# Model Based Clustering Vizualization

Journées Big Data Mining and Visualization 2014

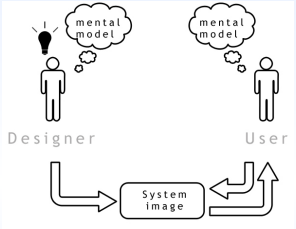
Serge Iovleff, Christophe Biernacki, Vincent Vandewalle, Komi Nagbe

Modal Team, Inria Lille Nord Europe, Université Lille 1 (Paul Painlevé), Université Lille 2 (EA 2694)

# Outline 1

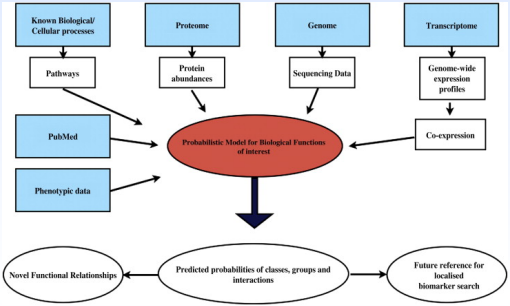
# Why do we need new visualization tools ?

Visualization is the process of making mental images, and the image will represent the model, which will serve as a pattern from which future things will emerge. Model Visualization should allow to




- Understand the model,
- Trust the Model,
- Compare different Models using Visualization.

*MΘdal* build *Probabilistic* models (Mixture Models) for heterogeneous data.




*"Overall, studies agree in that methodologies using carefully selected data of various types to predict particular classes, groups and interactions, perform better than when applied to a single type of data"* (From Georgia Tsiliki and Sophia Kossida, "Fusion methodologies for biomedical data", Journal of Proteomics, 2011).

# Outline 2

A mixture model is a generative models. The sample $\mathbf{X}$ is a realization of the following two-step process :

1. For each $i = 1, \ldots, n$

4. End For

A mixture model is a generative models. The sample $\mathbf{X}$ is a realization of the following two-step process :

1. For each $i = 1, \ldots, n$
2. $\mathbf{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K)$,
4. End For

A mixture model is a generative models. The sample $\mathbf{X}$ is a realization of the following two-step process :

1. For each $i = 1, \ldots, n$
2. $\mathbf{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K)$,
3. $\mathbf{x}_i | \mathbf{z}_i = k \sim h(.; \alpha_k)$
4. End For

A mixture model is a generative models. The sample $\mathbf{X}$ is a realization of the following two-step process :

1. For each $i = 1, \ldots, n$
2. $\mathbf{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K)$,
3. $\mathbf{x}_i | \mathbf{z}_i = k \sim h(.; \alpha_k)$
4. End For

Only the sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is observed, $\mathbf{x}_i \sim \sum_{k=1}^{K} \pi_k h(.; \alpha_k)$

A mixture model is a generative models. The sample $\mathbf{X}$ is a realization of the following two-step process :

1. For each $i = 1, \ldots, n$
2. $\mathbf{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K)$,
3. $\mathbf{x}_i | \mathbf{z}_i = k \sim h(.; \alpha_k)$
4. End For

Only the sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is observed, $\mathbf{x}_i \sim \sum_{k=1}^{K} \pi_k h(.; \alpha_k)$
Dedicated softwares (Rmixmod, mclust, rankcluster, functclustering, ...) estimate $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ by minimizing

$$\sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_k h(\mathbf{x}_i; \boldsymbol{\alpha}_k) \right)$$

A mixture model is a generative models. The sample $\mathbf{X}$ is a realization of the following two-step process :

1. For each $i = 1, \ldots, n$
2. $\mathbf{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K)$,
3. $\mathbf{x}_i | \mathbf{z}_i = k \sim h(.; \alpha_k)$
4. End For

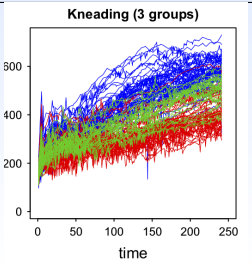
Only the sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is observed, $\mathbf{x}_i \sim \sum_{k=1}^{K} \pi_k h(.; \alpha_k)$
Dedicated softwares (Rmixmod, mclust, rankcluster, functclustering, ...) estimate $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ by minimizing

$$\sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_k h(\mathbf{x}_i; \boldsymbol{\alpha}_k) \right)$$

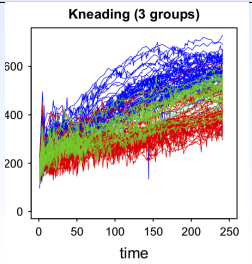
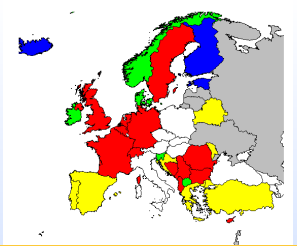
The maximum a posteriori rule provides the clusters

$$\hat{\mathbf{z}}_i = \arg\max_{k=1}^{K} t_{ik}, \qquad \text{with} \qquad t_{ik} = P(\mathbf{z}_i = k | \mathbf{x}_i; \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)}{\sum_{k=1}^{K} \hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)}$$

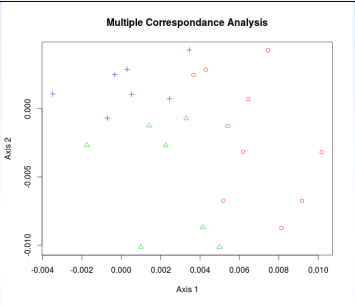
Mixture models are mainly used for Gaussian (quantitative) data. But have been extended to many other type of data

| Data type | Visualization |
| --- | --- |
| Functional data (Cookies) |  |
| | |

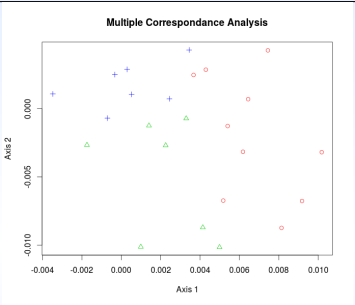
Mixture models are mainly used for Gaussian (quantitative) data. But have been extended to many other type of data

| Data type | Visualization |
|---|---|
| Functional data (Cookies) |  |
| Rank Data (Eurovision) |  |

Mixture models were mainly used for Gaussian (quantitative) data. But are extended to many other type of data

| Data type | Visualization |
| --- | --- |
| Categorical Data (Titanic) |  |

Mixture models were mainly used for Gaussian (quantitative) data. But are extended to many other type of data

| Data type | Visualization |
|---|---|
| Categorical Data (Titanic) |  |
| Heterogeneous Data (Artificial data set) | `> plot(mixmodCluster(heterodata,2))`<br>`Erreur dans plot(mixmodCluster(heterodata, 2)) :`<br>`  visualization for heterogeneous is not available yet` |

# Gaussian mixture models representation

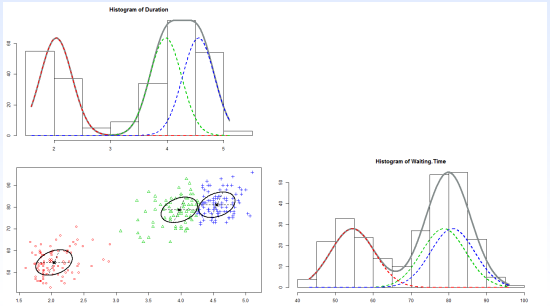



FIGURE : Geyser data set [Bowman, A. W]

This visualization allows to :

1. check if the classes are well-separated (should I try $K = 2$ groups ?)
2. check if the clusters behave similarly (orientations,...)

# Outline 3

# Main Idea

| Estimated Model | Bivariate Gaussian Model |
|---|---|
| $\sum_{k=1}^{K} \hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)$ | $\sum_{k=1}^{K} \hat{\pi}_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ |
| | |

**To retain**

Find the "closest" bivariate Gaussian mixture model. $\mathbf{Y} = (\mathbf{y}_i)_{i=1}^{n}$ give the visualization.

# Main Idea

| Estimated Model | Bivariate Gaussian Model |
|---|---|
| $\sum_{k=1}^{K} \hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)$ | $\sum_{k=1}^{K} \hat{\pi}_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ |
| *Difficult* to represent | *Easy* to represent |

**To retain**

Find the "closest" bivariate Gaussian mixture model. $\mathbf{Y} = (\mathbf{y}_i)_{i=1}^{n}$ give the visualization.

# Main Idea

| Estimated Model | Bivariate Gaussian Model |
|---|---|
| $\sum_{k=1}^{K} \hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)$ | $\sum_{k=1}^{K} \hat{\pi}_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ |
| *Difficult* to represent | *Easy* to represent |

Minimize in $(\mathbf{y}_i)_{i=1}^n, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1}^K$ the loss :

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \ell(h(\mathbf{x}_i|\hat{\boldsymbol{\alpha}}_k),\ \phi(\mathbf{y}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)).$$

with $\phi$ bivariate Gaussian density. Choice of $\ell$ ?

**To retain**

Find the "closest" bivariate Gaussian mixture model. $\mathbf{Y} = (\mathbf{y}_i)_{i=1}^n$ give the visualization.

# Outline 4

# Latent Logistic Visualization (I)

Take the $K$th class as a reference class and consider $l_{ik} = (\log(t_{ik}/t_{iK}))_{k=1}^{K-1}$, $i = 1, \ldots, n$.

**Recall**

$$t_{ik} = \frac{\hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)}{\sum_{k=1}^{K} \hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)}, \qquad h_{ik} = h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)$$

Consider an *isotropic* Gaussian mixture density on $\mathbb{R}^2$ with for each cluster the parameterized distribution

$$\phi(\mathbf{y}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k) \right\}. \tag{1}$$

And compute $\log(t^y_{ik}/t^y_{iK})$. Some computations give easily

$$l^y_{ik} = w_{0k} + \mathbf{w}'_k \mathbf{y}_i$$

with

$$w_{0k} = \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma} (\boldsymbol{\mu}_K - \boldsymbol{\mu}_k) \qquad \text{and} \qquad \mathbf{w}_k = (\boldsymbol{\mu}_K - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}.$$

# Latent Logistic Visualization (II)

Writing $\mathbf{1_n}$ the constant vector of 1 of size $n$, $\mathbf{w}_0 = (w_{0k})_{k=1}^{K-1}$, $\mathbf{Y} = (\mathbf{y}_i')_{i=1}^{n}$ and $\mathbf{W} = (\mathbf{w}_k)_{k=1}^{K-1}$ we minimize simultaneously in $\mathbf{y}$, $\mathbf{w}_0$ and $\mathbf{W}$ the trace norm.

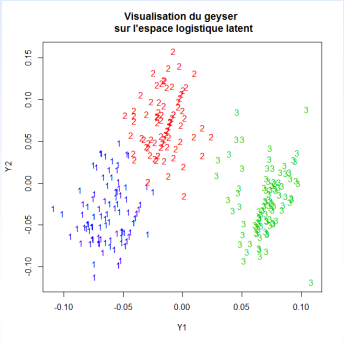
$$\|\mathbf{L} - \mathbf{L}^y\|^2 = \|\mathbf{L} - \mathbf{1}_n\mathbf{w}_0' - \mathbf{YW}\|^2, \qquad \mathbf{L} = (l_{ik})_{i=1,\,k=1}^{K,\,n}.$$

In this case it is well known that the solution for $\mathbf{w}_0$ is given by the empirical mean of the column of $\mathbf{Y}$ and for $\mathbf{Y}$ and $\mathbf{W}$ by the SVD of the column-centered matrix $\mathbf{L}$.

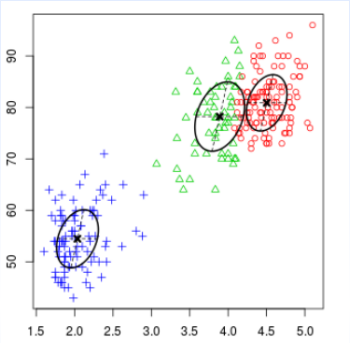
**To retain**

If $K = 3$ we get directly a planar representation.

# Latent Logistic Visualization : Geyser data set




Latent Logistic Visualization

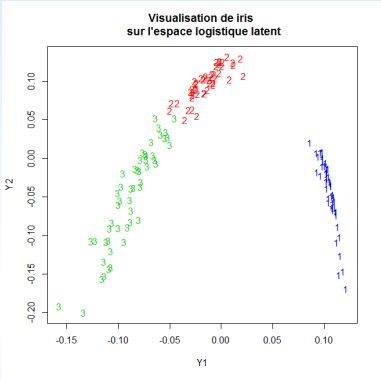


True Values

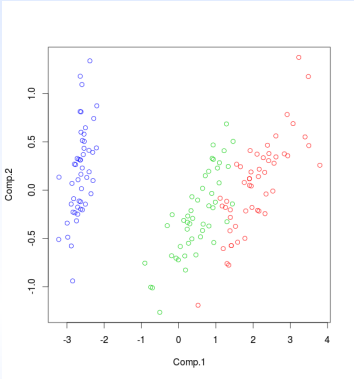
*Translations*, *Scalings* and *Rotations* (and *Colors*) can alter the visualization !

```
 wayting = 72.31438 + 195.60580 Y1 + 138.07009 Y2     R-squared:  0.9970
duration =  3.461   -  19.18 Y1    +   4.970   Y2     R-squared:  0.9999
```

# Latent Logistic Visualization : Iris data set




Latent Logistic Visualization

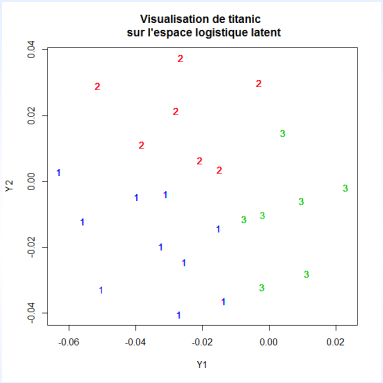


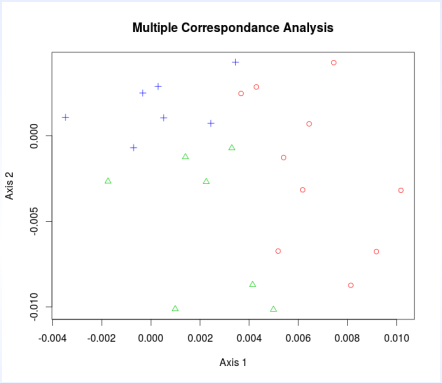

Main PCA plane

**Remark**

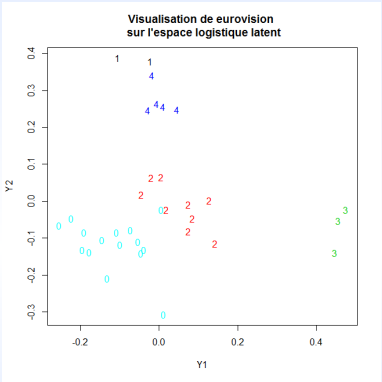
Iris data in $\mathbb{R}^4$. Main PCA plane is not the best way to represent the Iris data set.

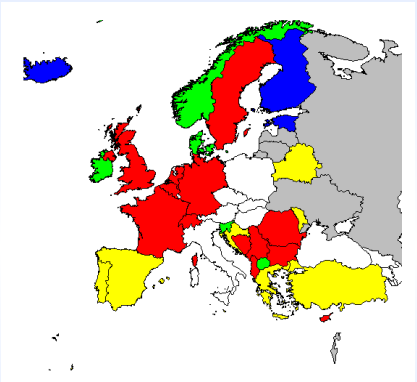
# Latent Logistic Visualization : Titanic data set (qualitative)



# Latent Logistic Visualization : Eurovision data set (ranks)



# Outline 5

# Least squares between densities

Find $(\hat{\mathbf{Y}}, \hat{a})$ such that :

$$(\hat{\mathbf{Y}}, \hat{a}) = \arg\min_{\mathbf{Y},a} \sum_{i=1}^{n} \sum_{k=1}^{K} ||h_{ik} - a\phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)||^2 \tag{2}$$

under the constraints

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} t_{ik} \mathbf{y}_i \tag{3}$$

and

$$\boldsymbol{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n} t_{ik} (\mathbf{y}_i - \boldsymbol{\mu}_k)(\mathbf{y}_i - \boldsymbol{\mu}_k)' \tag{4}$$

# Least squares between densities

The optimization algorithm iterate over $a$ and $\mathbf{y}_i$.
At iteration $r+1$ compute :

$$a^{(r+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} h_{ik}\phi(\mathbf{y}_i^{(r)};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{i=1}^{n}\sum_{k=1}^{K} h_{ik}^2}$$

And for $i$ from 1 to $n$ :

$$\begin{aligned}\mathbf{y}_j^{(r+1)} &= \arg\min_{\mathbf{y}_j\in\mathbb{R}^d} \sum_{i=1}^{j-1}\sum_{k=1}^{K} ||h_{ik} - a^{(r+1)}\phi(\mathbf{y}_i^{(r+1)};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)||^2 \\ &\quad + \sum_{k=1}^{K} ||h_{jk} - a^{(r+1)}\phi(\mathbf{y}_j;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)||^2 + \sum_{i=j+1}^{n}\sum_{k=1}^{K} ||h_{ik} - a^{(r+1)}\phi(\mathbf{y}_i^{(r)};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)||^2\end{aligned}$$

under the constraints (3) and (4).

# Outline 6

# Least squared between logs of densities

Find $(\hat{\mathbf{Y}}, \hat{\gamma})$ such that :

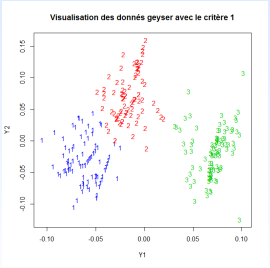
$$(\hat{\mathbf{Y}}, \hat{\gamma}) = \arg\max_{\mathbf{Y},\gamma} \sum_{i=1}^{n} \sum_{k=1}^{K} || \log h_{ik} - \gamma - \log \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)||^2 \tag{5}$$

under the constraints (3) et (4).

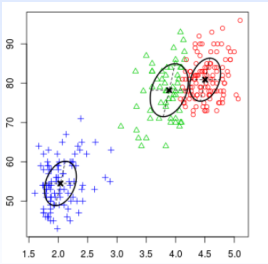
Use an alternative optimization algorithms over $\gamma$ and $\mathbf{y}_i$.

**Remark**

Criterion 1 and 2 do not give the same weights to the tail of the distribution.



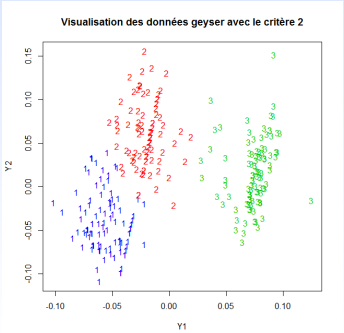

Criteria 1 : Visualization



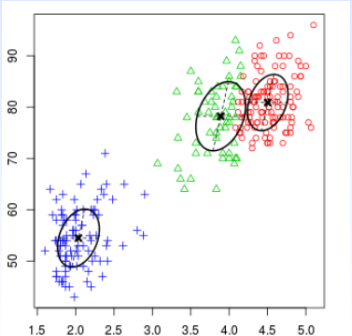
True Values

Multivariate Regression

```
 wayting = 71.94224  + 198.33749  Y1 + 143.59273  Y2     R-squared:  0.9946
duration =  3.454970 -  19.300149 Y1 +   4.805663 Y2     R-squared:  0.9952
```




Criteria 2 Visualization




True Values

Multivariate Regression

```
 wayting = 72.31438  + 181.13233 Y1 + 156.57605 Y2     R-squared:  0.997
duration =  3.461    -  19.58    Y1 +   3.066   Y2     R-squared:  0.9999
```

# Outline 7

The probability law of $X$ conditionally to the sample $\mathbf{X}$ is

$$p(X = \mathbf{x}_i, Z = k|\mathbf{X}) = \frac{\pi_k h_{ik}}{\sum_{i=1}^{n}\sum_{k=1}^{K} \pi_k h_{ik}}$$

We seek for the Gaussian distribution $P(Y = \mathbf{y}_i, Z = k)$ minimizing the Kullback–Leibler divergence :

$$\sum_{i=1}^{n}\sum_{k=1}^{K} P(X = \mathbf{x}_i, Z = k|\mathbf{X}) \log \frac{P(Y = \mathbf{y}_i, Z = k)}{P(X = \mathbf{x}_i, Z = k|\mathbf{X})}.$$

under the constraints (3) et (4). We get the criterion :

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\pi}_k f_{ik} \log \hat{\pi}_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) - \left(\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\pi}_k f_{ik}\right) \log \left(\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{\pi}_k \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

**To retain**

This criterion failed to give correct visualization : all the points in the same cluster are merged.

# Outline 8

Let $\psi$ map $\mathcal{X}$ into $\mathbb{R}^2$, we want to maximize :

$$\int_{\mathcal{X}} \sum_{k=1}^{K} \pi_k h(\mathbf{x}; \boldsymbol{\alpha}_k) \ln \phi(\psi(\mathbf{x}); \mu_k, \boldsymbol{\Sigma}_k) d\mathbf{x}.$$

**Remark**

It is required that $\phi(\psi(\mathbf{x}); \mu_k, \sigma_k)$ define a probability density $q(\mathbf{x}; \psi, \beta)$, i.e.

$$\int_{\mathcal{X}} \phi(\psi(\mathbf{x}); \mu_k, \boldsymbol{\Sigma}_k) d\mathbf{x} = 1, \qquad k \in \{1, \ldots, K\}.$$

Our aim is to minimize with respect to $(\psi, \beta)$, here $\beta = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)_{k=1}^{K}$, the Kullback-Leibler divergence between $f(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}})$ and $q(\mathbf{x}; \psi, \beta)$

$$\int_{\mathcal{X}} \sum_{k=1}^{K} \hat{\pi}_k h(\mathbf{x}; \hat{\boldsymbol{\alpha}}_k) \ln \frac{h(\mathbf{x}; \hat{\boldsymbol{\alpha}}_k)}{q(\mathbf{x}; \psi, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} d\mathbf{x}.$$

The (theoretical) criteria to maximize is thus

$$C(\psi, \beta) = \int_{\mathcal{X}} \sum_{k=1}^{K} \pi_k h(\mathbf{x}; \hat{\boldsymbol{\alpha}}_k) \ln \phi(\psi(\mathbf{x}); \mu_k, \boldsymbol{\Sigma}_k) d\mathbf{x} - \sum_{k=1}^{K} \pi_k \ln \left( \int_{\mathcal{X}} \phi(\psi(\mathbf{x}); \mu_k, \boldsymbol{\Sigma}_k) d\mathbf{x} \right)$$

The left integral is approximated using as importance sampling the distribution $\sum_{k=1}^{K} \pi_k f(\mathbf{x}; \hat{\boldsymbol{\alpha}}_k)$ :

$$\begin{aligned} \int_{\mathcal{X}} \sum_{k=1}^{K} \pi_k h(\mathbf{x}; \hat{\boldsymbol{\alpha}}_k) \ln \phi(\psi(\mathbf{x}); \mu_k, \boldsymbol{\Sigma}_k) d\mathbf{x} &\approx \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \pi_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k) \ln \phi(\psi(\mathbf{x}_i); \mu_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} f(\mathbf{x}_i; \hat{\theta}_{k'})} \quad (6) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \ln \phi(\psi(\mathbf{x}_i); \mu_k, \boldsymbol{\Sigma}_k) \quad (7) \end{aligned}$$

The right integral is approximated using as importance sampling the distribution $\sum_{k=1}^{K} \pi_k f(\mathbf{x}; \hat{\boldsymbol{\alpha}}_k)$ :

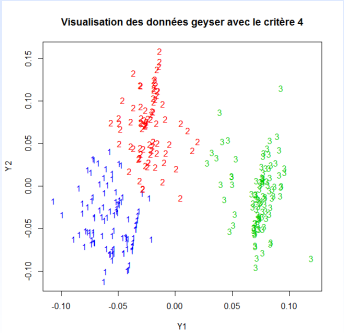
$$\int_{\mathcal{X}} \phi(\psi(\mathbf{x}); \mu_k, \boldsymbol{\Sigma}_k) d\mathbf{x} \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(\psi(\mathbf{x}_i); \mu_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{g} \hat{\pi}_k h(\mathbf{x}_i; \hat{\boldsymbol{\alpha}}_k)}$$

Putting everything together, The criteria to maximize become :

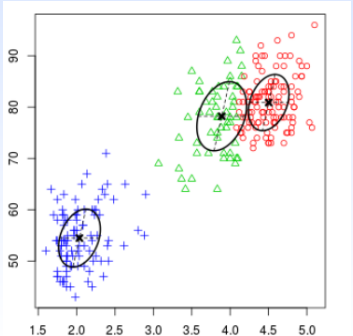
$$C(\mathbf{Y}, \alpha) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \ln \phi(\mathbf{y}_i; \mu_k, \mathbf{\Sigma}_k) - \sum_{k=1}^{K} \pi_k \ln \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(\mathbf{y}_i; \mu_k, \mathbf{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k h_{ik}} \right)$$

**Remark**

The left term is interpreted as the expectation of the completed likelihood.
The right term is interpreted as a penalization term.




Criteria 2 Visualization

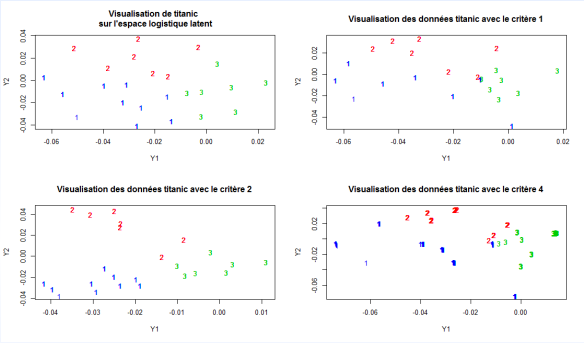


True Values

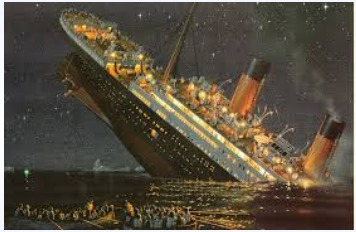
Multivariate Regression

```
 wayting = 71.86512  + 180.02938  Y1 + 159.04190  Y2     R-squared:  0.9943
duration =  3.462434 -  19.543821 Y1 +   2.684186 Y2     R-squared:  0.9945
```

# Thank You !




Titanic results



We will try to arrive...