# EXPLORATORY TOOLS IN MODEL-BASED CLUSTERING

**L.A. García-Escudero, A. Gordaliza, C. Matrán and A. Mayo-Iscar**
Dpto. de Estadística e I. O. Universidad de Valladolid
(SPAIN)

BoSanTouVal 2009

## 1.- Introduction

- **Two key questions:**
  - ◇ **How to adequately choose the number of groups in a clustering problem?**
  - ◇ **How to measure the strength of data point cluster assignments?**
- It is **impossible to answer** these two questions without:
  - ◇ Stating clearly which is the **probabilistic model assumed**.
  - ◇ Putting constrains on the **allowed clusters scatters**.
  - ◇ Stating clearly what we **understand by noise**.

# 2.- Model Based Clustering:

- Many statistical practitioners view the **Cluster Analysis** as a collection of **mostly heuristic techniques** for partitioning multivariate data.

- This view relies on the fact that most of the cluster techniques are not explicitly based on a probabilistic model:

  *"...lead the naive investigator into believing that he or she did not make any assumption at all, and that the results therefore are 'objective'..." (Flury 1997, page 123)*

⇒ A **properly stated underlying probabilistic model is convenient**

- **Two model-based clustering** approaches:

  - **Mixture approach:**

$$\prod_{i=1}^{n} \left[ \sum_{j=1}^{k} \pi_j \phi(x_i; \theta_j) \right]$$

(assign $x_i$ to cluster $j$ whenever $\pi_j \phi(x; \theta_j) > \pi_l \phi(x; \theta_l)$ for $l \neq j$).

  - **"Crisp" (0-1) approach:**

$$\prod_{j=1}^{k} \prod_{i \in R_j} \phi(x_i; \theta_j)$$

($R_j$ indexes of the $x_i$'s assigned to cluster $j$).

◇ **Mixture approach:**

$$\prod_{i=1}^{n}\left[\sum_{j=1}^{k}\pi_j\phi(x_i;\theta_j)\right] \Rightarrow \text{EM-algorithm}$$

(assign $x_i$ to cluster $j$ whenever $\pi_j\phi(x;\theta_j) > \pi_l\phi(x;\theta_l)$ for $l \neq j$).

◇ **"Crisp" (0-1) approach:**

$$\prod_{j=1}^{k}\prod_{i\in R_j}\phi(x_i;\theta_j) \Rightarrow \text{CEM-algorithm}$$

($R_j$ indexes of the $x_i$'s assigned to cluster $j$).

- **Noise** in real problems ⇒ **Robust Clustering**

- Two **robust clustering approaches** providing "theoretical well-based clustering criterion in presence of outliers" (Bock 2002):

  ⋄ **Mixture modeling:** The noise is fitted through mixture components (Fraley and Raftery, Peel and McLachlan,...)

  ⋄ **Trimming approach:** A fraction $\alpha$ of most outlying data is trimmed. (Gallegos and Ritter, Cuesta-Albertos et al., García-Escudero et al., Neykov et al.,...).

- **Noise** in real problems $\Rightarrow$ **Robust Clustering**

- Two **robust clustering approaches** providing "theoretical well-based clustering criterion in presence of outliers" (Bock 2002):

  ◇ **Mixture modeling:** The noise is fitted through mixture components (Fraley and Raftery, Peel and McLachlan,...)

  ◇ **Trimming approach:** A fraction $\alpha$ of most outlying data is trimmed. (Gallegos and Ritter, Cuesta-Albertos et al., García-Escudero et al., Neykov et al.,...).

  $\Rightarrow$ *We will focus on the trimming approach!!*

# 3.- TCLUST methodology

- **Spurious-Outlier Model** (Gallegos 2001 and Gallegos and Ritter 2005):

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j} f(x_i;\mu_j,\Sigma_j)\right]\left[\prod_{i\notin R} g_{\psi_i}(x_i)\right]$$

◇ $f(x_i;\mu,\Sigma)$ is a $p$-variate normal *p.d.f.*.

◇ $R=\cup_{j=1}^{k}R_j$ contains $[n(1-\alpha)]$ **regular** data.

◇ $g_{\psi_i}$ are some *p.d.f.*'s for the **non-regular** data.

- If **no conditions** are possed on $\Sigma_i$'s $\Rightarrow$ **Not a well-defined problem.**

- **Restrictions are needed:**
  - ◇ *Same spherical covariance matrices* (i.e., $\Sigma_j = \sigma^2 \cdot I$) $\Rightarrow$ Trimmed $k$-means (Cuesta-Albertos et al. 1997).
  - ◇ *Same (not necessarily spherical) covariance matrices* ($\Sigma_j = \Sigma$) $\Rightarrow$ Determinantal criteria (Gallegos and Ritter 2005).
  - ◇ *Different covariances but with equal scales* ($|\Sigma_1| = ... = |\Sigma_g|$) $\Rightarrow$ Heterogeneous robust clustering (Gallegos 2001, 2003)

- **A different constrain:**

$$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l(\Sigma_j) \text{ and } m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l(\Sigma_j),$$

  where $\lambda_l(\Sigma_j)$ are the **eigenvalues** of the $\Sigma_j$.

- Fix a **constant** $\mathbf{c}$ such that

$$\boxed{M_n/m_n \leq \mathbf{c} \text{ (Eigenvalues-ratio restriction)}.}$$

  ◇ $\mathbf{c}$ **controls the strength** of the restriction:
    · $\mathbf{c} = 1 \Rightarrow$ Trimmed $k$-means.
    · Large $\mathbf{c} \Rightarrow$ An almost unrestricted solution.

- It extends Hathaway's restrictions $\sigma_i^2 \leq \mathbf{c} \cdot \sigma_j^2$ for $1 \leq i, j \leq k$.

- **Weights:** We consider **group weights** $\pi_j \in [0,1]$.

- **Trimming + Eigenvalue restrictions + Weights $\Rightarrow$ TCLUST**

  (García-Escudero et al. (2008) *Annals of Statistics*, **36**, 1324-1345)

  ◇ **Existence** of both theoretical and sample solutions.
  ◇ **Consistency**.
  ◇ Feasible **algorithm**.

## 3. Guidance in choosing $k$

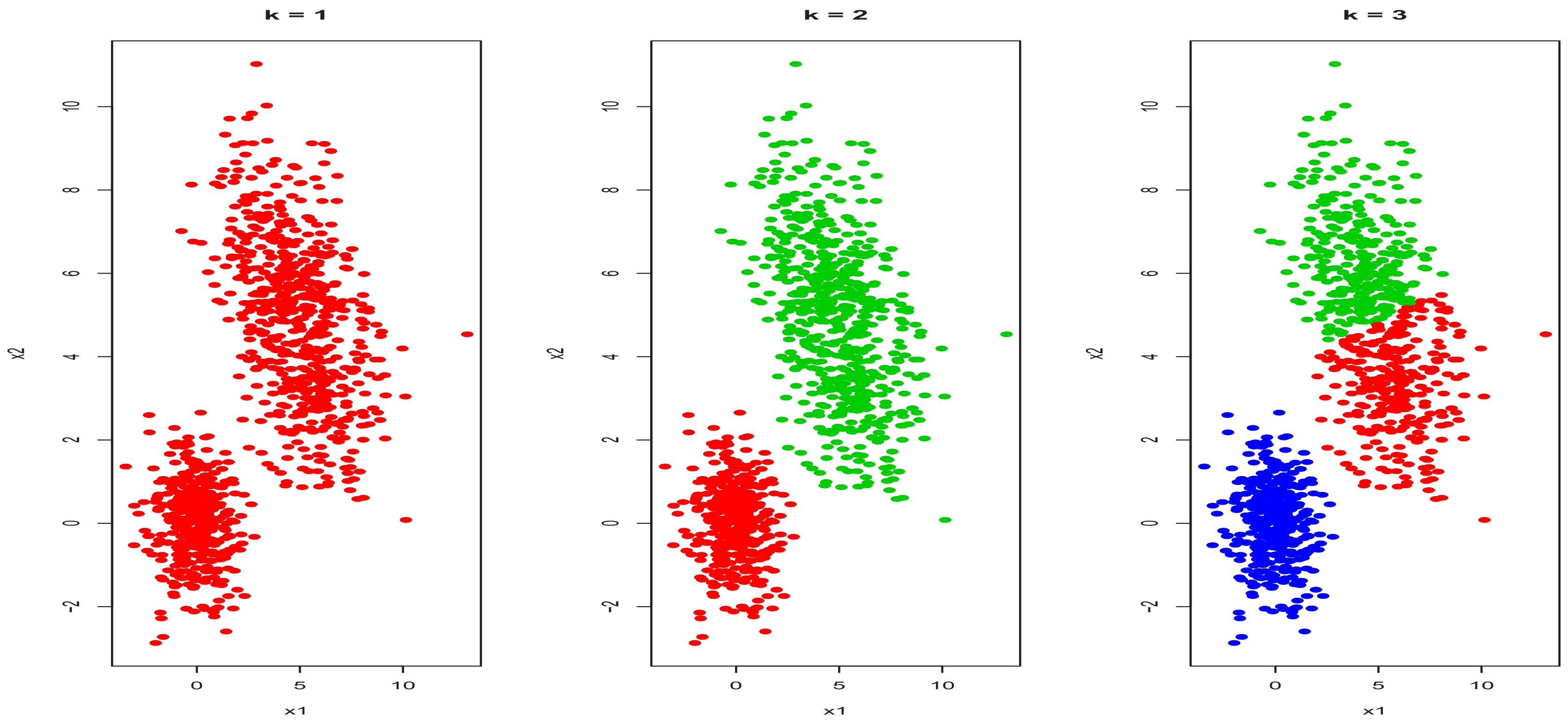
- Many procedures for choosing $k$ in "crisp" clustering are based on **monitoring** the size of the (log-) "likelihoods" :

$$k \mapsto \sum_{j=1}^{k} \sum_{i \in R_j} \log \phi(x_i; \widehat{\theta}_j) \text{ for } k = 1, 2, \ldots \, .$$

- **Examples:**
  - ◇ $\Sigma_j = \sigma^2 I$ ($k$-means) $\Rightarrow$ Friedman and Rubin 1967, Engelman and Hartigan 1969, Calinski and Harabasz 1974,...
  - ◇ $\Sigma_j = \Sigma$ (determinant criterium) $\Rightarrow$ Marriot 1971,...

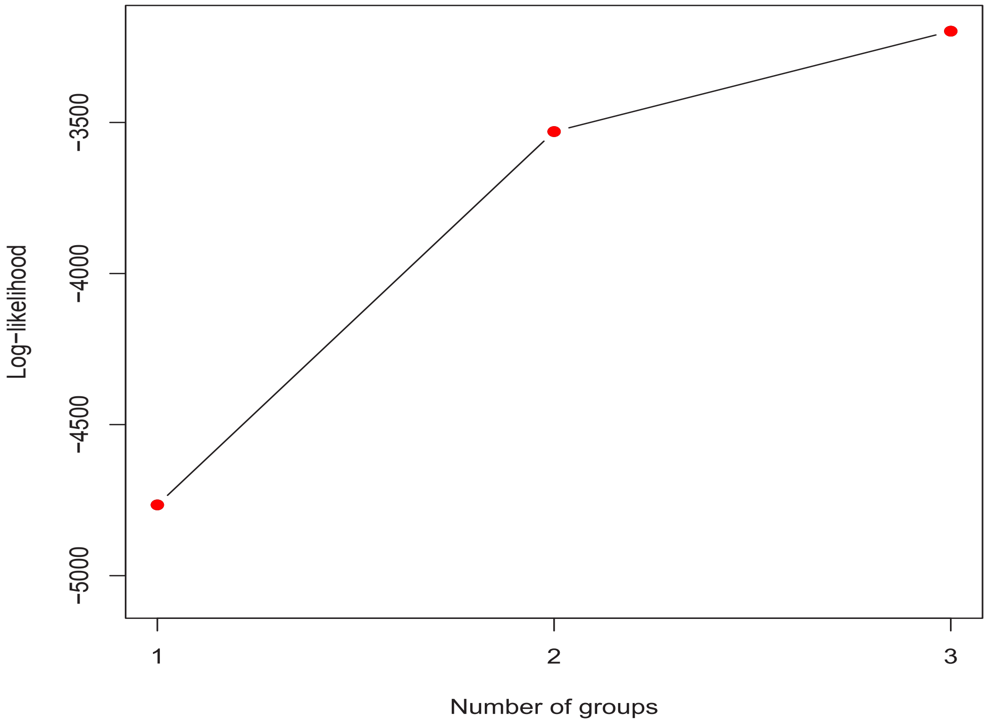
- **Trimmed versions** can be also considered $\Rightarrow$ García-Escudero et al 2003.

- **Drawback:** The log-likelihoods **strictly increase** when increasing $k$:



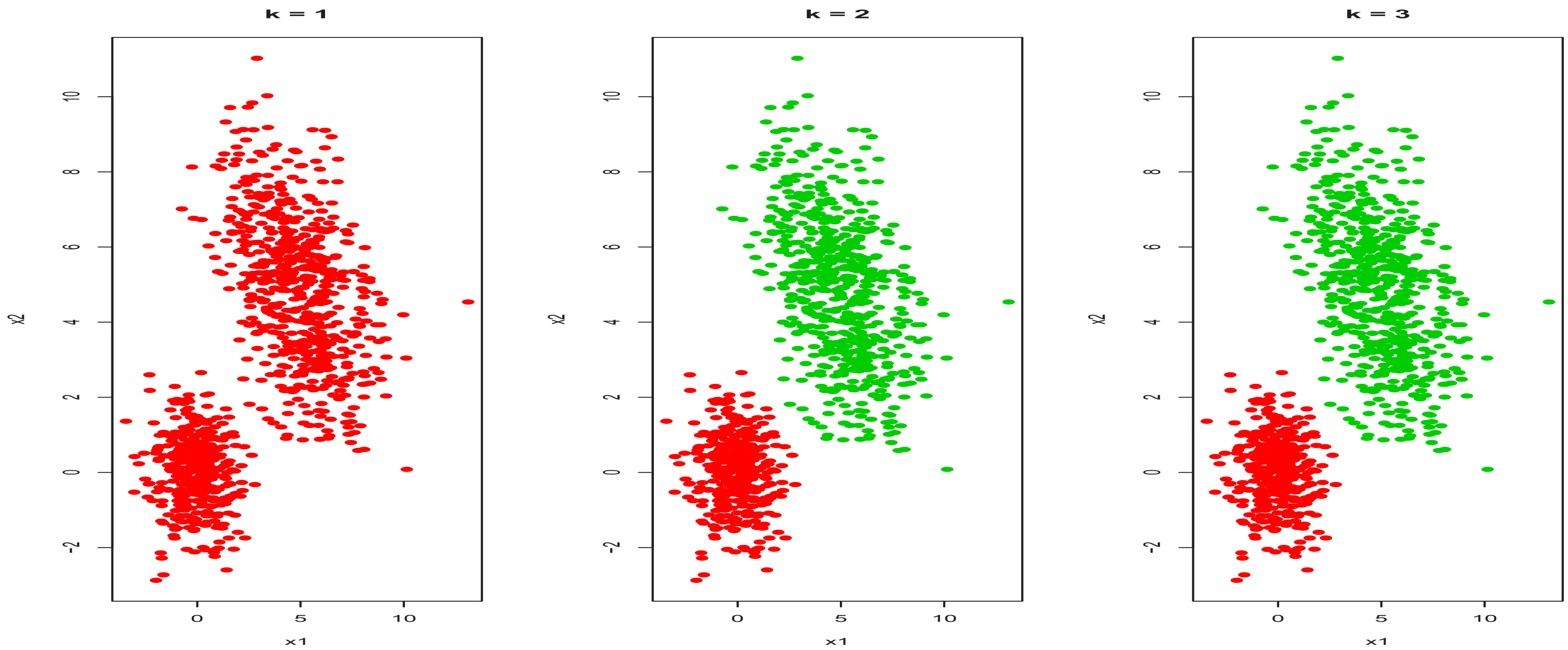

- Log-likelihoods: -4765.8 $(k=1) <$ -3530.1 $(k=2)$ $\boxed{<}$ -3197.7 $(k=3)$

- **Figure:**




- **Solutions:**
  - ◇ Searching for an "elbow".
  - ◇ Nonlinear transformation by Sugar and James 2003.

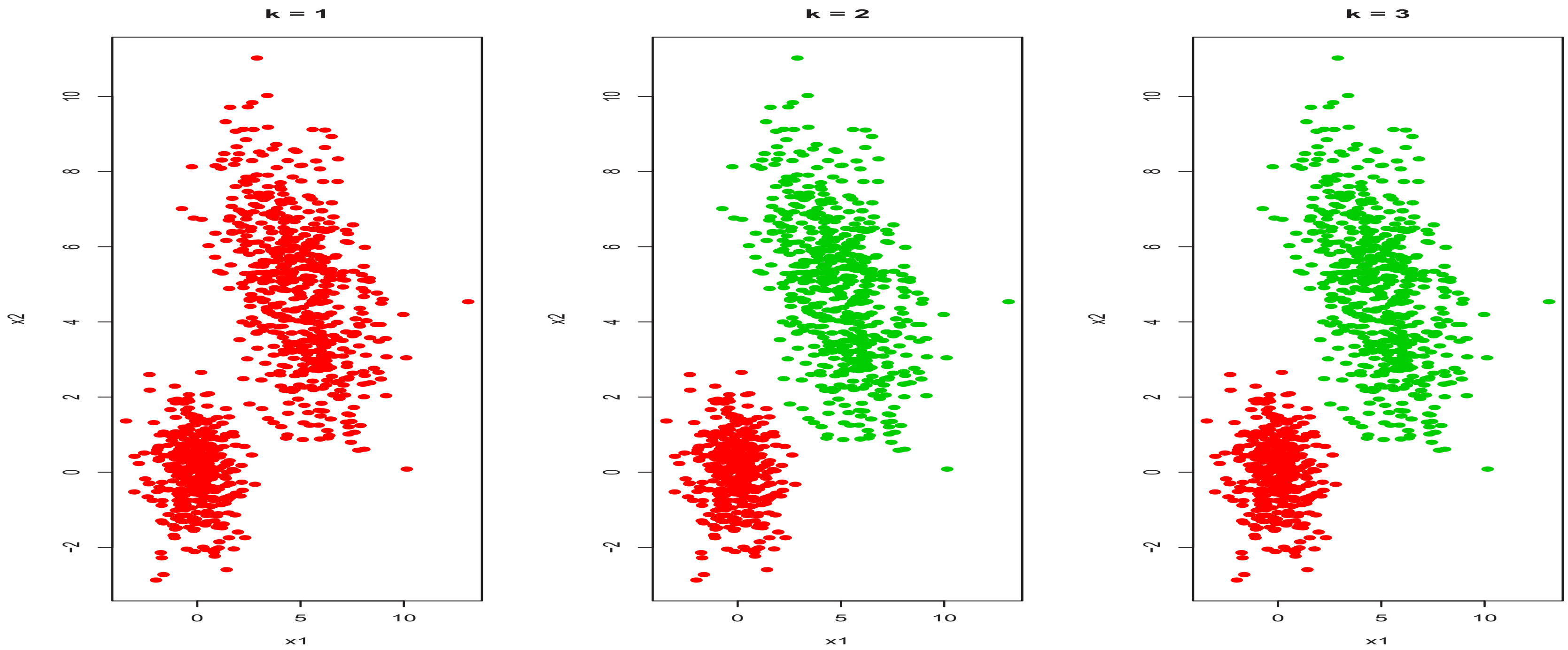
- The **TCLUST does not suffer from this problem**:




- Log-likelihoods: -4765.8 $(k = 1) <$ -4203.2 $(k = 2)$ $\boxed{=}$ -4203.2 $(k = 3)$

- Recall the presence of **weights** (which can be **set to zero**) $\Rightarrow \pi_3 = 0$ *in* $k = 3$ *solution!.*

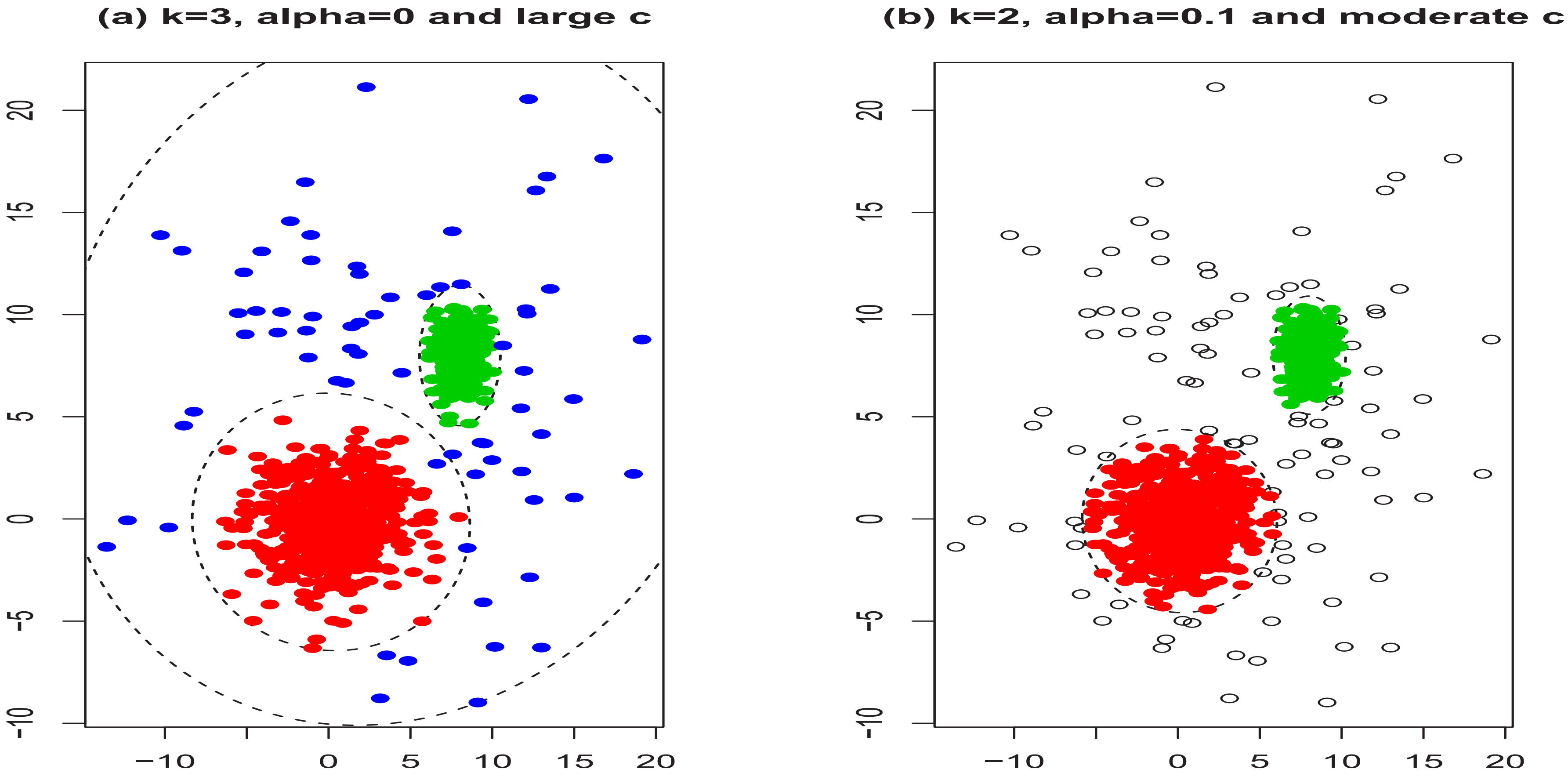
- The **TCLUST does not suffer from this problem**:




- Log-likelihoods: -4765.8 $(k = 1) <$ -4203.2 $(k = 2)$ $\boxed{=}$ -4203.2 $(k = 3)$

- *This fact was already noticed by Bryant (1991) when dealing with the so-called <u>Penalized</u> Classification Maximum Likelihood...*

- **Importance of the "trimming" and the scatter constrain:**




⇒ "∘" are trimmed points in the figure on the right.

## 4.- Classification Trimmed Likelihood Curves

- **Based on the TCLUST methodology**, we monitor:

$$(\alpha, k) \mapsto \mathcal{L}(\alpha, k) := \sum_{j=1}^{k} n_j \log \widehat{\pi}_j + \sum_{j=1}^{k} \sum_{i \in R_j} \log \phi(x_i; \widehat{\theta}_j),$$

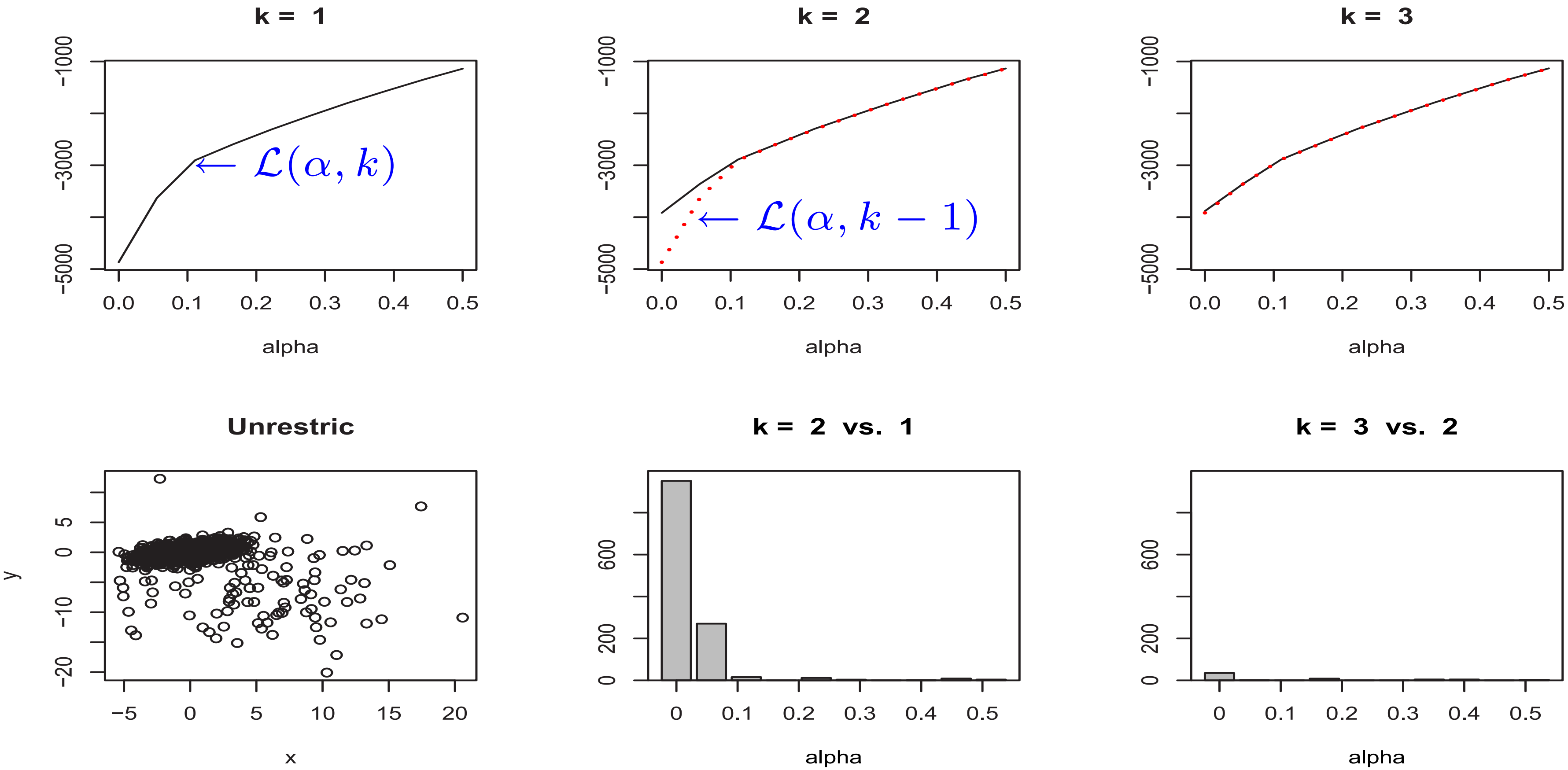
  when $k = 1, 2, ...$ and $\alpha \in (0, 1)$.

- Smallest $k$ such that $\mathcal{L}(\alpha, k) \simeq \mathcal{L}(\alpha, k+1)$ (for almost every $\alpha$) $\Rightarrow k$ is a good **choice for the number of groups**.

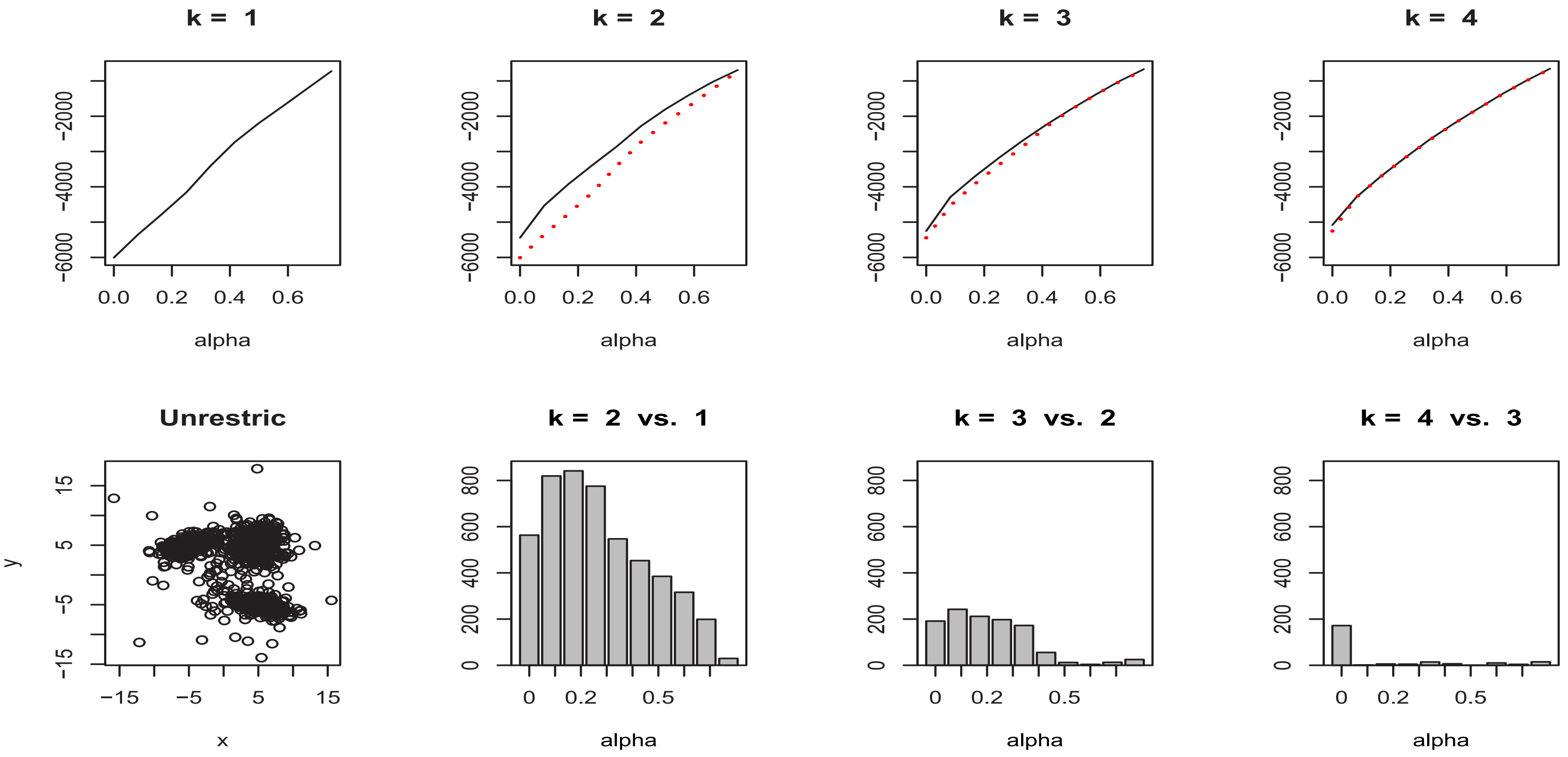
- $\mathcal{L}(\alpha, k)$ increase very fast till $\alpha \leq \alpha_0 \Rightarrow \alpha_0$ is a good **choice for the trimming level**.

- They provide **information about the group sizes**.

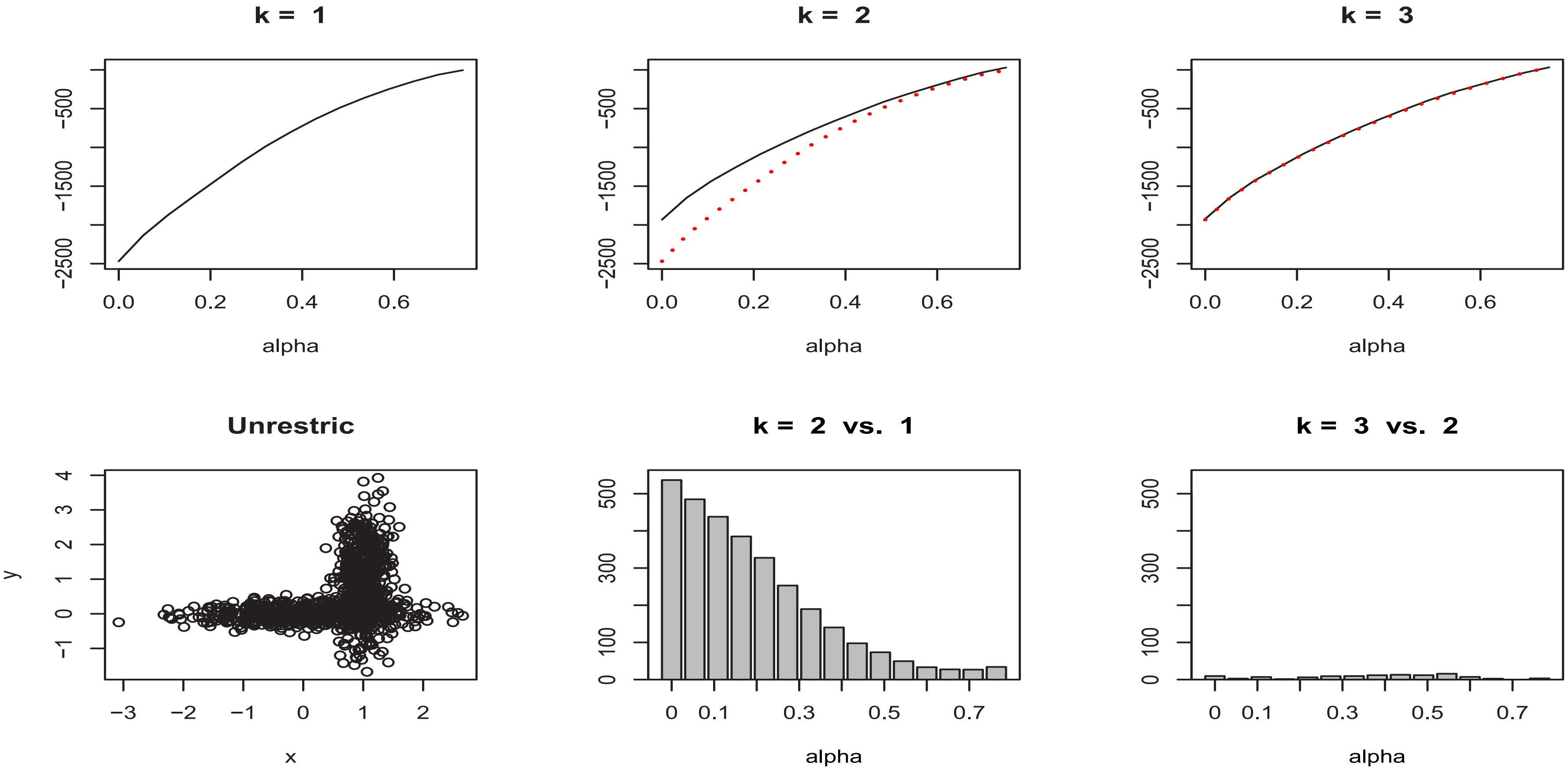
**Example 1:** Mixture $0.9 \cdot N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix} \right) + 0.1 \cdot N\left( \begin{pmatrix} 5 \\ -5 \end{pmatrix}, \begin{pmatrix} 30 & 0 \\ 0 & 30 \end{pmatrix} \right)$



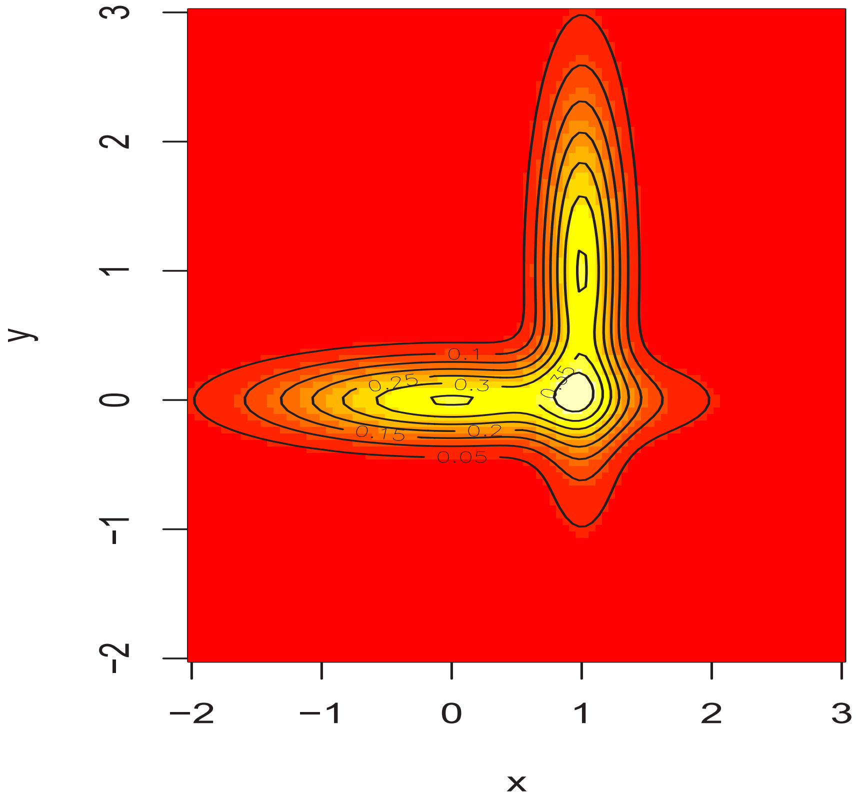
**Example 2:** $0.3 \cdot N\left(\binom{-5}{5}, \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}\right) + 0.3 \cdot N\left(\binom{5}{-5}, \begin{pmatrix} 4 & -1 \\ -1 & 1 \end{pmatrix}\right) + 0.3 \cdot N\left(\binom{5}{5}, \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}\right) + 0.1 \cdot N\left(\binom{0}{0}, \begin{pmatrix} 30 & 0 \\ 0 & 30 \end{pmatrix}\right)$

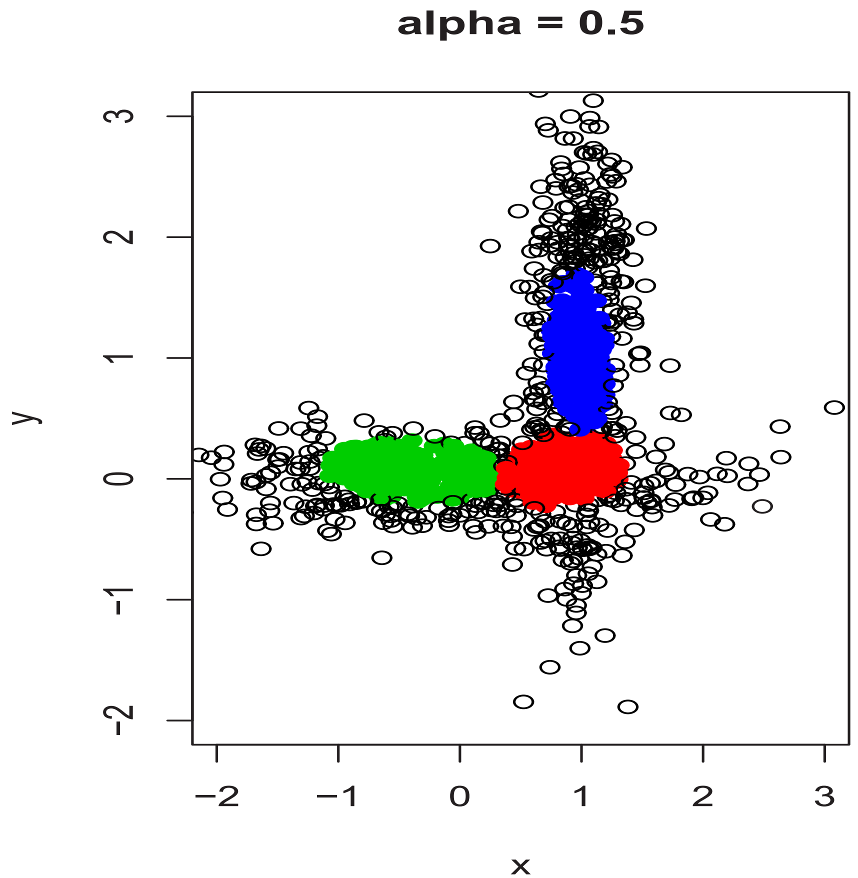


**Example 3:** "The topography of multivariate normal mixtures" (Ray and Lindsay 2005) ⇒ Mixture with 2 components.



Mixture with 2 components with 3 modes:



## 5.- Strength of cluster assignments

- **Confirmatory tool:** *Were satisfactory the choices made for $k$ and $\alpha$?*
- The **strength of the cluster assignment** of observation $x_i$ to group $j$:

$$\boxed{D_j(x_i, \hat{\theta}) = \pi_j \phi(x_i, \hat{\theta}_j)}$$

- If $D_{(1)}(x, \hat{\theta}) \leq ... \leq D_{(k)}(x, \hat{\theta})$, define some **Bayes factors** as:

$$\text{BF}(i) = \log\left(D_{(k-1)}(x_i; \hat{\theta}) / D_{(k)}(x_i; \hat{\theta})\right).$$

  ◇ Small BF$(i)$ $\Rightarrow$ Clear cluster assignment for the observation $x_i$.

- **Bayes factors for trimmed points:** Given the maximum possible strength:

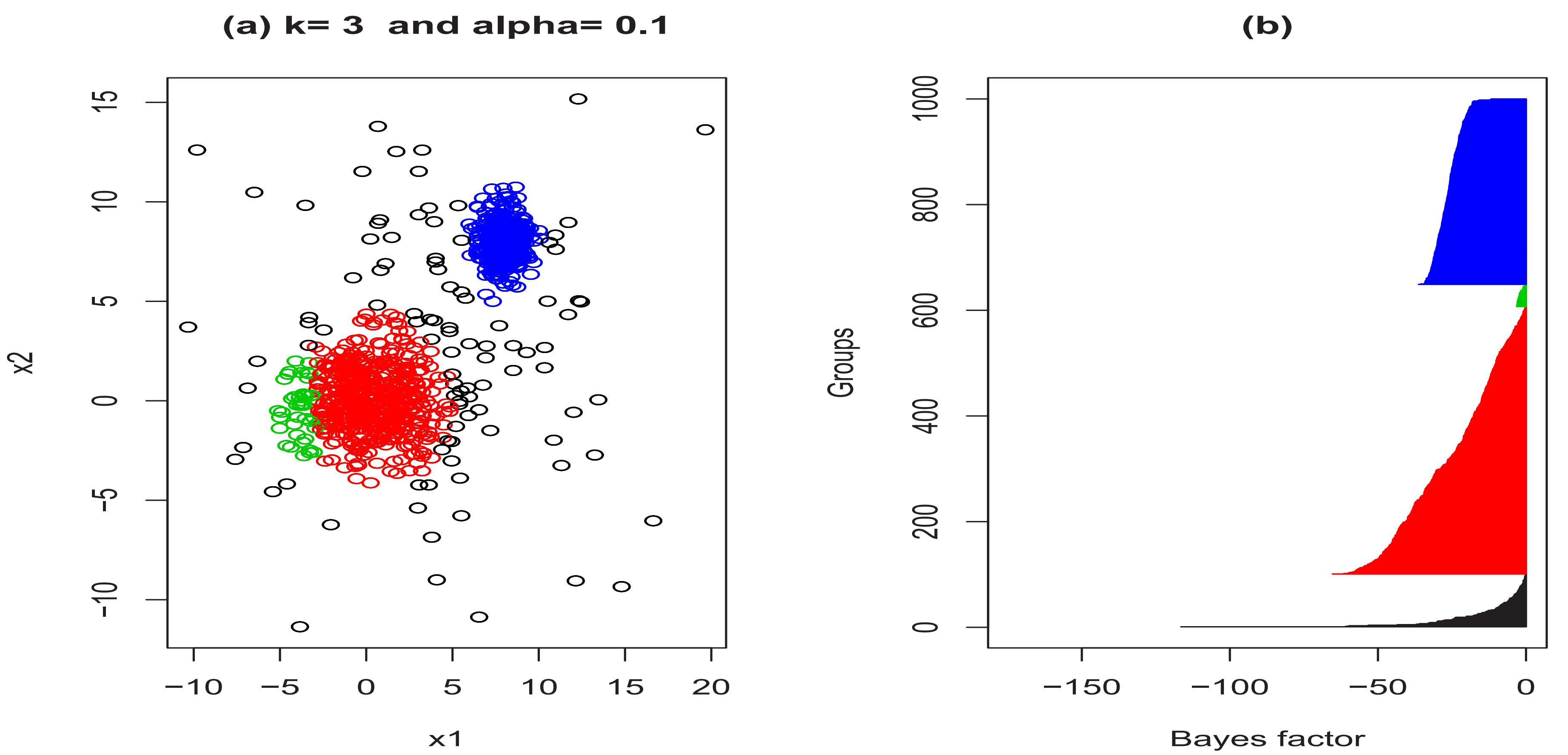
$$d_i = \max_{j=1,\ldots,k} \{\pi_j \phi(x_i, \widehat{\theta_j})\} = D_{(k)}(x_i, \hat{\theta}), \text{ we have:}$$

◇ $\boxed{d_{(1)} \leq \ldots \leq d_{([n\alpha])}} \leq \ldots \leq d_{(n)} \Rightarrow \boxed{[n\alpha] \text{ observations to be trimmed.}}$

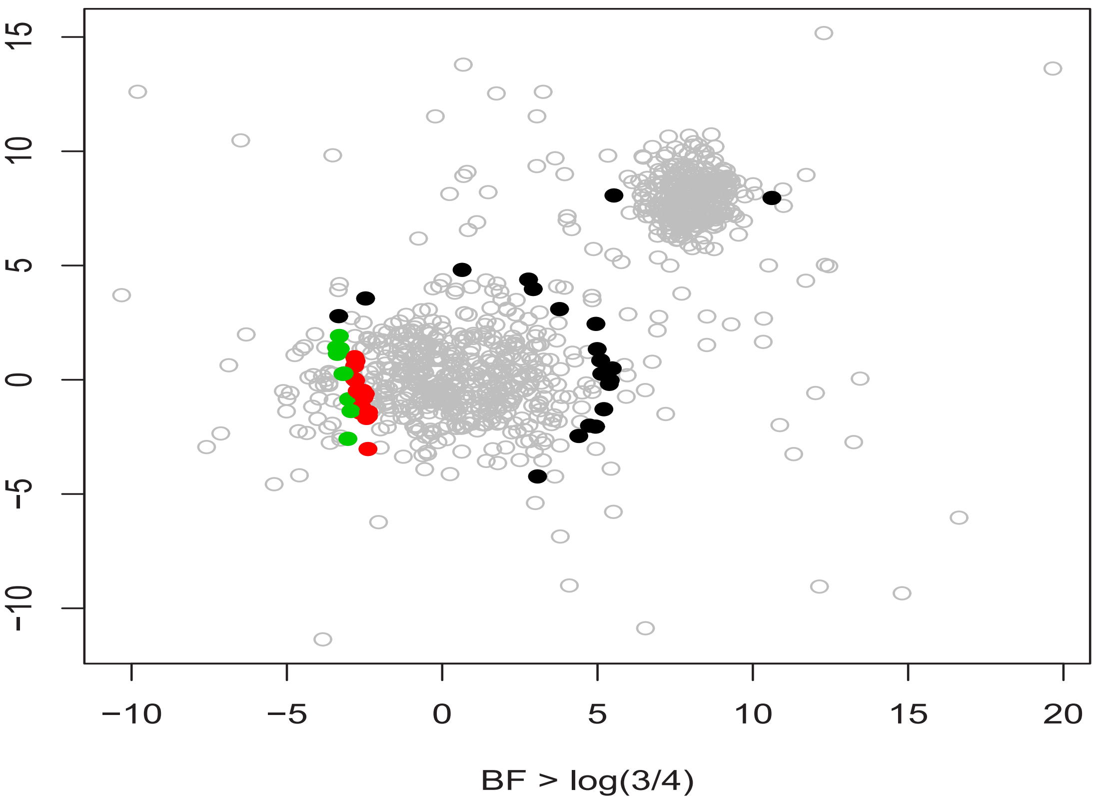
◇ Bayes factors for trimmed data $\Rightarrow$ BF$(i) = \log \left( D_{(k)}(x_i, \hat{\theta}) / d_{([n\alpha])} \right)$.

◇ Small BF$(i)$ $\Rightarrow$ More clearly observation $i$ should be trimmed.

- **Graphical display I: "Silhouette" plot**



- **Graphical display II: “Most Doubtful assignments”**
  - ◇ Label observations $i$'s with BF$(i) \geq \log(3/4)$:




*[PCA, discriminant or Bhattacharyya coordinates (Hennig and Christlieb 2002) if $p > 2$...]*