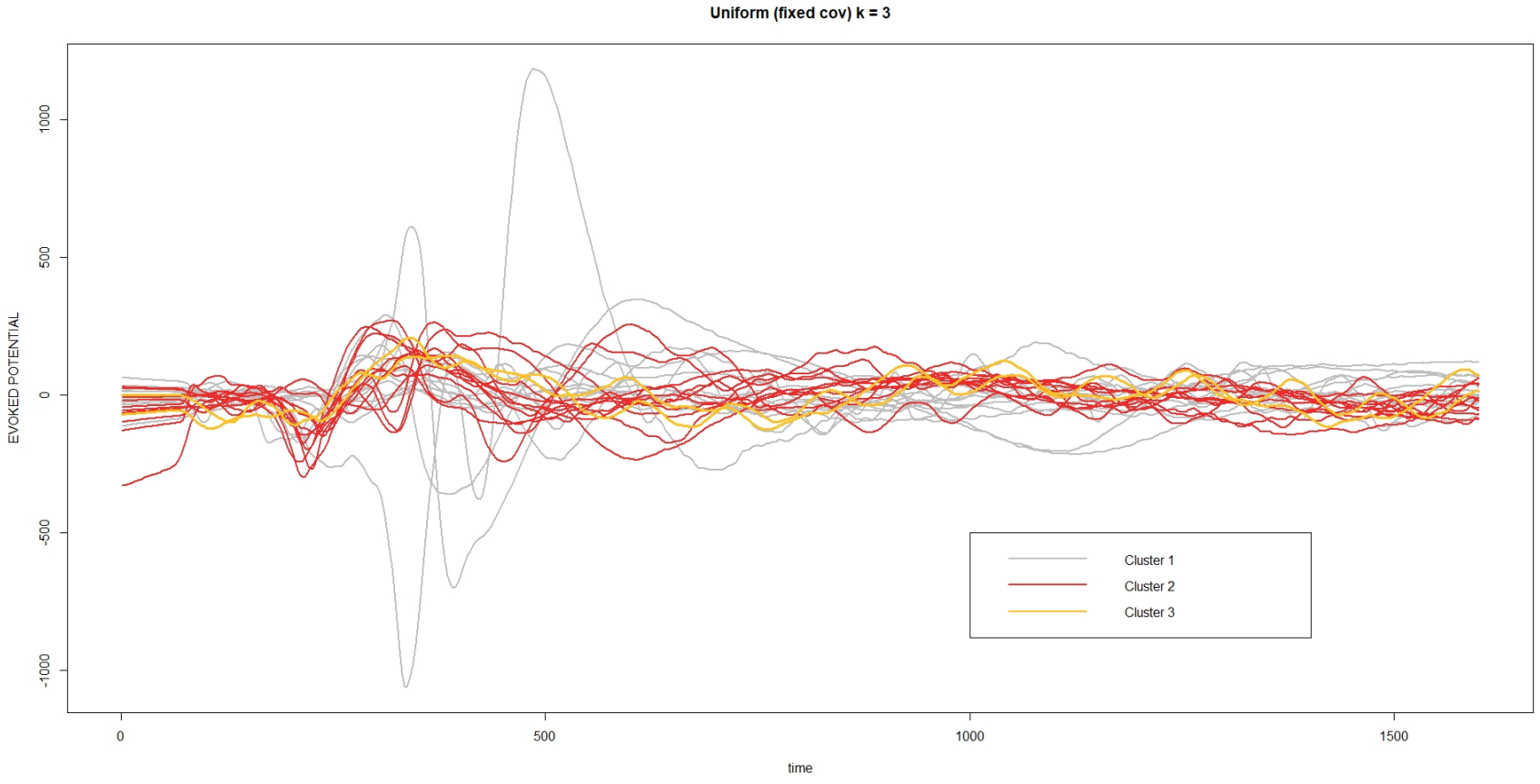




Uniform prior with fixed covariance and K=2

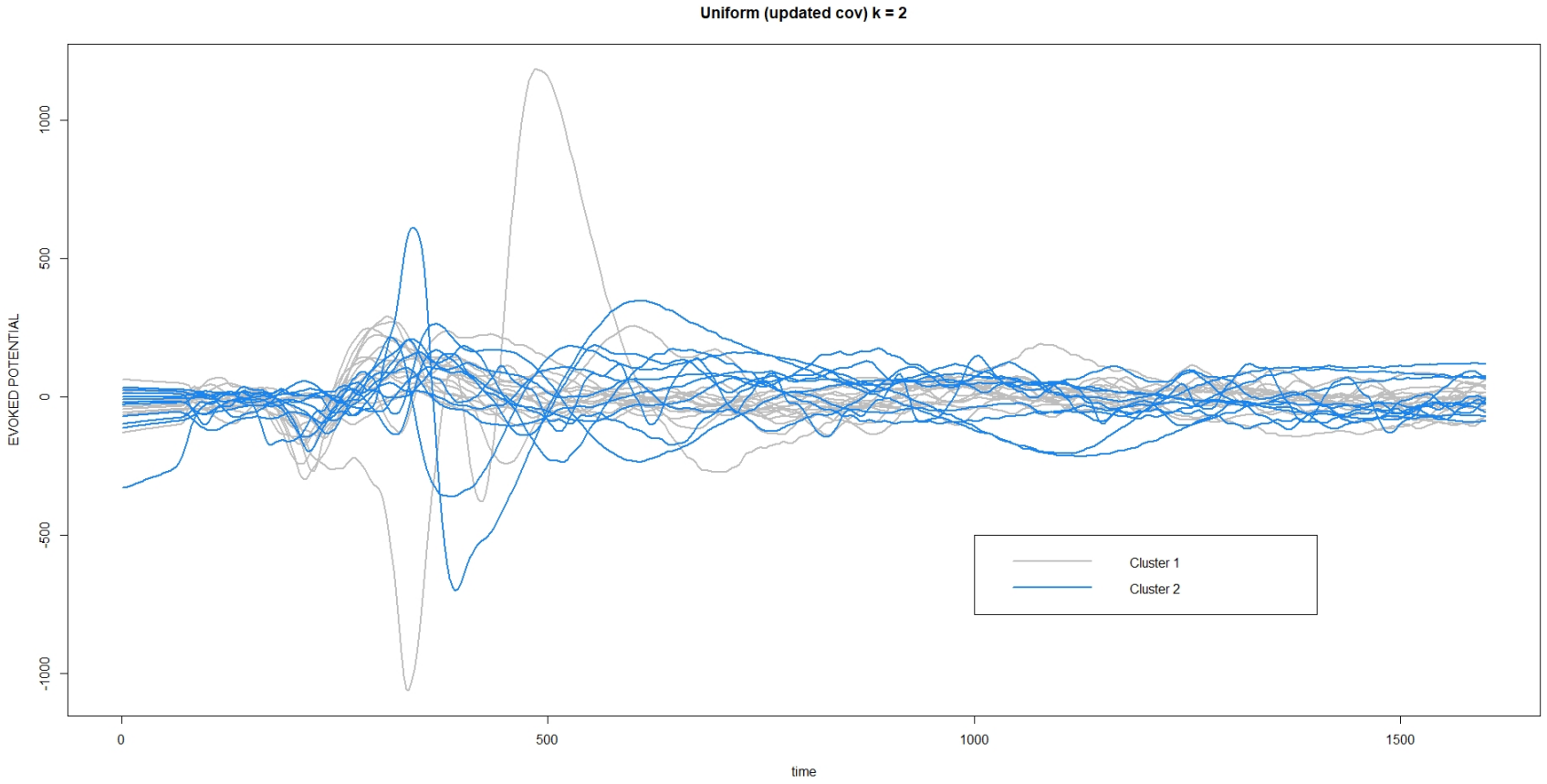


Uniform prior with fixed covariance and K=3

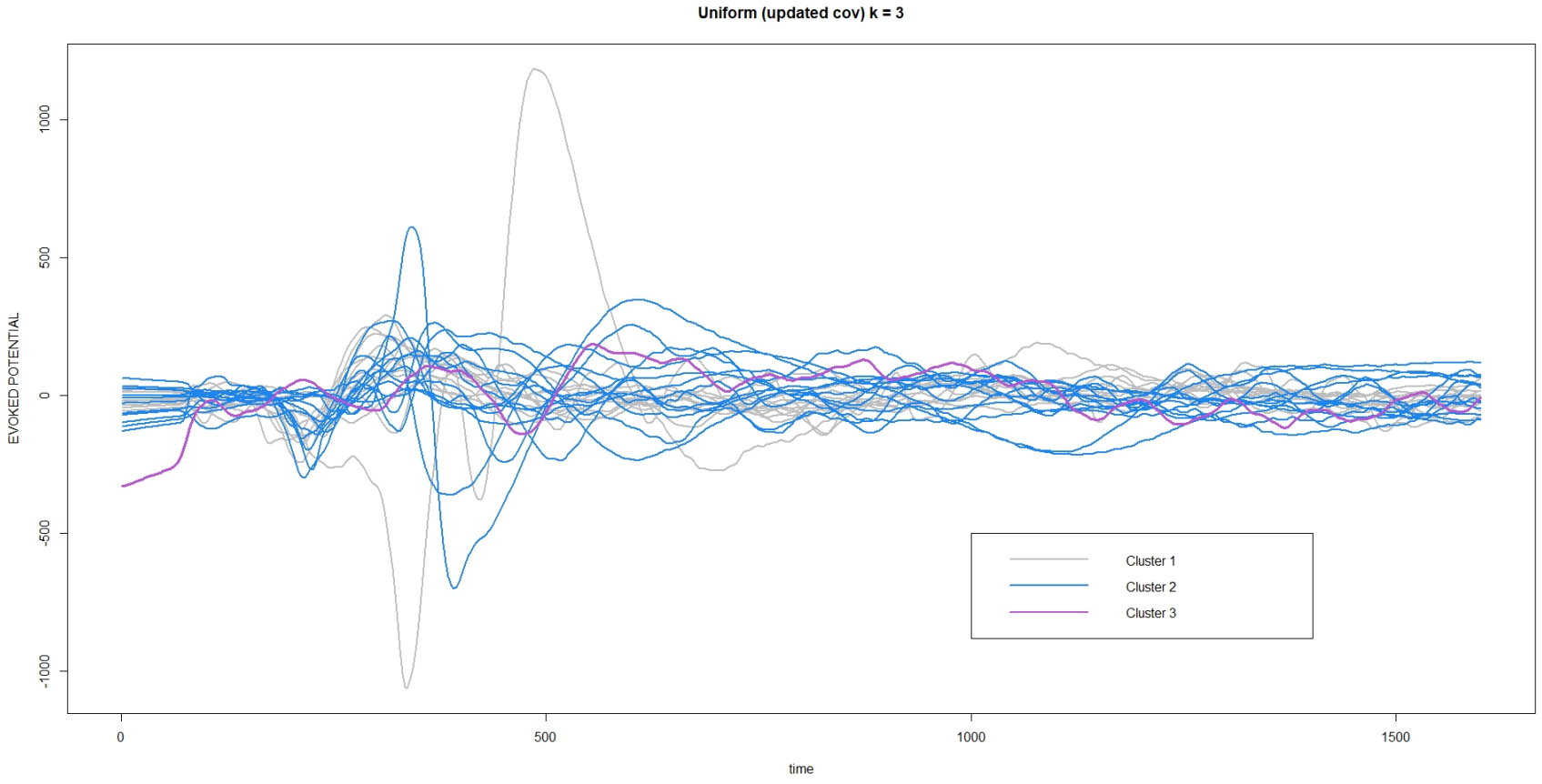


Generation of a small cluster and minimal loss reduction

Uniform prior with updated covariance and K=2

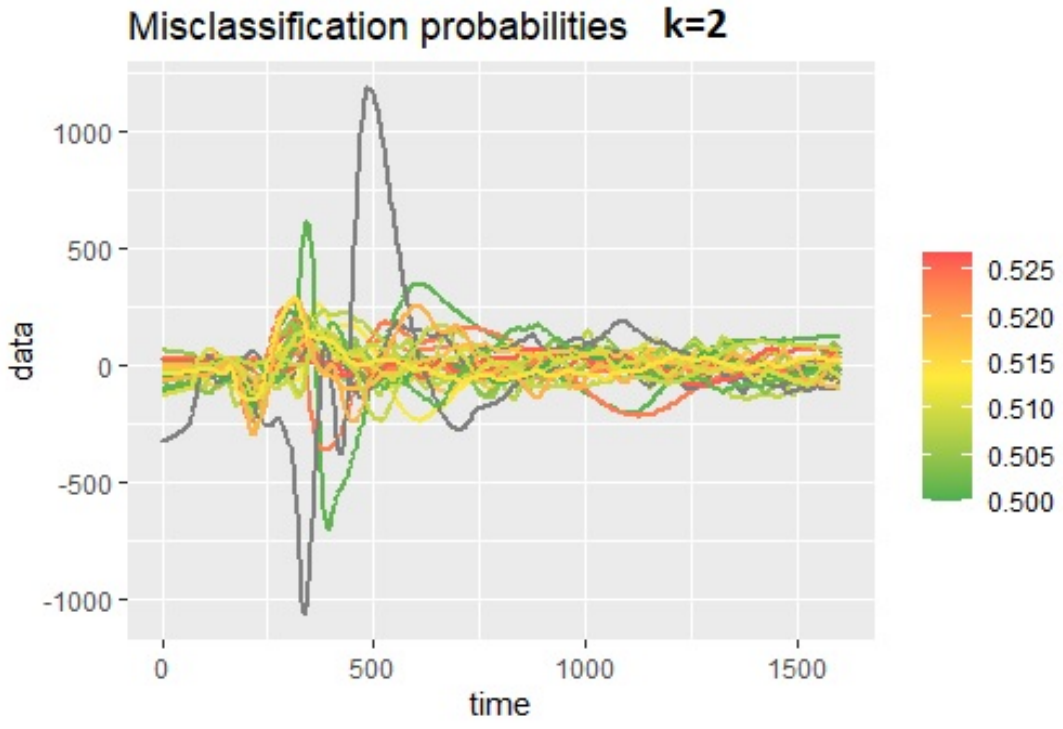


Uniform prior with updated covariance and K=3

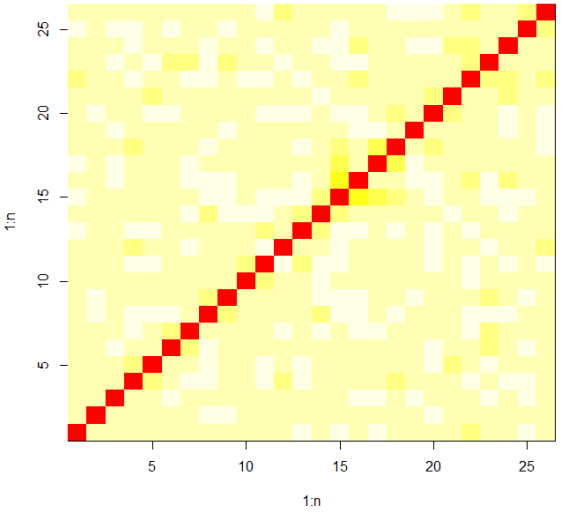


Generation of single unit cluster and extremely minimal loss reduction

Gibbs sampling results on clinical data

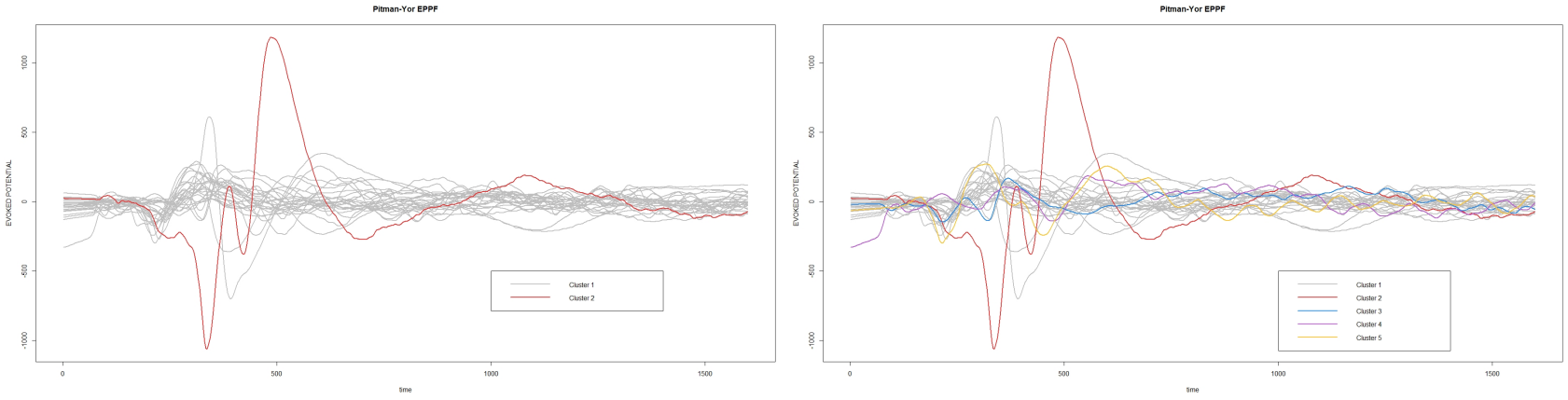


(a) Misclassification probability



(b) Similarity matrix

Clustering with PY-EPPF prior



(a) $\sigma = 0.75$ **(b)** $\sigma = 0.5$

Generation of single unit clusters depending on the value of the discount parameter σ

Conclusions

Considerations about algorithms and real data:

- Best model: Uniform prior model with updating covariance (K = 2).
- None of the models showed correlation with the clinical evaluations of patients.

Possible solutions/developments:

- Consider a functional distance not so highly influenced by the covariance structure of the data (for the clinical data)
- Modify the likelihood to accommodate the case $N_k = 1$ and/or introduce a penalization term for small clusters

References

References

[1] José R. Berrendero, Beatriz Bueno-Larraz, and Antonio Cuevas. "On Mahalanobis Distance in Functional Settings". In: *Journal of Machine Learning Research* 21 (2020), pp. 1–33.

[2] Antonio Canale et al. "On the Pitman–Yor process with spike and slab prior specification". In: (2017).

[3] Markus Herdin et al. "Correlation Matrix Distance, a Meaningful Measure for Evaluation of Non-Stationary MIMO Channels". In: (2005), p. 2.

[4] Tommaso Rigon, Amy H. Herring, and David B. Dunson. "A generalized Bayes framework for probabilistic clustering". In: *arXiv:2006.05451* (2020).

[5] Sara Wade and Zoubin Ghahramani. "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)". In: *Bayesian Anal.* 13 (2) (2018), pp. 559–626.

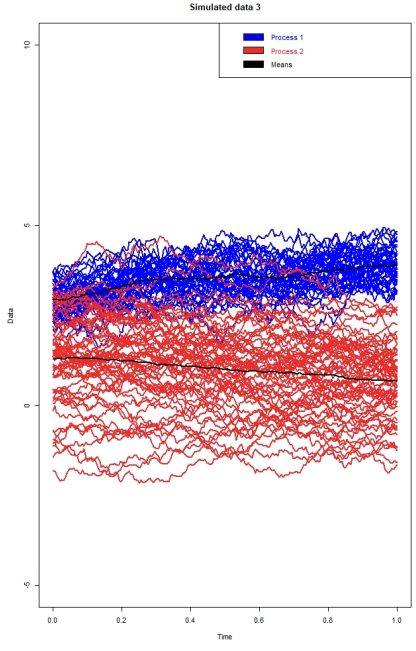
Challenges

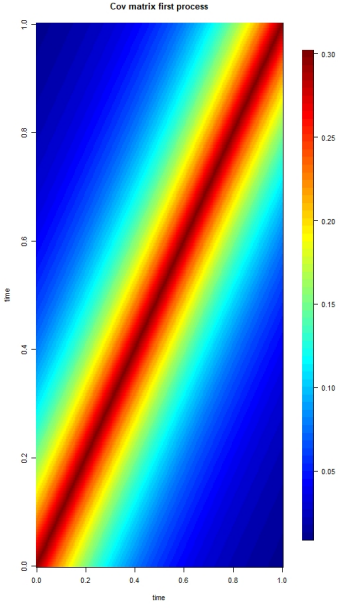
Simulated Data 3

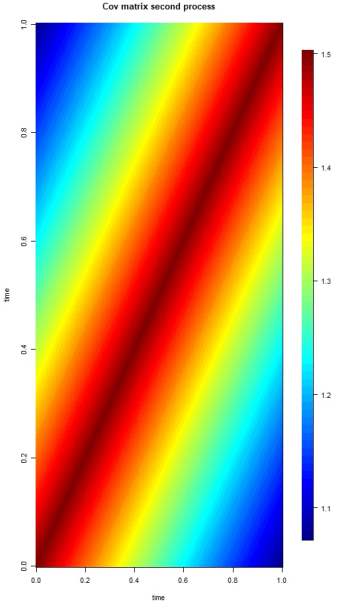
Data clustered in two groups:

- First process: $X(t) = sin(t) + 3 + \epsilon_1(t)$
  with $\epsilon_1(t) \sim \text{GP}(0, \mathcal{C}_1)$ $\quad \mathcal{C}_1(s,t) = 0.3 \cdot exp(-|s-t|/0.3)$ $\quad t \in [0,1]$
- Second process: $X(t) = cos(t) + \epsilon_2(t)$
  with $\epsilon_2(t) \sim \text{GP}(0, \mathcal{C}_1)$ $\quad \mathcal{C}_1(s,t) = 1.5 \cdot exp(-|s-t|/3)$ $\quad t \in [0,1]$

Characteristics: similar means and different covariance matrices in GP

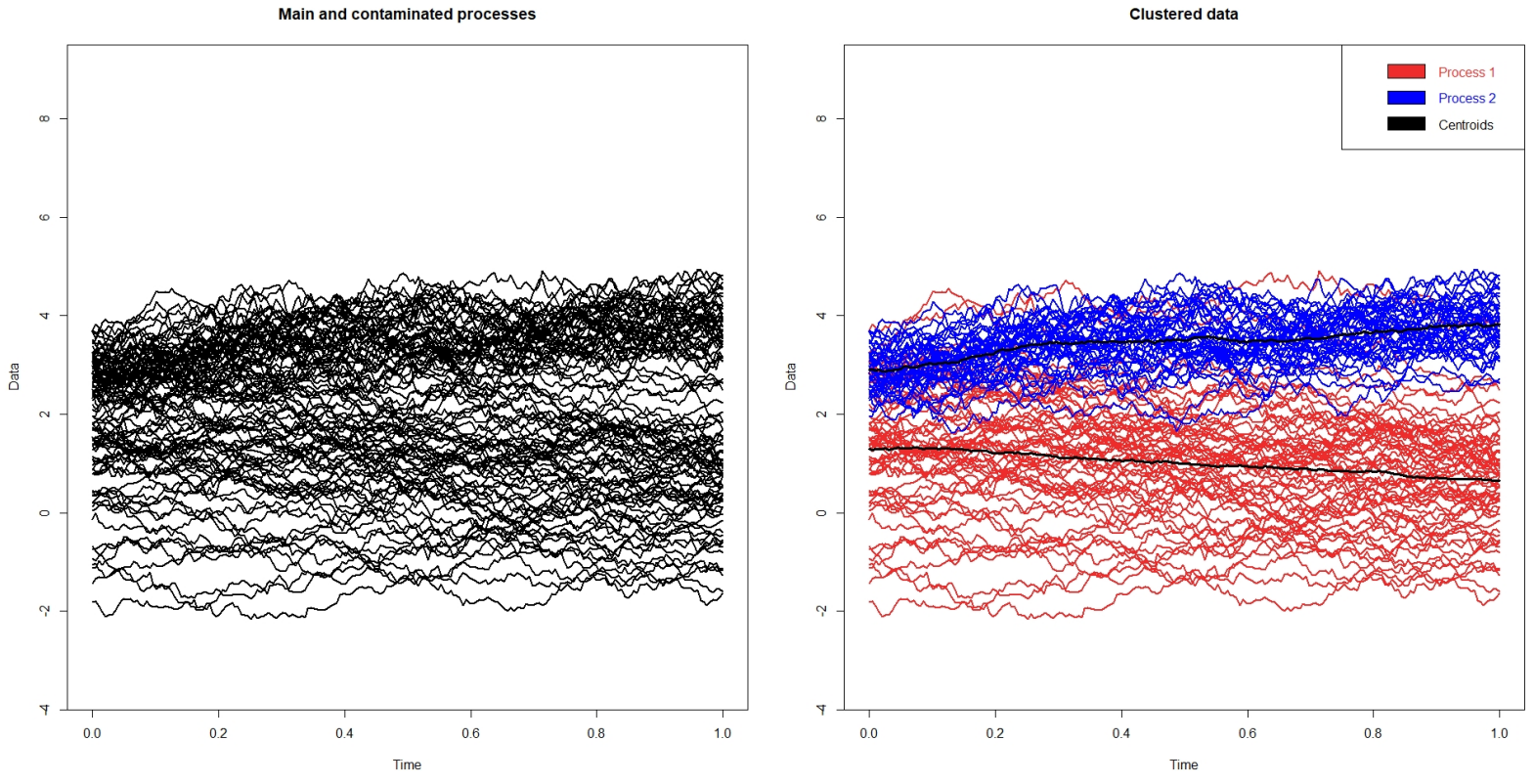






Clustering on Simulated data 3

To test the effectiveness of "Mahalanobis distance clustering algorithm with uniform prior and covariance updating" in this situation:



4 observations up to 100 have been wrongly classified: best result after lots of simulations

Union of similar clusters

Idea: Merge clusters if both centroids distance and covariance clusters distance are small

Algorithm 4: Union of similar clusters

Input : Optimal centroids $\mathbf{m}_1, ..., \mathbf{m}_k$

Output: United centroids where needed

Compute centroid distances matrix d : $d_{ij} = \|\mathbf{m}_i - \mathbf{m}_j\|_2^2 \quad \forall i, j$;

Set $\varepsilon = 0.5 \cdot (\text{median}(d) + \text{mean}(d))$;

if $d_{ij} < \varepsilon$ **then**

  Compute cluster covariances distance: $d_{cov} = 1 - \frac{\text{tr}\{C_i C_j\}}{||C_i||_f ||C_j||_f}$;

  if $d_{cov} < 0.05$ **then**

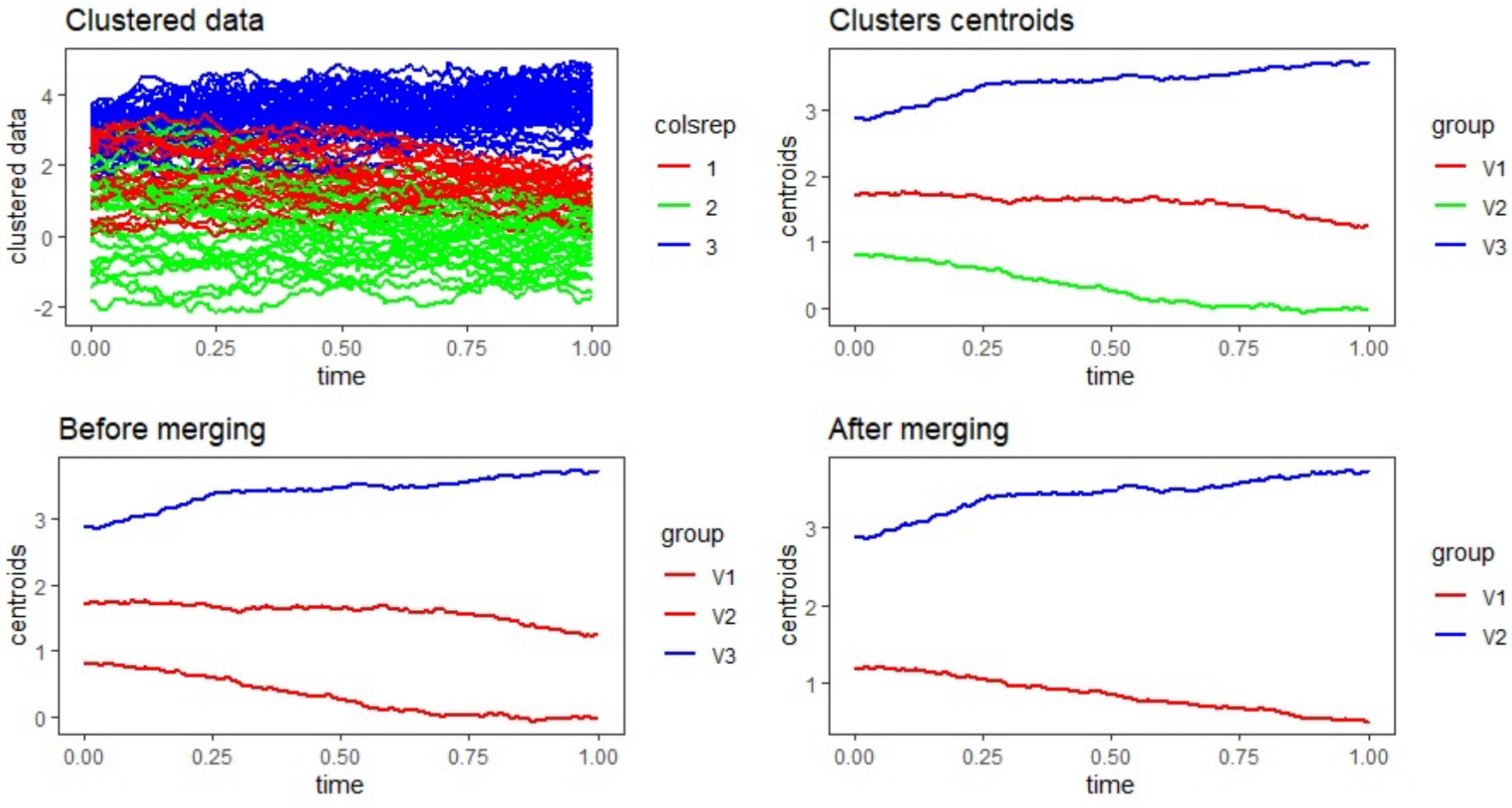
    Merge clusters i, j;

  end

end

Recompute the centroids and repeat until $d_{i_{new} j_{new}} > \varepsilon_{new} \quad \forall i_{new}, j_{new}$

Application on Simulated Data 3



Result: only 5 observations are misclassified