Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

# Learning Mixtures of Gaussian Processes through Random Projection

**Emmanuel O. Akeweje & Mimi Zhang**
School of Computer Science and Statistics

# Functional Data

# and

# Functional Data Cluster Analysis

# Examples of Functional Data

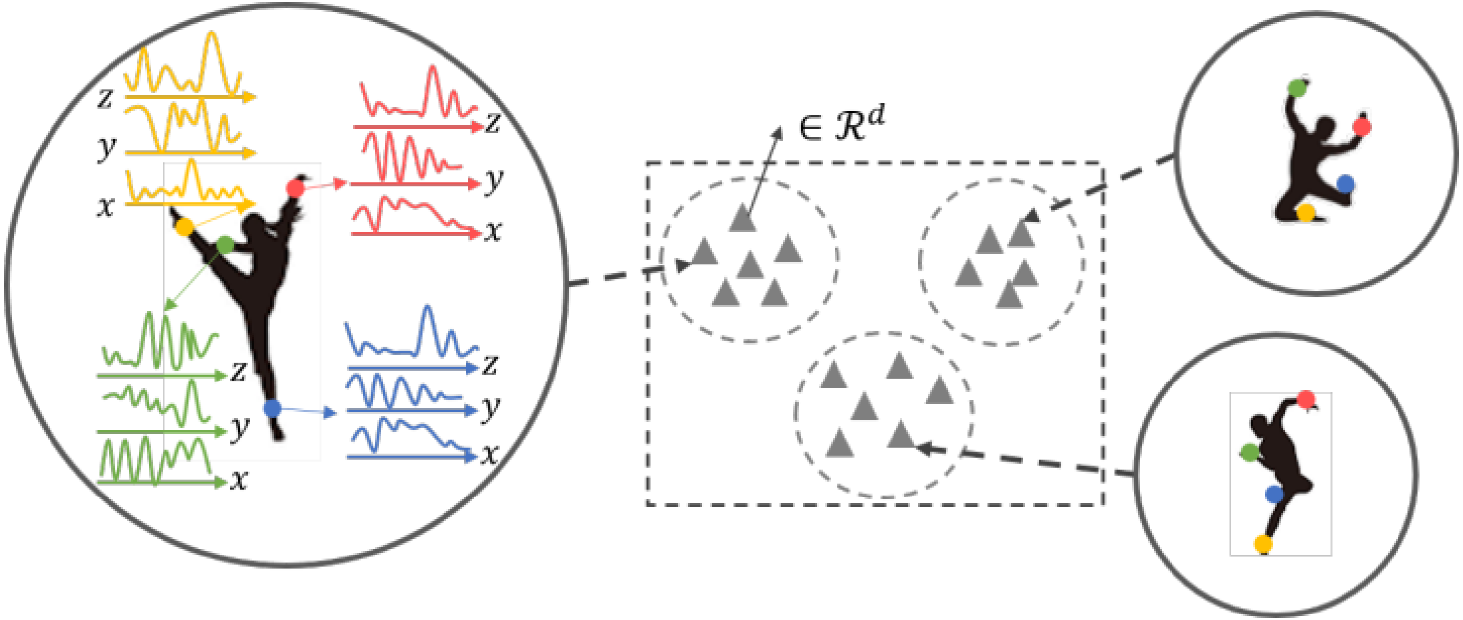



Image source [2]

# Examples of Functional Data

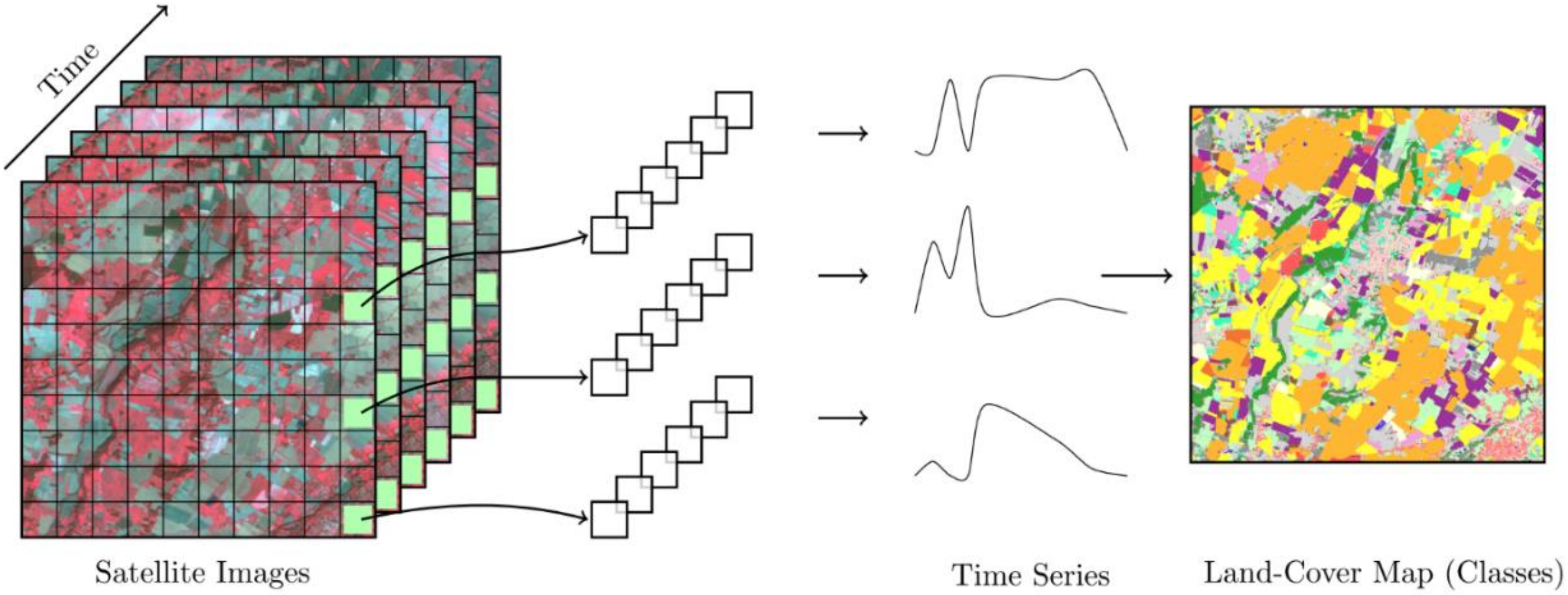



Image source [3]

# Examples of Functional Data

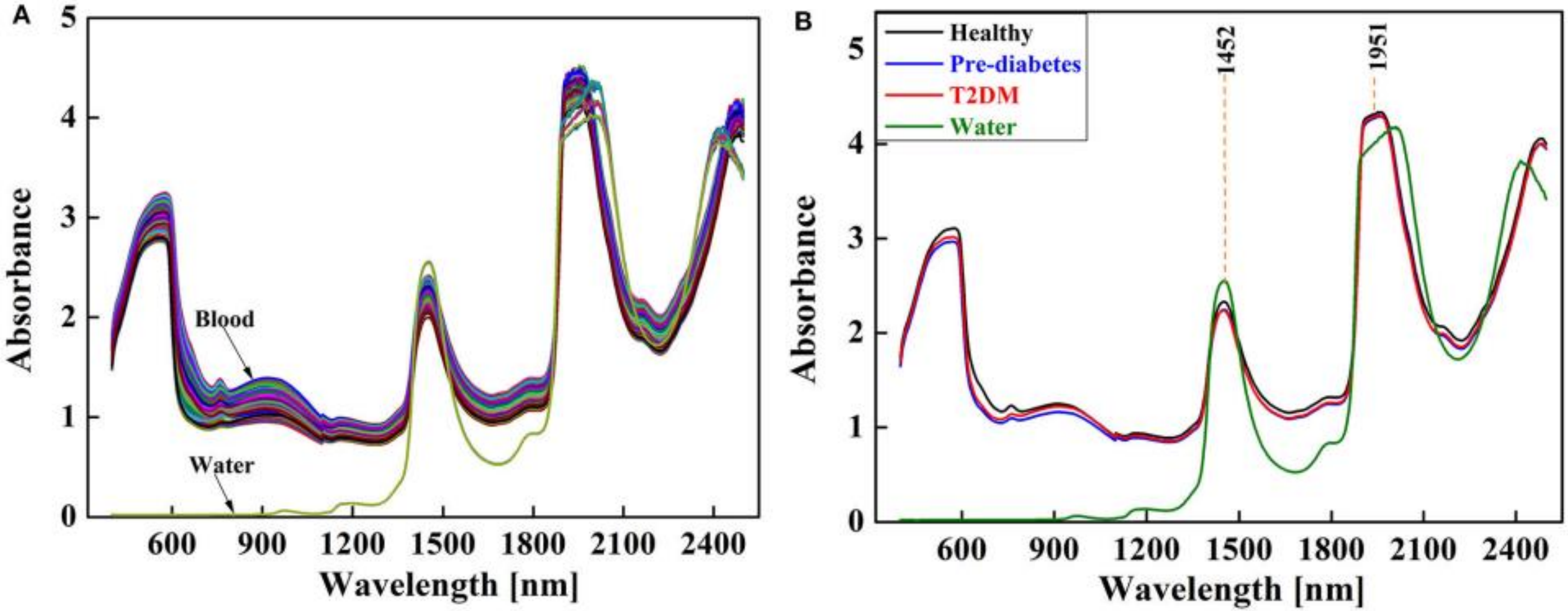


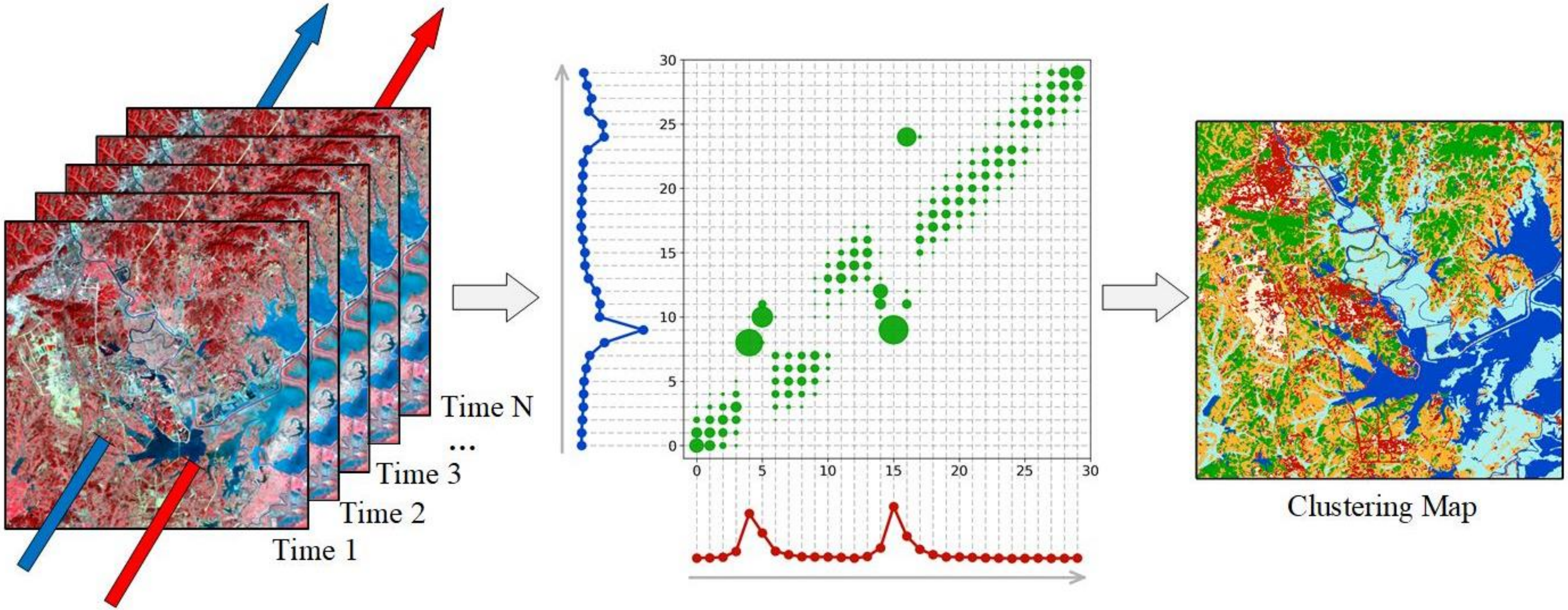

Image source [4]

# Reasons for Considering Functional Data

**Functional data analysis is about curves, surfaces or anything varying over a continuum.**

- **It is more natural to think through modelling problems in a functional form.**
- **The functional form informs us the values of f(t) for t at nearby locations, its derivatives, resilient to noise contamination.**
- **The focus is on analysing relations among the random elements, rather than properties of individual random elements.**

# Examples of Functional Data Clustering




**land cover mapping**

Image source [5]

# Examples of Functional Data Clustering

RNA-seq data were generated from a human cell line at 1, 3, 5, 7, 9, and 11 hours after treatment with the synthetic gluco-corticoid (GC) dex.

Clustering methods partition time series gene expression data into disjoint clusters based on the similarity of expression response.

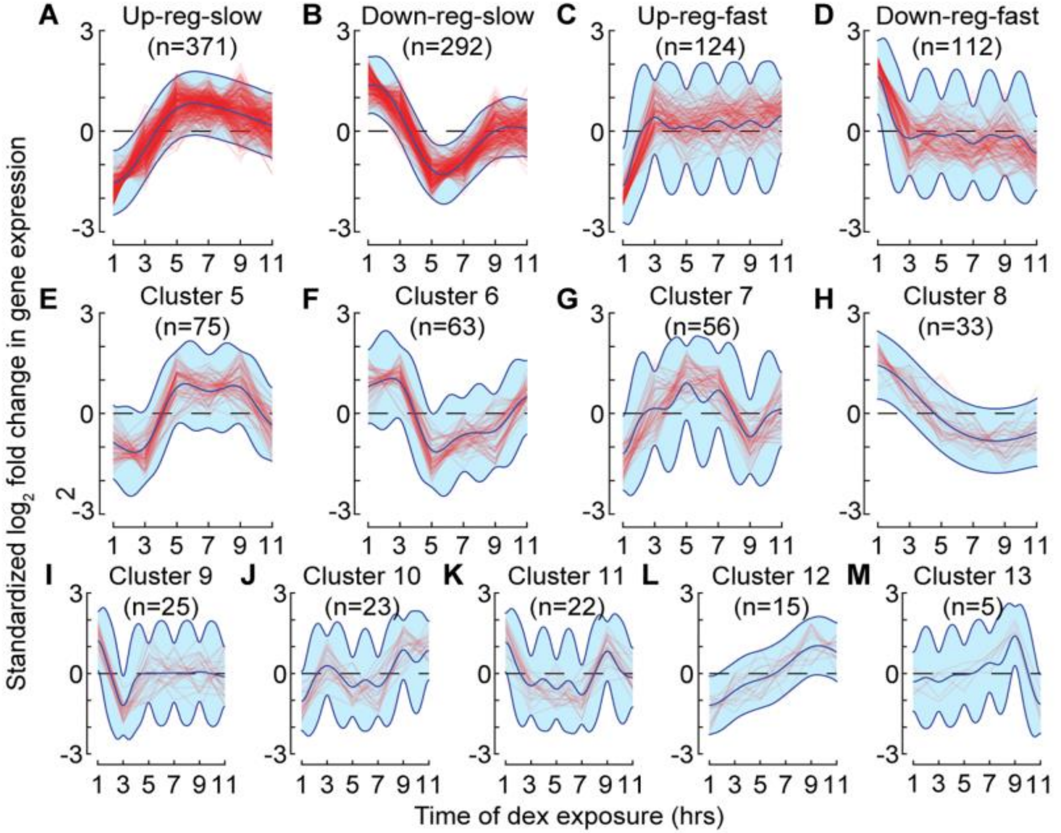
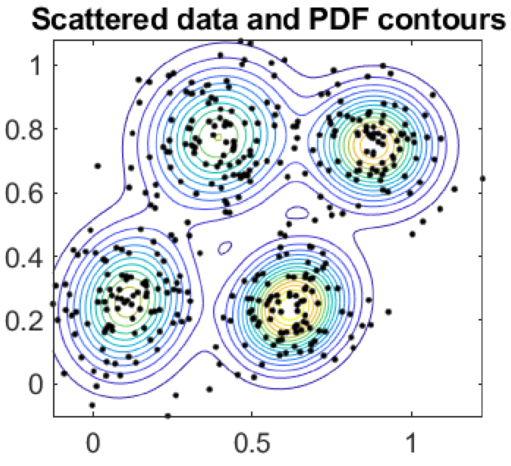
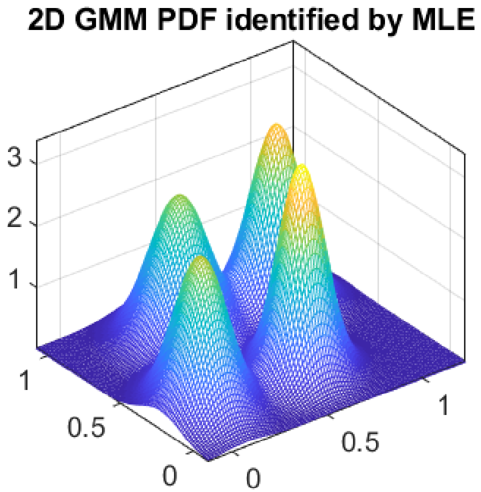



Image source [6]

# Gaussian Process (GP) Mixture

# Gaussian Mixture Model (GMM)




$$X \sim \sum_{k=1}^{K} \pi_k N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,.$$

Clustering is done by assigning each $\boldsymbol{x}_i$ to the mixture component (i.e., cluster) to which it is most likely to belong a posteriori.

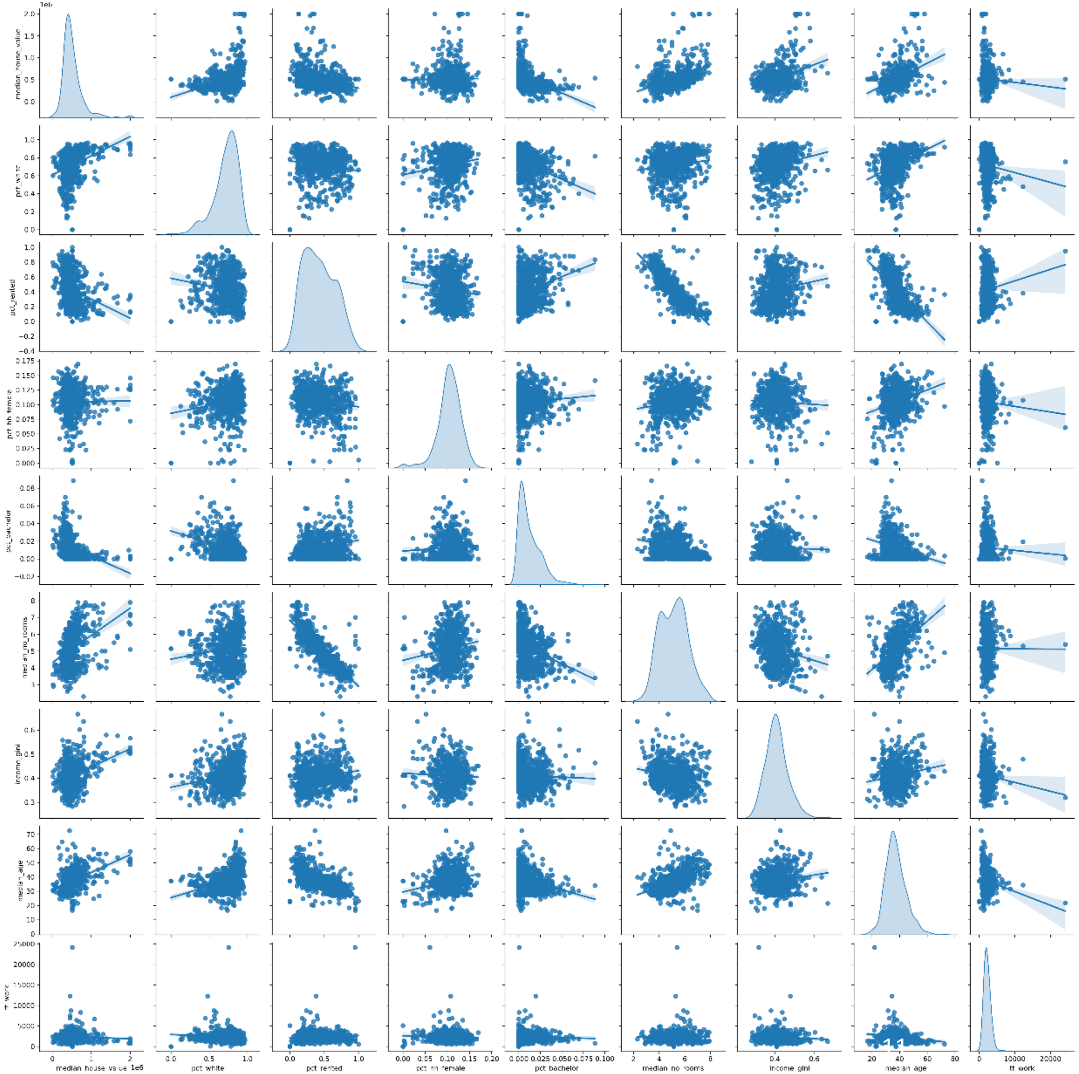
# High Dimension

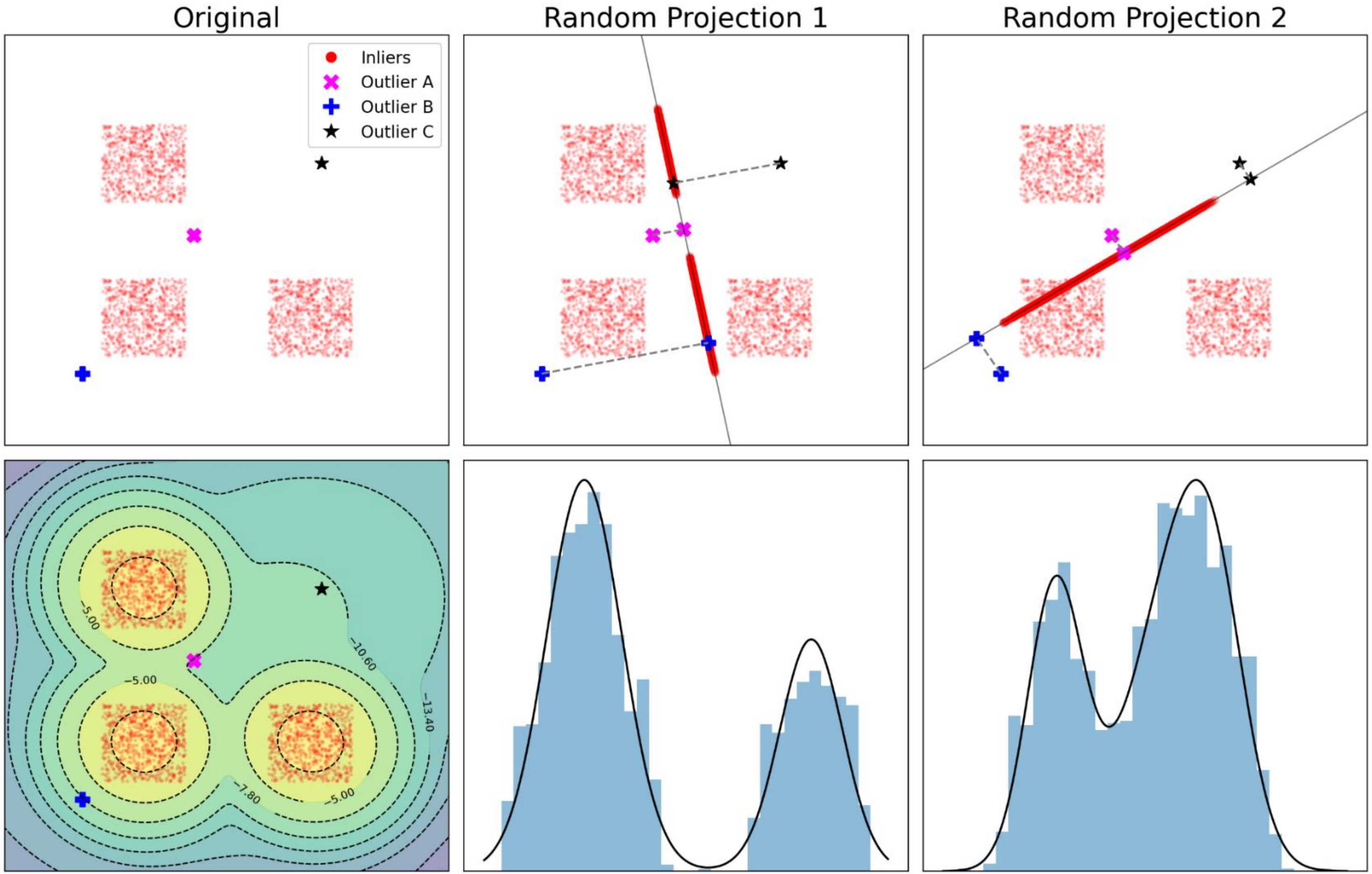
**Data Sparsity**:

- When dimensionality increases, data points that were close together in lower dimensions become increasingly separated.

**Distance Metric Problems**:

- Most pairs of points become nearly equidistant from each other and from a reference point.





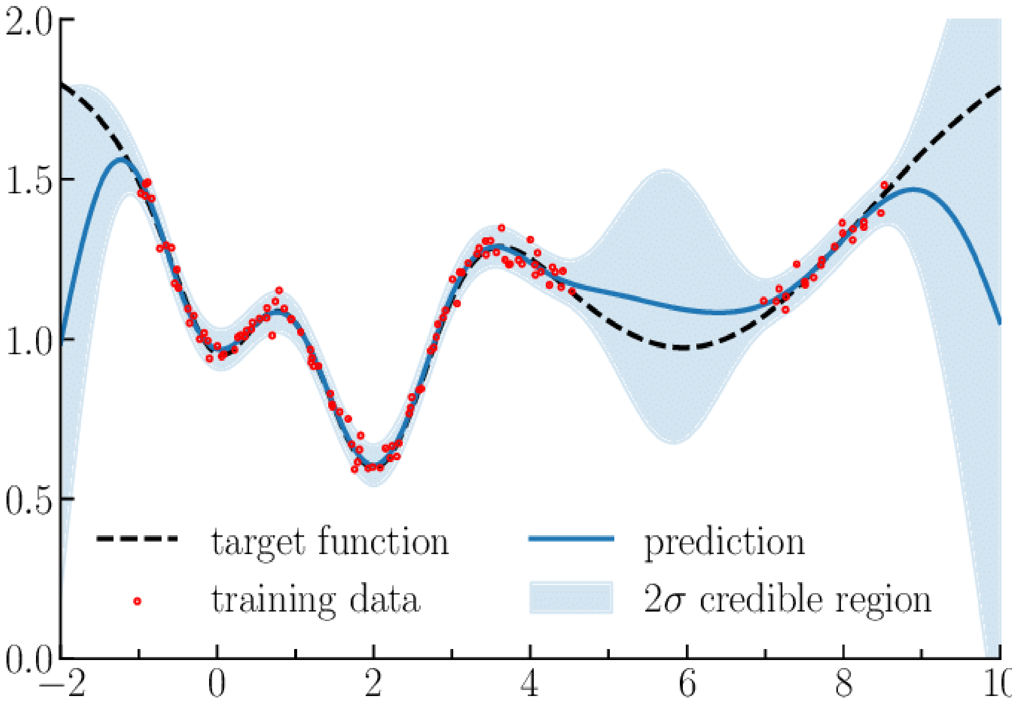
Original
Random Projection 1
Random Projection 2
Inliers
Outlier A
Outlier B
Outlier C
-5.00
-5.00
-10.60
-13.40
-7.80
-5.00

# Gaussian Process (GP)

A Gaussian process $GP(\mu, \Sigma)$ is a stochastic process:

$X \sim GP(\mu, \Sigma)$, then $\forall\, \boldsymbol{t} = (t_1, \ldots, t_m)^T$,

$$X(\boldsymbol{t}) \sim N\big(\mu(\boldsymbol{t}), \Sigma(\boldsymbol{t}, \boldsymbol{t})\big).$$



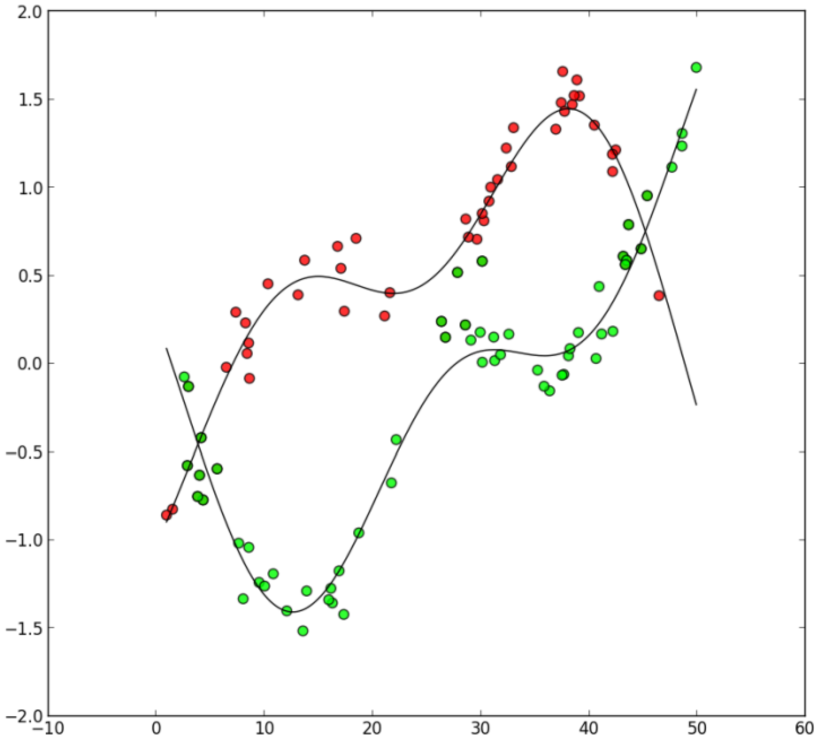
# Gaussian Process (GP) Mixture

$\{(x_i, z_i): i = 1, \dots, n\} \sim_{iid} (X, Z)$

$\Pr(Z = k) = \pi_k$

$[X|Z = k] = X_k \sim GP(\mu_k, \Sigma_k)$

$x_i(\boldsymbol{t}) \sim \sum_{k=1}^{K} \pi_k N\left(\mu_k(\boldsymbol{t}), \Sigma_k(\boldsymbol{t}, \boldsymbol{t})\right)$



# From

# GP Mixture

# to

# Univariate Gaussian Mixture Model

# Projecting onto One Dimension

**Gaussian Random Variable**

$[X|Z = k] = X_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$:

$$\boldsymbol{\Sigma}_k = \sum_{r=1}^{\infty} \lambda_{kr} \boldsymbol{b}_{kr} \boldsymbol{b}_{kr}^T .$$

$\forall\, \boldsymbol{y} \in \mathbb{R}^p$,

$$\langle X_k, \boldsymbol{y} \rangle \sim N\left( \langle \boldsymbol{\mu}_k, \boldsymbol{y} \rangle, \sum_{r=1}^{\infty} \lambda_{kr} \langle \boldsymbol{b}_{kr}, \boldsymbol{y} \rangle^2 \right).$$

$\langle X, \boldsymbol{y} \rangle \sim \sum_{k=1}^{K} \pi_k N(\langle \boldsymbol{\mu}_k, \boldsymbol{y} \rangle, \sum_{r=1}^{\infty} \lambda_{kr} \langle \boldsymbol{b}_{kr}, \boldsymbol{y} \rangle^2)$

**Gaussian Random Function**

$[X|Z = k] = X_k \sim GP(\mu_k, \Sigma_k)$:

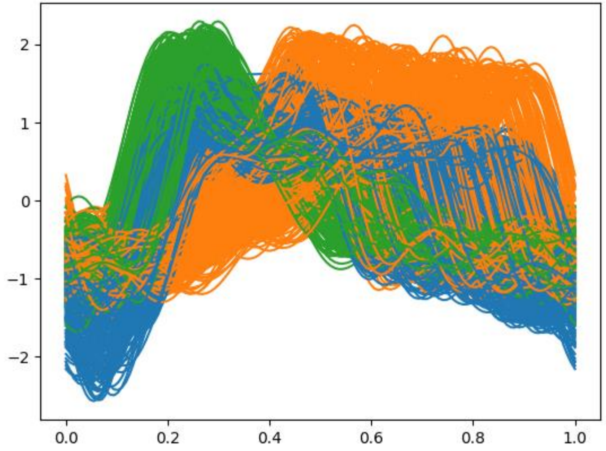
$$\Sigma_k(s,t) = \sum_{r=1}^{\infty} \lambda_{kr} b_{kr}(s) b_{kr}(t) .$$

$\forall\, y \in \mathcal{H}(T, \mathbb{R})$,

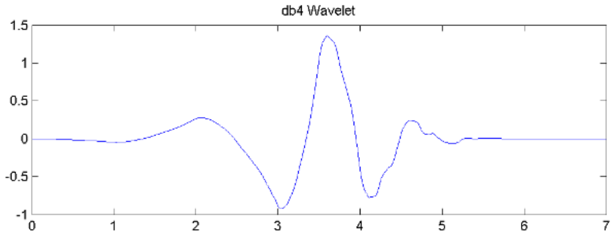
$$\langle X_k, y \rangle \sim N\left( \langle \mu_k, y \rangle, \sum_{r=1}^{\infty} \lambda_{kr} \langle b_{kr}, y \rangle^2 \right).$$

$\langle X, y \rangle \sim \sum_{k=1}^{K} \pi_k N(\langle \mu_k, y \rangle, \sum_{r=1}^{\infty} \lambda_{kr} \langle b_{kr}, y \rangle^2)$

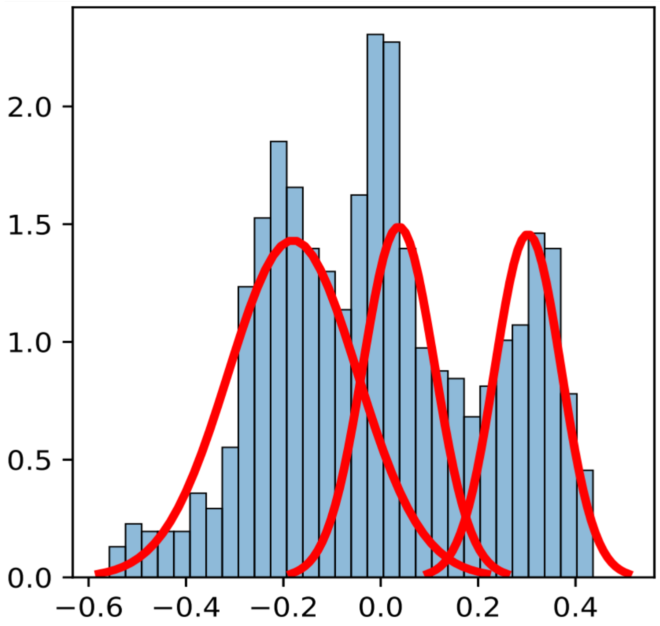
# Random Projection



$\{x_i\}_{i=1}^{n}$




$y$



$\{\langle x_i, y\rangle\}_{i=1}^{n}$

# How to generate the projection function $y$?

**Fixed:**

- wavelets
- B-splines
- Fourier
- $\{b_1, \dots, b_m\}$ from $\Sigma(s,t) = \sum_{r=1}^{\infty} \lambda_r b_r(s) b_r(t)$

**Random:**

- Ornstein-Uhlenbeck process
- $y(t) = \sum_{r=1}^{m} a_r b_r(t)$ where $a_r \sim N(0, \lambda_r)$

# The GPmix Algorithm

**Input:** The raw data $\mathcal{D} = \{y_i(\underline{t}_i)\}_{i=1}^n$, the projection functions $\{\beta_v\}_{v=1}^V$, and the number of clusters $K$.
**Output:** The learned cluster labels $\{z_i\}_{i=1}^n$.

Estimate the population mean function $\mu$ and the $n$ sample functions $\{x_i\}_{i=1}^n$.

**for** $v = 1, \ldots, V$ **do**

Calculate the $n$ projection coefficients:

$$\alpha_{iv} = \langle x_i - \mu, \beta_v \rangle, \quad 1 \le i \le n.$$

Train a univariate GMM from the data $\{\alpha_{iv}\}_{i=1}^n$, denoted by $\sum_{k=1}^K \pi_{vk} \phi(\alpha; u_{vk}, \sigma_{vk}^2)$.

Obtain the cluster membership matrix $\mathbf{M}_v$:

$$m_{ik}^v = \frac{\pi_{vk} \phi(\alpha_{iv}; u_{vk}, \sigma_{vk}^2)}{\sum_{j=1}^K \pi_{vj} \phi(\alpha_{iv}; u_{vj}, \sigma_{vj}^2)}, \quad 1 \le i \le n, 1 \le k \le K.$$

Construct a binary membership indicator matrix $\mathbf{B}_v$:

$$b_{ik}^v = \begin{cases} 1, & \text{if } k = \arg\max_{1 \le j \le K} \{m_{ij}^v\}; \\ 0, & \text{otherwise.} \end{cases}$$

Calculate the weight $w_v (> 0)$: $\sum_{v=1}^V w_v = 1$.

**end for**

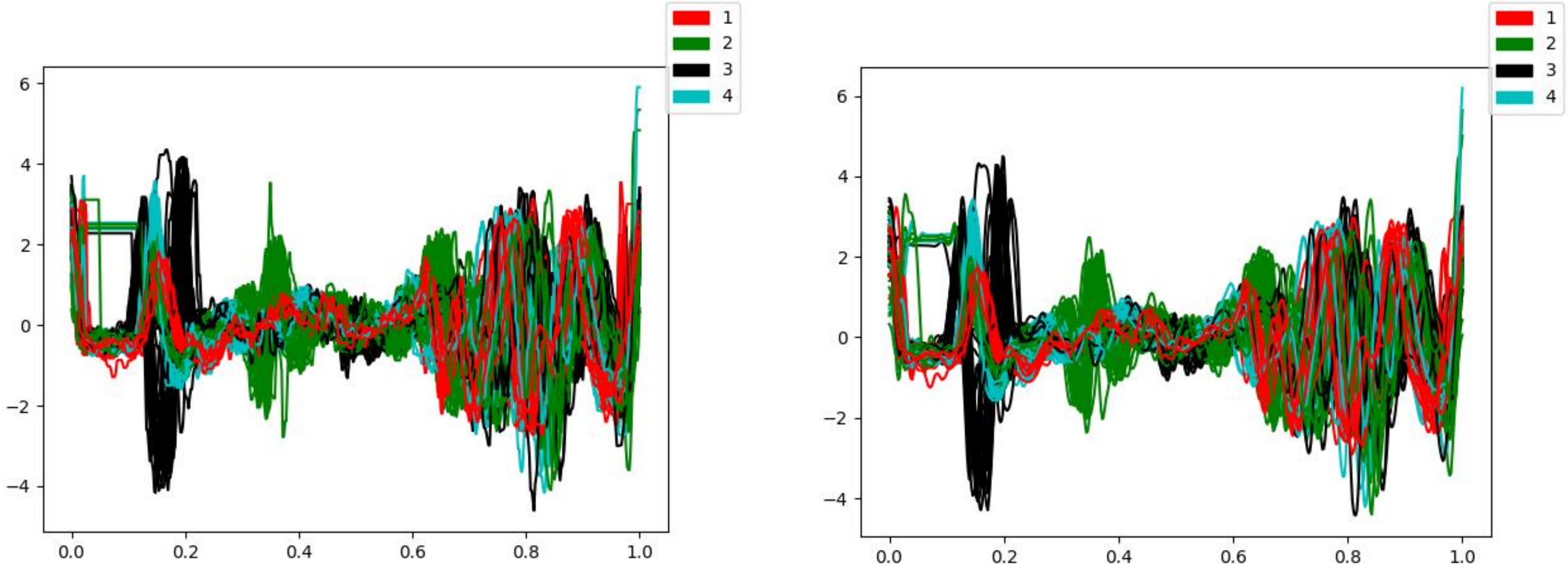
Apply a multivariate clustering method on the affinity matrix $\mathbf{A} = \sum_{v=1}^V w_v \mathbf{B}_v \mathbf{B}_v^T$ and return the identified cluster labels $\{z_i\}_{i=1}^n$.

from raw data to smooth functions

projecting functional data & learning a univariate GMM from the projection coefficients

extracting a consensus clustering from the multiple GMMs

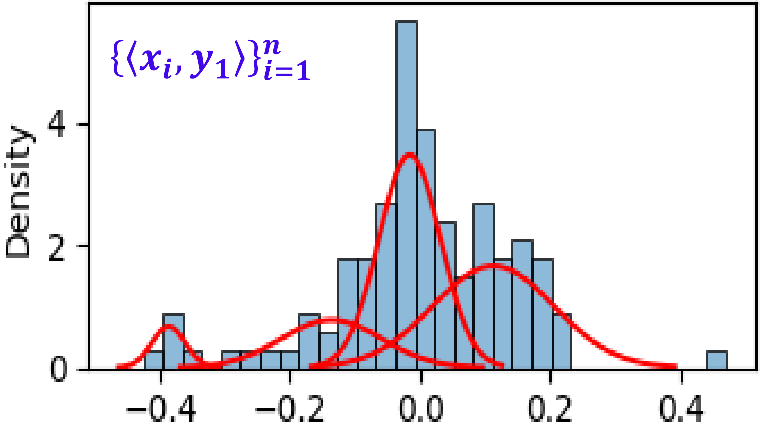
# Algorithm - Smooth




FaceFour from UEA & UCR Time Series Classification Repository.

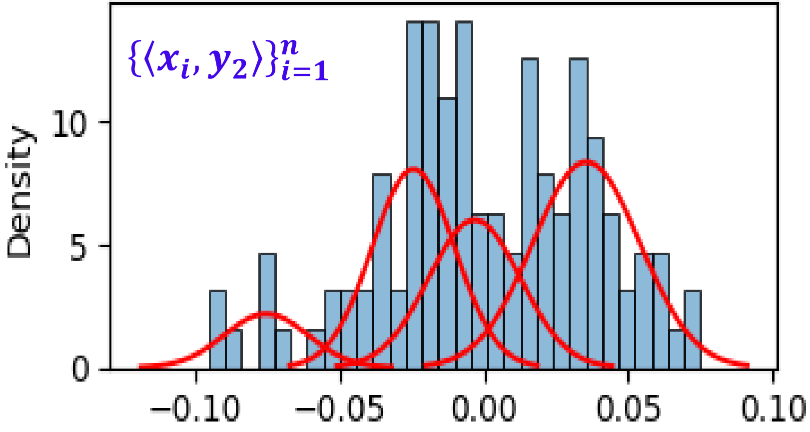
# Algorithm - Projection

Projection function $y_1$ (wavelet: sym17)




Univariate GMM $\sum_{k=1}^{4} \pi_k N\left(u_{1k}, \sigma_{1k}^2\right)$

Projection function $y_2$ (wavelet: sym17)



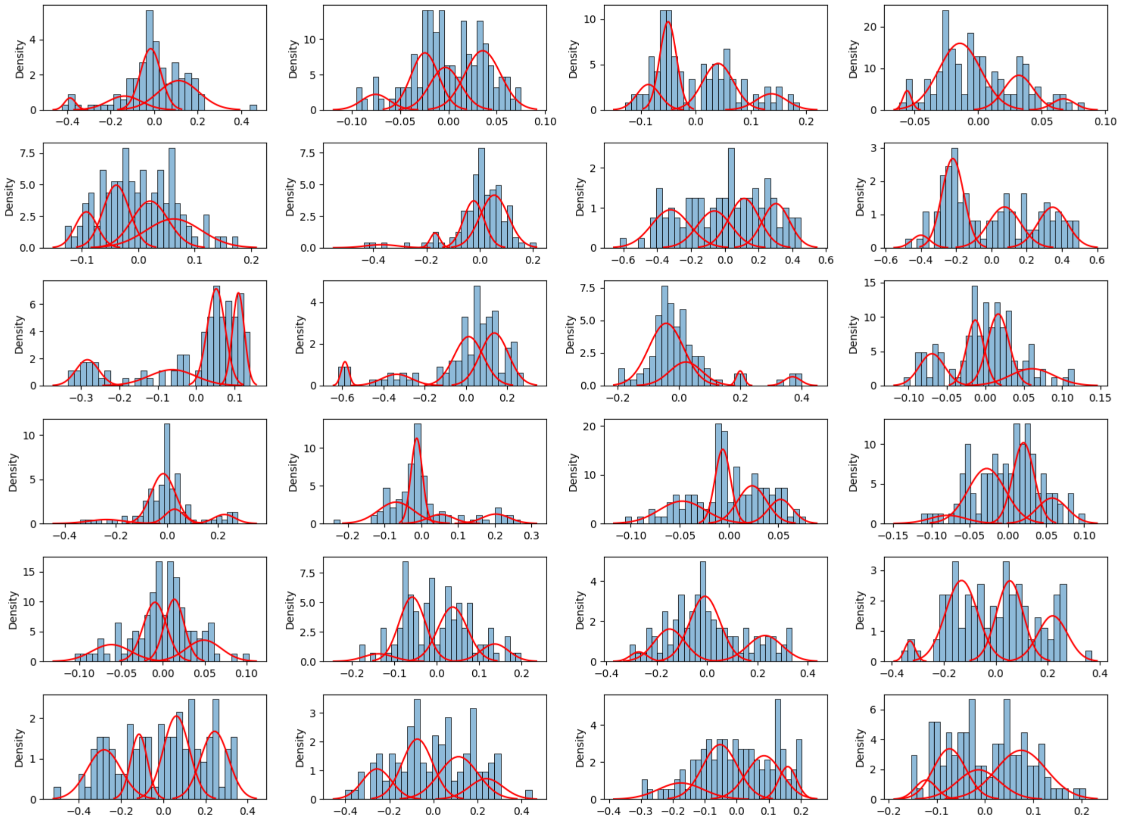

Univariate GMM $\sum_{k=1}^{4} \pi_k N\left(u_{2k}, \sigma_{2k}^2\right)$

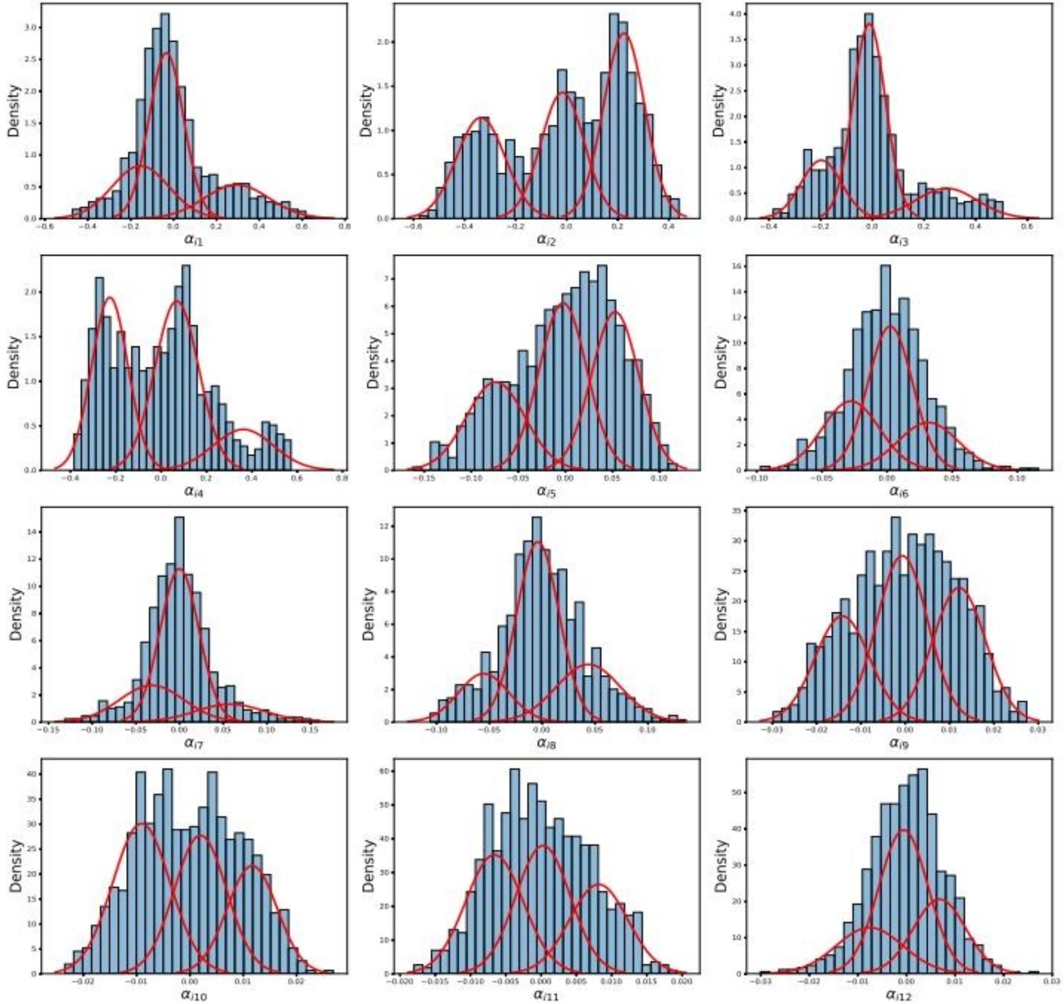
# Algorithm - Ensemble

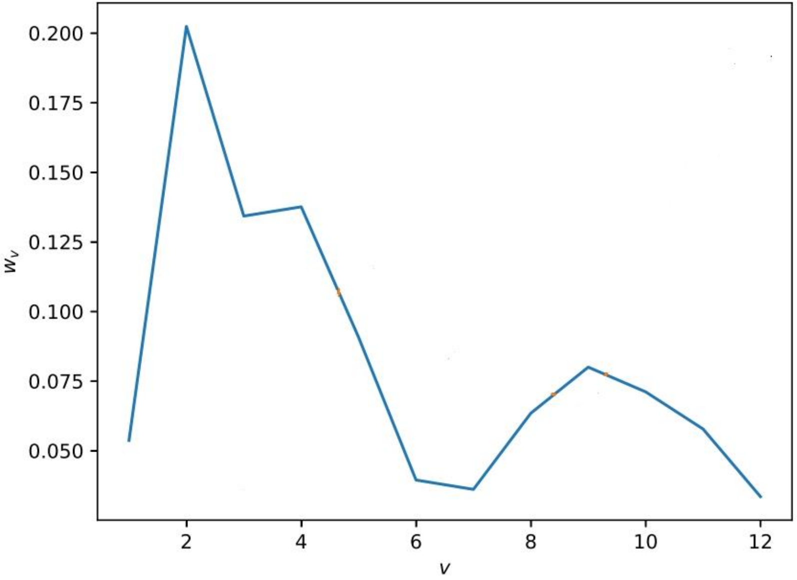
Perform cluster analysis on

$$A = \sum_{r=1}^{m} w_r B_r B_r^T,$$

- $B_r$ is the membership indicator matrix obtained from r-th GMM.
- $w_r$ is a data-driven weight on the r-th GMM.






**Base clustering weights $w_r$, calculated according to the overlapping degree of mixture components.**

# Theoretical Analysis

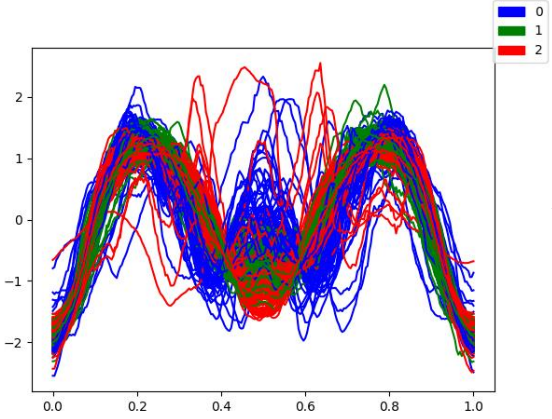
**Conditions for the identifiability of GP mixtures.**

**The probability that a 1-dimensional random projection achieves a separation of $\epsilon$ or higher among the mixture components.**

**Sample complexity and computational complexity of the learning problem are in every way polynomial.**

# Experimental Results

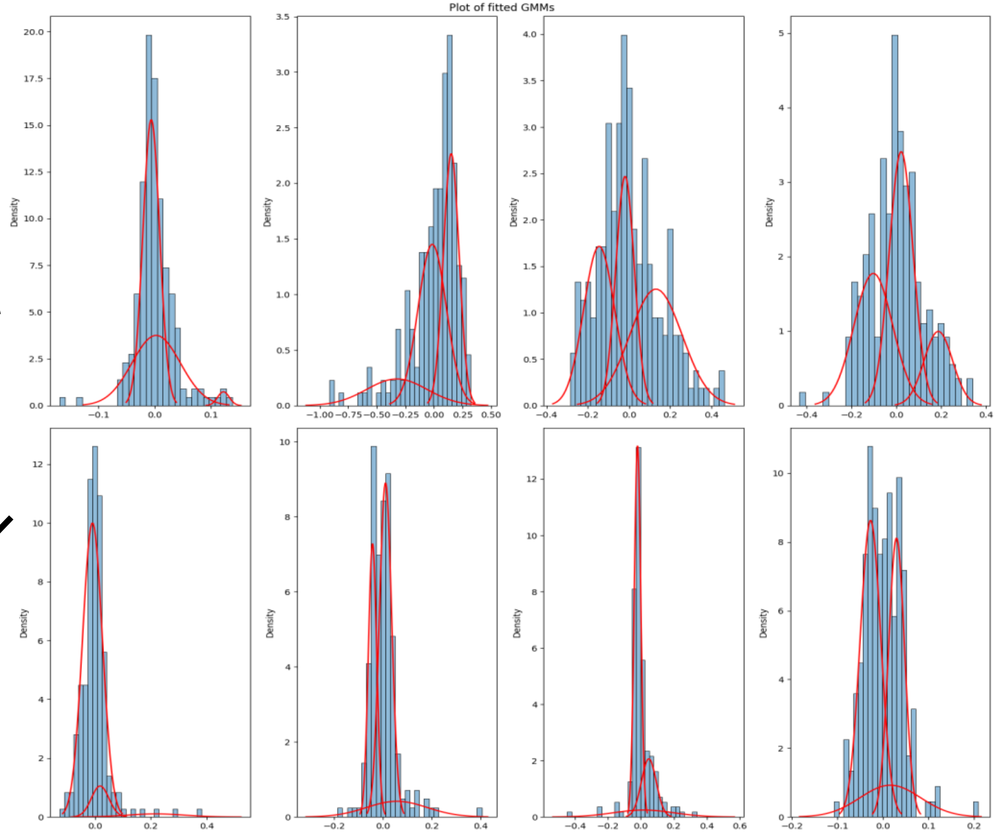
# Real Data Analysis

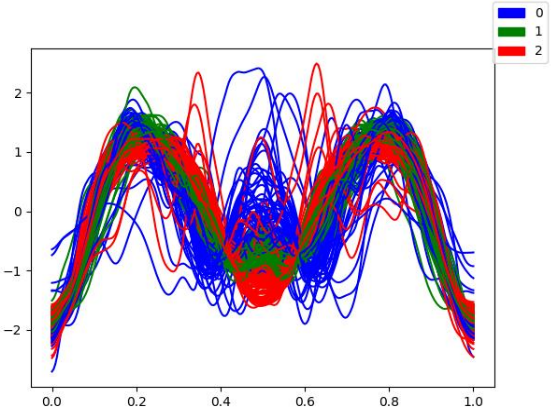



**ArrowHead**

**wavelet: db10**

**AMI: 0.37**
**ARI: 0.36**
**CCA: 0.67**



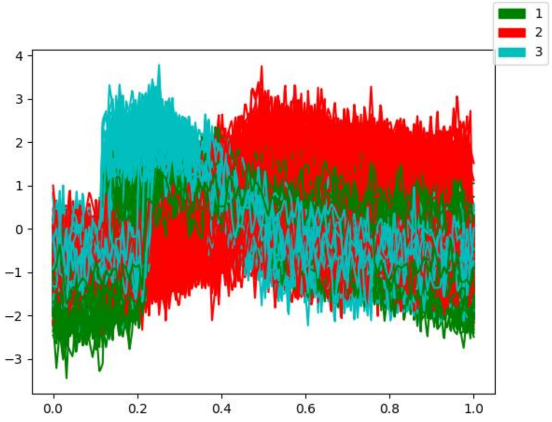



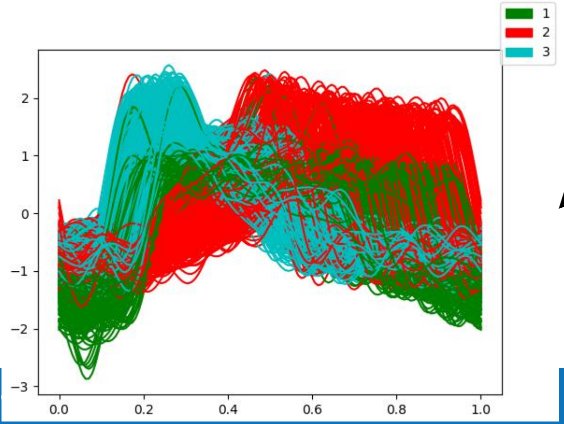
# Real Data Analysis

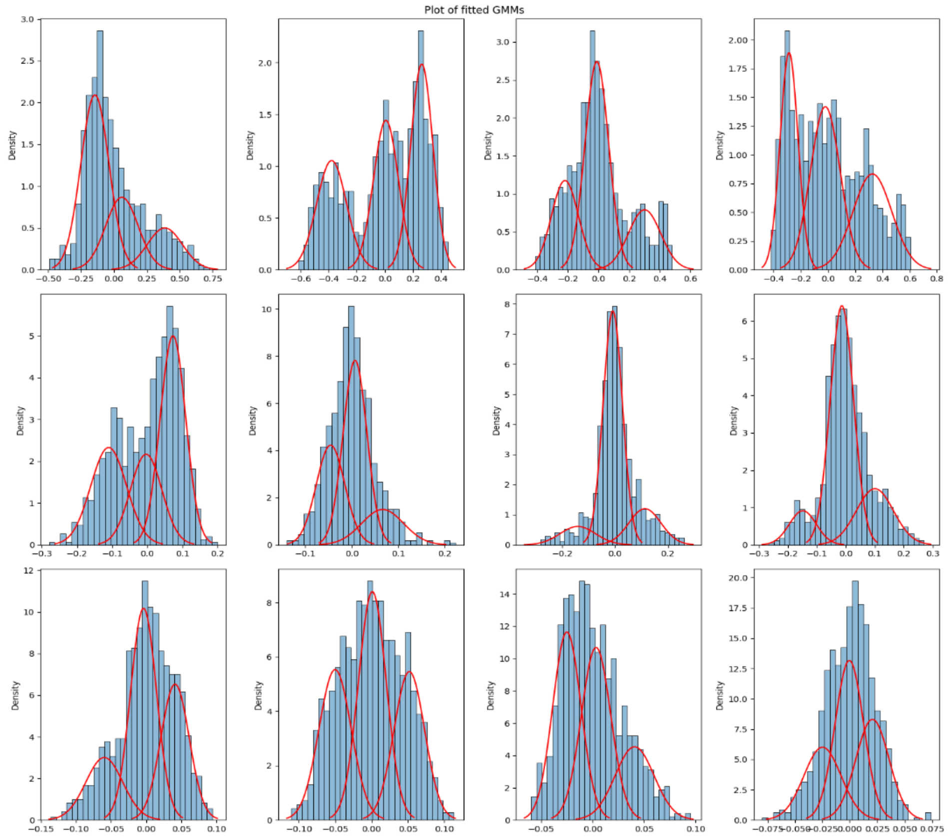
**CBF**

**wavelet: Haar**

**AMI: 0.84**
**ARI: 0.87**
**CCA: 0.89**



# Real Data Analysis

| Data | Projection Function | No. Projections | AMI | ARI | CCA |
|---|---|---|---|---|---|
| FaceFour | wavelet: bior2.4 | 64 | 0.77 | 0.76 | 0.86 |
| ECG200 | wavelet: Haar | 6 | 0.37 | 0.38 | 0.83 |
| CBF | wavelet: Haar | 14 | 0.84 | 0.87 | 0.89 |
| Symbols | Fourier | 16 | 0.75 | 0.66 | 0.69 |
| Meat | wavelet: bior2.4 | 10 | 0.70 | 0.69 | 0.84 |
| Diatom Size Reduction | OU | 32 | 0.94 | 0.95 | 0.98 |
| Trace | wavelet: db35 | 8 | 0.49 | 0.43 | 0.68 |
| ArrowHead | wavelet: db10 | 8 | 0.37 | 0.36 | 0.67 |
| GunPoint | wavelet: bior2.4 | 6 | 0.34 | 0.25 | 0.75 |

# Real Data Analysis

Benchmarking with:

funFEM
funHDDC
Funclust (from Funclustering)
FClust (from fdapace)
kmeans_align (from fdasrvf)
FADPclust

| DATA | GPMIX | FEM | HDD | CLU | FC | KM | ADP |
|---|---|---|---|---|---|---|---|
| AH | **0.37** | 0.25 | 0.22 | 0.05 | 0.24 | 0.28 | 0.19 |
| | **0.36** | 0.29 | 0.21 | 0.01 | 0.25 | 0.26 | 0.18 |
| BC | **0.24** | 0.03 | 0.06 | 0.22 | 0.08 | 0.10 | 0.08 |
| | **0.29** | 0.04 | 0.07 | 0.15 | 0.10 | 0.10 | 0.10 |
| CBF | **0.84** | 0.37 | 0.47 | 0.01 | 0.53 | 0.34 | 0.40 |
| | **0.87** | 0.35 | 0.44 | 0.00 | 0.44 | 0.31 | 0.31 |
| DSR | **0.94** | 0.79 | 0.82 | 0.00 | 0.83 | 0.72 | 0.78 |
| | **0.95** | 0.83 | 0.86 | 0.01 | 0.86 | 0.73 | 0.82 |
| ECG | **0.37** | 0.15 | 0.17 | 0.03 | 0.37 | 0.17 | 0.07 |
| | **0.38** | 0.26 | 0.28 | 0.03 | 0.37 | 0.28 | 0.14 |
| FF | **0.77** | 0.47 | 0.40 | 0.06 | 0.56 | 0.50 | 0.44 |
| | **0.76** | 0.41 | 0.36 | 0.08 | 0.54 | 0.45 | 0.32 |
| GUP | **0.34** | 0.00 | 0.00 | 0.02 | 0.00 | 0.15 | 0.01 |
| | **0.25** | 0.00 | 0.00 | 0.02 | 0.00 | 0.07 | 0.01 |
| MEAT | 0.70 | **0.93** | 0.54 | 0.36 | 0.54 | 0.66 | 0.72 |
| | 0.69 | **0.95** | 0.44 | 0.37 | 0.49 | 0.69 | 0.69 |
| SB | **0.32** | 0.08 | 0.00 | 0.03 | 0.12 | 0.07 | 0.03 |
| | **0.30** | 0.00 | 0.00 | 0.03 | 0.04 | 0.07 | 0.05 |
| SYM | 0.75 | 0.63 | 0.77 | 0.00 | **0.85** | 0.69 | 0.37 |
| | 0.66 | 0.53 | 0.67 | 0.00 | **0.80** | 0.62 | 0.30 |

# Ref


[1] E. Akeweje and M. Zhang. *Learning Mixtures of Gaussian Processes through Random Projection*. In 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 2024.

[2] T. Hsieh, Y. Sun, S. Wang, and V. Honavar. *Functional autoencoders for functional data representation learning*. In Proceedings of the 2021 SIAM International Conference on Data Mining (SDM), pages 666–674, 2021.

[3] C. Tan, G. Webb, and F. Petitjean. *Indexing and classifying gigabytes of time series under time warping*. In Proceedings of the 2017 SIAM International Conference on Data Mining (SDM), pages 282–290, 2017.

[4] Y. Li, L. Guo, L. Li, C. Yang, P. Guang, F. Huang, Z. Chen, L. Wang, and J. Hu. *Early diagnosis of type 2 diabetes based on near-infrared spectroscopy combined with machine learning and aquaphotomics*. Frontiers in Chemistry, 8, 2020.

[5] Z. Zhang, P. Tang, W. Zhang and L. Tang. *Satellite image time series clustering via time adaptive optimal transport*. Remote Sensing, 13(19), 3993, 2021.

[6] I. McDowel, D. Manandhar, C. Vockley, A. Schmid, T. Reddy and B. Engelhardt. *Clustering gene expression time series data using an infinite Gaussian process mixture model*. PLOS Computational Biology 14(1): e1005896, 2018.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

# Thank You