Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.

Jean-Patrick Baudry

Directeur de thèse : Gilles Celeux

Université Paris-Sud 11
Projet SELECT (INRIA)

3 Décembre 2009

Model Selection for Clustering. How Many Classes?

Jean-Patrick Baudry

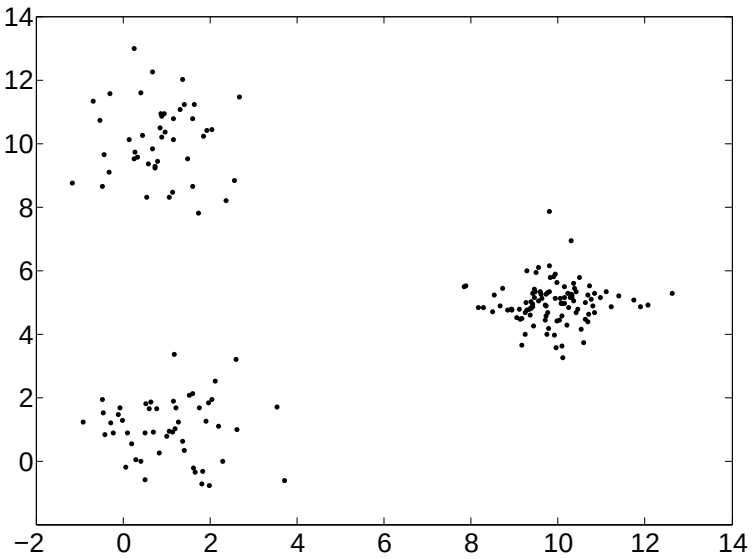
Advisor: Gilles Celeux

Université Paris-Sud 11
Project SELECT (INRIA)

December 3, 2009

Table of contents

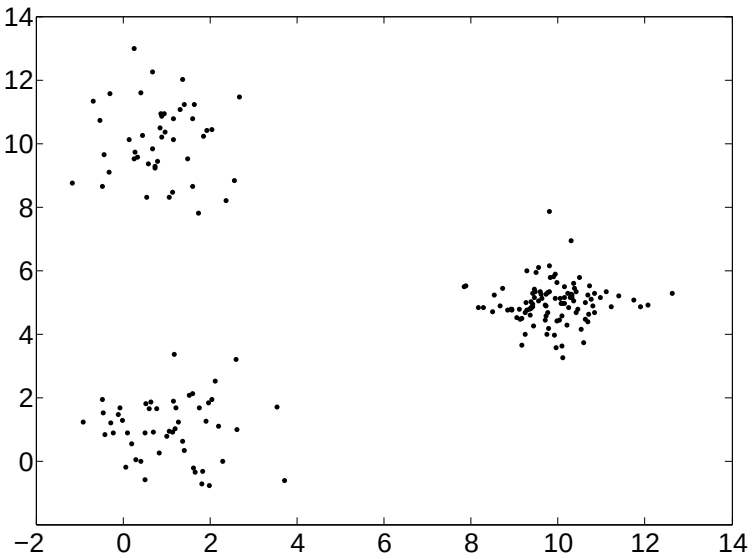
Clustering



Data: $x_1, \ldots, x_n \in \mathbb{R}^d$.

Aim: designing K classes.

Clustering

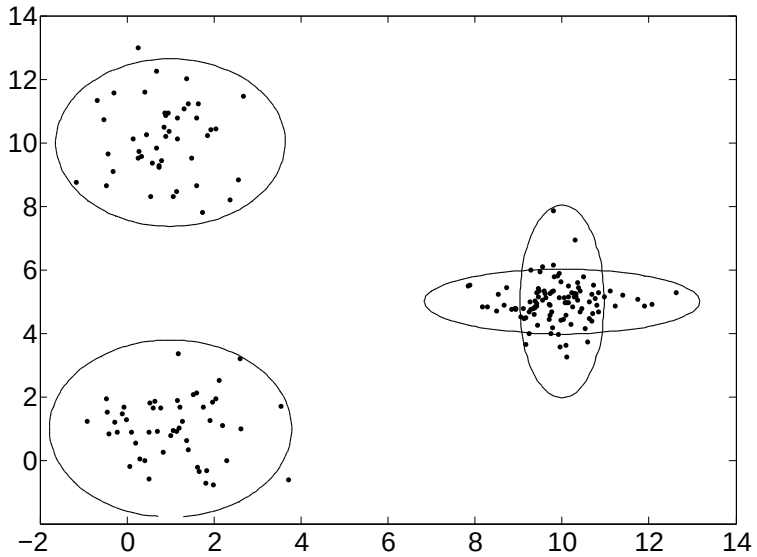


Data: $x_1, \ldots, x_n \in \mathbb{R}^d$.

Aim: designing K classes

? ?

Clustering



200 observations from a four-component Gaussian mixture

Data: $x_1, \ldots, x_n \in \mathbb{R}^d$.

Aim: designing K classes.

Model-Based Clustering

Statistical Approach: $(x_1, \ldots, x_n)$ realization of $(X_1, \ldots, X_n)$ i.i.d. $\sim f^{\wp}$.

- Fit a mixture model to the data.

Model-Based Clustering

- Fit a mixture model to the data.

$$\mathcal{M}_K = \left\{ \sum_{k=1}^{K} \pi_k \phi(\,.\,;\omega_k) \,\middle|\, (\pi_1, \ldots, \pi_K, \omega_1, \ldots, \omega_K) \in \Theta_K \right\},$$

$$\text{with} \begin{cases} \Theta_K \subset \Pi_K \times \left(\mathbb{R}^d \times \mathbb{S}_+^d\right)^K \\ \Pi_K = \left\{(\pi_1, \ldots, \pi_K) \in [0,1] : \sum_{k=1}^{K} \pi_k = 1\right\}. \end{cases}$$

Let us denote:

- $f(\,.\,;\theta) = \sum_{k=1}^{K} \pi_k \phi(\,.\,;\omega_k)$, for all $\theta \in \Theta_K$.
- $D_K = \dim(\Theta_K)$, "number of free parameters".

- Mixture form $\leftrightarrow$ Choice of constraints on Θ_K.
- One model $\leftrightarrow$ One number of components K.

Model-Based Clustering

- Fit a mixture model to the data.

- Design classes according to the rule

  "One Gaussian component = One class"

  .

Model-Based Clustering

- Fit a mixture model to the data. Usually:

$$\widehat{\theta}_K^{\text{MLE}} \in \underset{\theta \in \Theta_K}{\text{argmax}} \underbrace{\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(X_i; \omega_k)}_{\log \mathrm{L}(\theta)}.$$

  Study based on a good estimation of the sample distribution.

- Design classes according to the rule

  "One Gaussian component = One class"

  based on the Maximum A Posteriori:

$$\forall x, \forall k, \forall \theta \in \Theta_K, \qquad \tau_k(x; \theta) = \frac{\pi_k \phi(x; \omega_k)}{\sum_{k'=1}^{K} \pi_{k'} \phi(x; \omega_{k'})}.$$

$$\widehat{z}^{\text{MAP}}(\widehat{\theta}_K^{\text{MLE}}) = \underset{1 \leq k \leq K}{\text{argmax}} \ \tau_k(x; \widehat{\theta}_K^{\text{MLE}}).$$

Choosing the Number of Classes: Model Selection

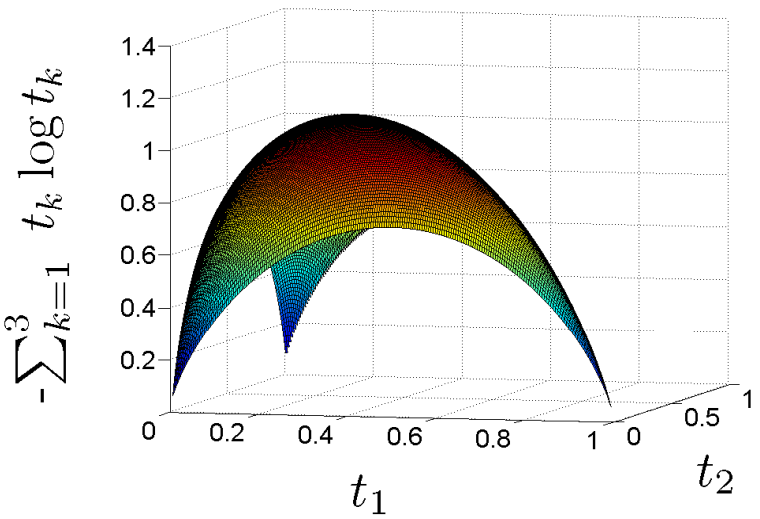
Penalized Likelihood Criteria.

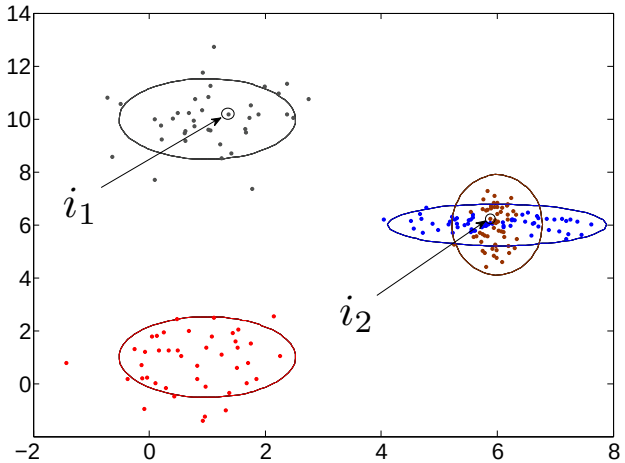
- Efficiency: minimize $d_{KL}\left(f^{\wp}, f(\,.\,;\widehat{\theta}_K^{\text{MLE}})\right)$.
  - AIC : $\hat{K}^{\text{AIC}} = \text{argmin}_{1\leq K\leq K_M}\left\{-\log \text{L}(\widehat{\theta}_K^{\text{MLE}}) + D_K\right\}$;
  - Slope heuristics (Birgé and Massart, 2006).
- Identification: minimize $\min_{\theta\in\Theta_K} d_{KL}\left(f^{\wp}, f(\,.\,;\theta)\right)$.
  - BIC : $\hat{K}^{\text{BIC}} = \underset{1\leq K\leq K_M}{\text{argmin}}\left\{-\log \text{L}(\widehat{\theta}_K^{\text{MLE}}) + \frac{D_K}{2}\log n\right\}$.
- A criterion adapted to clustering:
  - ICL (Biernacki, Celeux, Govaert, 2000) :
    $\hat{K}^{\text{ICL}} = \underset{1\leq K\leq K_M}{\text{argmin}}\left\{-\log \text{L}(\widehat{\theta}_K^{\text{MLE}}) + \text{ENT}(\widehat{\theta}_K^{\text{MLE}}) + \frac{D_K}{2}\log n\right\}$.

Entropy: Measure of the Assignment Confidence

$$\mathrm{ENT}(\theta; x) = -\sum_{k=1}^{K} \tau_k(x;\theta) \log \tau_k(x;\theta) \in [0, \log K].$$

$$\mathrm{ENT}(\theta) = \sum_{i=1}^{n} \mathrm{ENT}(\theta; x_i).$$





$\mathrm{ENT}(\widehat{\theta}_4^{\mathrm{MLE}}; x_{i_1})$ close to 0.
$\mathrm{ENT}(\widehat{\theta}_4^{\mathrm{MLE}}; x_{i_2})$ close to $\log 2$.

Choosing the Number of Classes: Model Selection

Penalized Likelihood Criteria.

- Efficiency: minimize $d_{KL}\left(f^{\wp}, f(\,.\,;\widehat{\theta}_K^{\text{MLE}})\right)$.
  - AIC : $\hat{K}^{\text{AIC}} = \text{argmin}_{1\leq K\leq K_M}\left\{-\log \text{L}(\widehat{\theta}_K^{\text{MLE}}) + D_K\right\}$;
  - Slope heuristics (Birgé and Massart, 2006).
- Identification: minimize $\min_{\theta\in\Theta_K} d_{KL}\left(f^{\wp}, f(\,.\,;\theta)\right)$.
  - BIC : $\hat{K}^{\text{BIC}} = \underset{1\leq K\leq K_M}{\text{argmin}}\left\{-\log \text{L}(\widehat{\theta}_K^{\text{MLE}}) + \frac{D_K}{2}\log n\right\}$.
- A criterion adapted to clustering:
  - ICL (Biernacki, Celeux, Govaert, 2000) :
    $\hat{K}^{\text{ICL}} = \underset{1\leq K\leq K_M}{\text{argmin}}\left\{-\log \text{L}(\widehat{\theta}_K^{\text{MLE}}) + \text{ENT}(\widehat{\theta}_K^{\text{MLE}}) + \frac{D_K}{2}\log n\right\}$.

Conditional Classification Likelihood

The classification log-likelihood for the complete data $(\underline{X}, \underline{Z})$ in model $\mathcal{M}_K$:

$$\log \mathrm{L_c}\left(\theta; (\underline{X}, \underline{Z})\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \pi_k \phi(X_i; \omega_k).$$

A key relation:

$$\log \mathrm{L_c}(\theta) = \log \mathrm{L}(\theta) + \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{ik} \log \tau_k(X_i; \theta).$$

Considering the conditional expectation of this relation:

Definition

The conditional classification log-likelihood is

$$\log \mathrm{L_{cc}}(\theta) = \log \mathrm{L}(\theta) - \mathrm{ENT}(\theta).$$

Contrast Minimization for Clustering

Contrast:

$$-\log \mathsf{L}_{\mathsf{cc}}(\theta) = -\log \mathsf{L}(\theta) + \mathsf{ENT}(\theta).$$

Associated loss:

$$\mathbb{E}_{f^{\wp}}\left[-\log \mathsf{L}_{\mathsf{cc}}(\theta)\right] - \min_{\theta} \mathbb{E}_{f^{\wp}}\left[-\log \mathsf{L}_{\mathsf{cc}}(\theta)\right]$$
$$\longleftrightarrow d_{KL}\left(f^{\wp}, f(\,.\,;\theta)\right) + \mathbb{E}_{f^{\wp}}\left[\mathsf{ENT}(\theta)\right].$$

Approximation in the model $\mathcal{M}_K$:

$$\Theta_K^0 = \underset{\theta \in \Theta_K}{\operatorname{argmin}} \left\{ d_{\mathsf{KL}}\left(f^{\wp}, f(\,.\,;\theta)\right) + \mathbb{E}_{f^{\wp}}\left[\mathsf{ENT}(\theta)\right] \right\}.$$

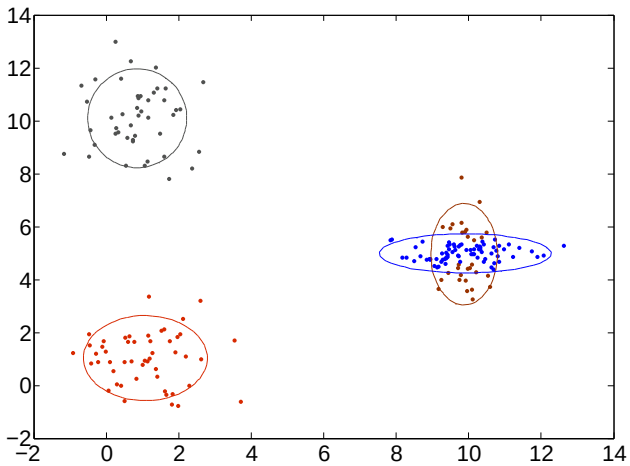
Even if $f^{\wp} \in \mathcal{M}_K$, there is no reason that $f^{\wp} \in \Theta_K^0$.

Estimation: MLccE

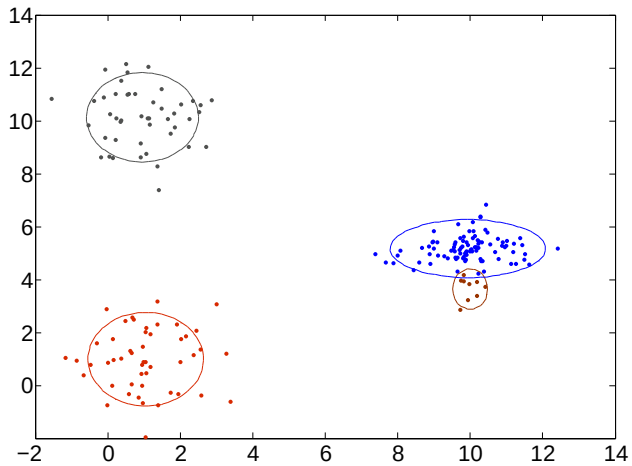
Definition

The minimum empirical contrast estimator, called "Maximum conditional classification Likelihood Estimator", is defined by

$$\widehat{\theta}^{\mathrm{MLccE}} \in \underset{\theta \in \Theta_K}{\operatorname{argmin}} \left\{ -\log \mathrm{L}_{\mathrm{cc}}(\theta) \right\}.$$



MLE



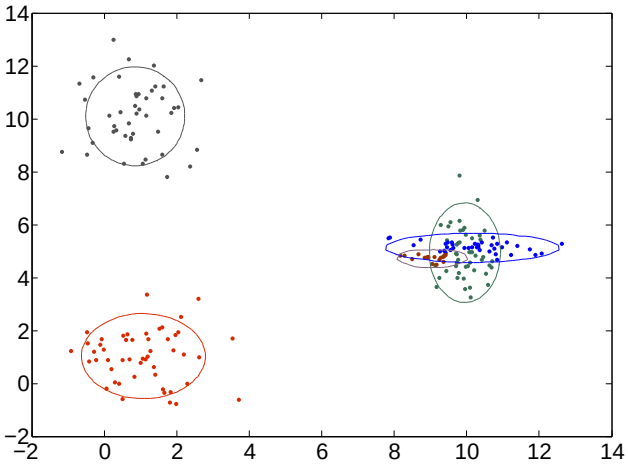
MLccE

Estimation: MLccE

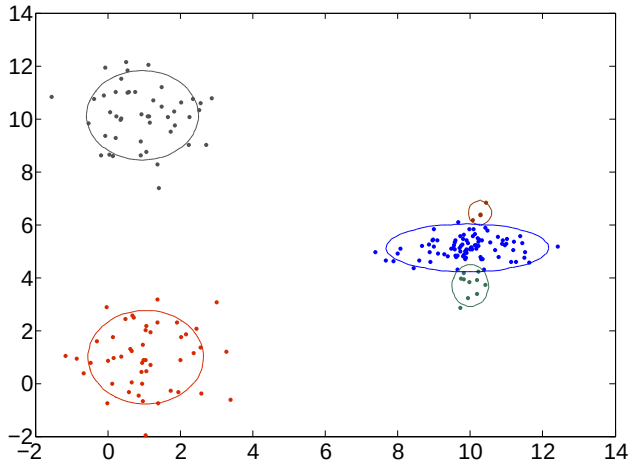
Definition

The minimum empirical contrast estimator, called "Maximum conditional classification Likelihood Estimator", is defined by

$$\widehat{\theta}^{\mathrm{MLccE}} \in \underset{\theta \in \Theta_K}{\operatorname{argmin}} \left\{ -\log \mathrm{L}_{\mathrm{cc}}(\theta) \right\}.$$



MLE



MLccE

MLccE Properties

Theorem

Let $K \in \mathbb{N}^*$. Assume

- Θ_K compact and convex;
- $H'_K(x) = \sup_{\theta \in \Theta_K} \left\| \left(\frac{\partial \log L_{cc}}{\partial \theta} \right)_{(\theta;x)} \right\|_\infty < \infty$ a.s. and $\|H'_K\|_1 < \infty$.
- $\widehat{\theta}^{\mathrm{MLccE}} \in \Theta_K$ such that for all $\theta_K^0 \in \Theta_K^0$,

$$-\log L_{cc}(\widehat{\theta}_K^{\mathrm{MLccE}}) \leq -\log L_{cc}(\theta_K^0) + o_{\mathbb{P}}(1).$$

Then

$$d(\widehat{\theta}_K^{\mathrm{MLccE}}, \Theta_K^0) \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

With $d(\theta, \widetilde{\Theta}) = \inf_{\widetilde{\theta} \in \widetilde{\Theta}} \|\theta - \widetilde{\theta}\|_\infty$.

Computing MLccE: L_{cc}-EM

The L_{cc}-EM algorithm for MLccE:

$$\widehat{\theta}_K^{\text{MLccE}} = \underset{\theta \in \Theta_K}{\text{argmax}} \left\{ \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(X_i; \omega_k) - \text{ENT}(\theta) \right\}.$$

Initialization: Random, Small_L_{cc}_EM, CEM, Km1...

Iteration $\theta^j \to \theta^{j+1}$:

E Step Compute $Q(\theta, \theta^j) = \mathbb{E}_{\theta^j}\left[\log L_c(\theta; \underline{X}, \underline{Z}) | \underline{X} = \underline{x}\right]$.
This amounts to computing $\tau_k(x_i; \theta^j)$.

M Step Maximization of $Q(\theta, \theta^j) - \text{ENT}(\theta)$ with respect to $\theta \in \Theta_K$:

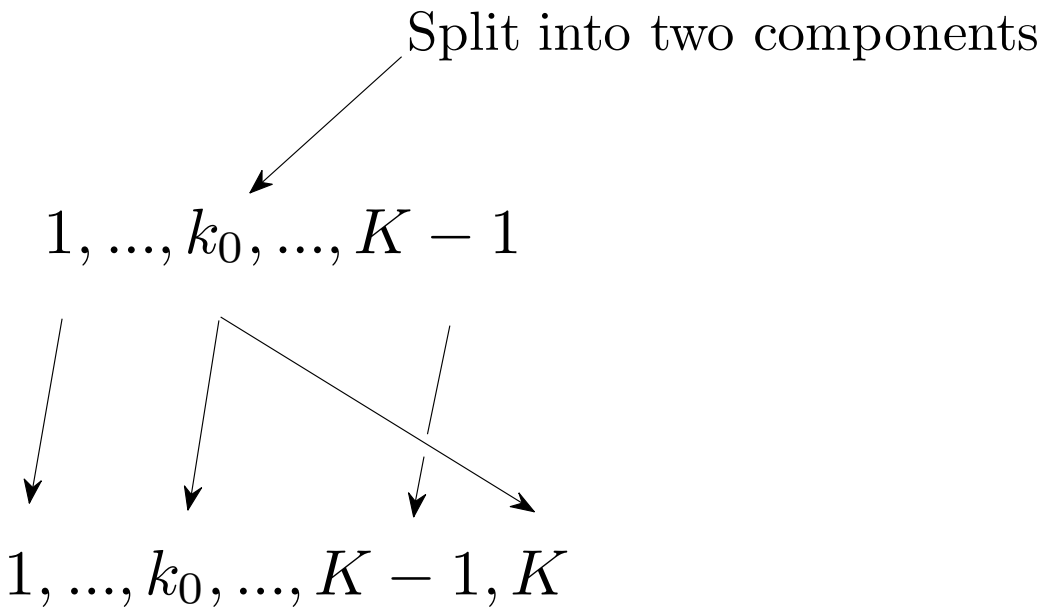
$$\theta^{j+1} \in \underset{\theta \in \Theta_K}{\text{argmax}} \left\{ \underbrace{\mathbb{E}_{\theta^j}\left[\log L_c(\theta; \underline{X}, \underline{Z}) | \underline{X} = \underline{x}\right] - \text{ENT}(\theta)}_{\log L(\theta) + \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_k(x_i;\theta^j) + \tau_k(x_i;\theta)) \log \tau_k(x_i;\theta)} \right\}$$

Computing MLccE: L_{cc}-EM

The L_{cc}-EM algorithm for MLccE:

$$\widehat{\theta}_K^{\mathsf{MLccE}} = \underset{\theta \in \Theta_K}{\mathsf{argmax}} \Big\{ \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(X_i; \omega_k) - \mathsf{ENT}(\theta) \Big\}.$$

Initialization: Random, Small_L_{cc}_EM, CEM, Km1...



Consistent Model Selection

Identification point of view:

$$K_0 = \min \operatorname*{argmin}_{1 \leq K \leq K_M} \mathbb{E}_{f^{\wp}} \left[-\log \mathrm{L}_{\mathrm{cc}}(\Theta_K^0) \right].$$

Procedures are considered such that

$$\widehat{K} = \operatorname*{argmin}_{1 \leq K \leq K_M} \left\{ -\log \mathrm{L}_{\mathrm{cc}}\left(\widehat{\theta}_K^{\mathrm{MLccE}}\right) + \mathrm{pen}(K) \right\}.$$

Consistent Model Selection

Theorem

Let us consider the model family $\left(\mathcal{M}_K\right)_{K\in\{1,\dots,K_M\}}$. Let us assume:

- $\forall K,\ \Theta_K$ is compact and convex.
- $\forall K, \forall\theta\in\Theta_K, \forall\theta^0_{K_0}\in\Theta^0_{K_0}$,

$$\mathbb{E}_{f^0}\left[-\log \mathsf{L}_{\mathsf{cc}}(\theta)\right] = \mathbb{E}_{f^0}\left[-\log \mathsf{L}_{\mathsf{cc}}(\theta^0_{K_0})\right] \Longleftrightarrow -\log \mathsf{L}_{\mathsf{cc}}(\theta;x) = -\log \mathsf{L}_{\mathsf{cc}}(\theta^0_{K_0};x) \text{ a.s.}$$

- $\forall K, H_K(x) = \sup_{\theta\in\Theta_K}\left|\log \mathsf{L}_{\mathsf{cc}}(\theta;x)\right| < \infty$ a.s. and $\|H_K\|_\infty < \infty$.
- $\forall K, H'_K(x) = \sup_{\theta\in\Theta_K}\left\|\left(\frac{\partial \log \mathsf{L}_{\mathsf{cc}}}{\partial\theta}\right)_{(\theta;x)}\right\|_\infty < \infty$ a.s. and $\|H'_K\|_2 < \infty$.
- $\forall K, \forall\theta^0_K\in\Theta^0_K, \frac{\partial^2}{\partial\theta^2}\left(\mathbb{E}_{f^0}\left[-\log \mathsf{L}_{\mathsf{cc}}(\theta)\right]\right)_{|\theta^0_K}$ is nonsingular.

Let $\mathrm{pen}:\{1,\dots,K_M\}\to\mathbb{R}^+$ such that $\begin{cases}\mathrm{pen}(K) = o_{\mathbb{P}}(n) \text{ as } n\to\infty \\ \left(\mathrm{pen}(K) - \mathrm{pen}(K')\right) \xrightarrow[n\to\infty]{\mathbb{P}} \infty \text{ if } K' < K.\end{cases}$

Then

$$\mathbb{P}\left[\widehat{K} \neq K_0\right] \xrightarrow[n\to\infty]{} 0.$$

A New Light on ICL

- Analogy with model selection criteria in the usual observed likelihood framework.
- A good identification criterion, by analogy with BIC:

$$\widehat{K}^{\mathsf{L_{cc}\text{-}ICL}} = \underset{K\in\{1,\ldots,K_M\}}{\operatorname{argmin}} \left\{ -\log \mathsf{L_{cc}}(\widehat{\theta}_K^{\mathsf{MLccE}}) + \frac{D_K}{2} \log n \right\}.$$

- ICL is an approximation of $\mathsf{L_{cc}}$-ICL.

A New Light on ICL

- Analogy with model selection criteria in the usual observed likelihood framework.
- A good identification criterion, by analogy with BIC:

$$\widehat{K}^{\mathsf{L_{cc}\text{-}ICL}} = \underset{K\in\{1,\ldots,K_M\}}{\operatorname{argmin}} \left\{ -\log \mathsf{L_{cc}}(\widehat{\theta}_K^{\mathsf{MLccE}}) + \frac{D_K}{2}\log n \right\}.$$

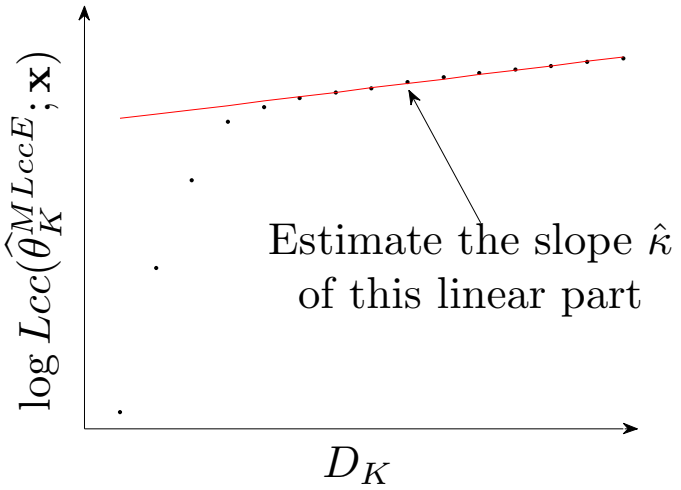
- ICL is an approximation of $\mathsf{L_{cc}}$-ICL :

$$\widehat{K}^{\mathsf{ICL}} = \underset{K\in\{1,\ldots,K_M\}}{\operatorname{argmin}} \left\{ -\log \mathsf{L_{cc}}(\widehat{\theta}_K^{\mathsf{MLE}}) + \frac{D_K}{2}\log n \right\}.$$

Slope Heuristics
(Birgé and Massart, 2006)

"*Data-driven Slope Estimation*"
Assume an "optimal" penalty is known up to a constant κ_{opt}:

$$\text{pen}(K) = \kappa_{\text{opt}} D_K.$$

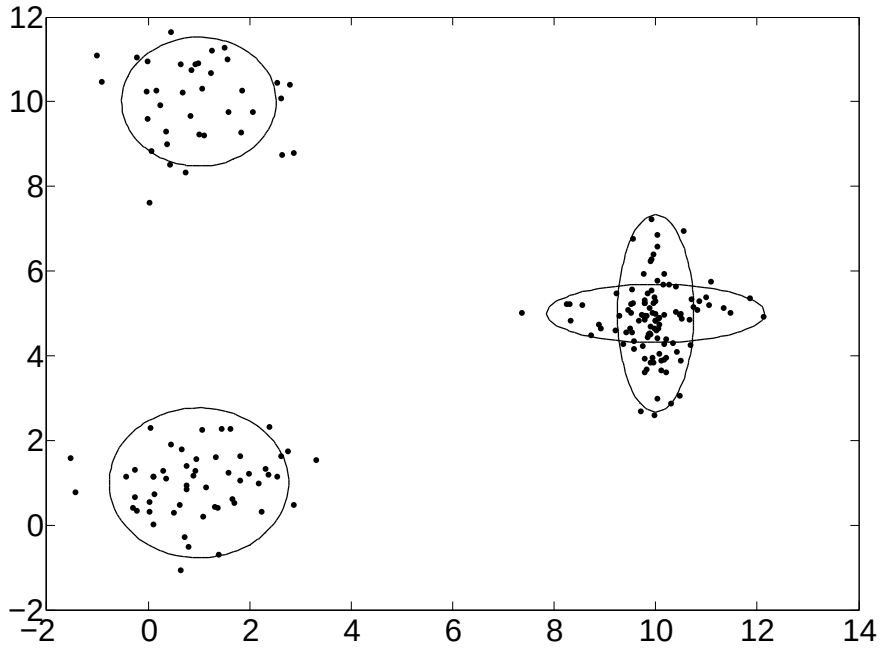


$$\widehat{K}^{\text{SHLcc}} = \underset{K \in \{1,\ldots,K_M\}}{\text{argmin}} \left\{ -\log \text{L}_{\text{cc}}(\widehat{\theta}_K^{\text{MLccE}}) + 2\widehat{\kappa} D_K \right\}$$

This data-driven procedure may be applied:

- to the usual observed likelihood contrast;
- to the conditional classification likelihood contrast.

"Cross" Dataset



- Simulated data in $\mathbb{R}^2$.
- Sample size: 200.
- Number of components: 4.
- Diagonal mixture models fitted: $f^{\wp} \in \mathcal{M}_4$.

"Cross" Dataset: Results

Selected number of components	2	3	4	5	6	7	8	9	10–20
AIC	0	0	1	1	2	2	3	3	88
BIC	0	4	91	5	0	0	0	0	0
SHL	0	2	84	10	3	0	0	0	1
ICL	0	96	3	1	0	0	0	0	0
L_{cc}-ICL	0	99	1	0	0	0	0		
SHL_{cc}	2	79	8	8	3	0	0		

Results for 100 experiments

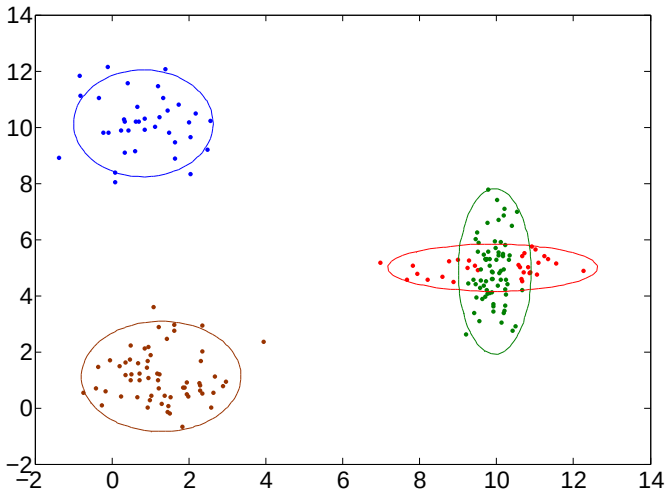
L	Risk $\times 10^3$	$\frac{\text{Risk}}{\text{Oracle Risk}}$
Oracle	59	1
AIC	506	8.03
BIC	65	1.10
(ICL)	156	2.62
SHL	69	1.17

"Oracle" number of components: 4

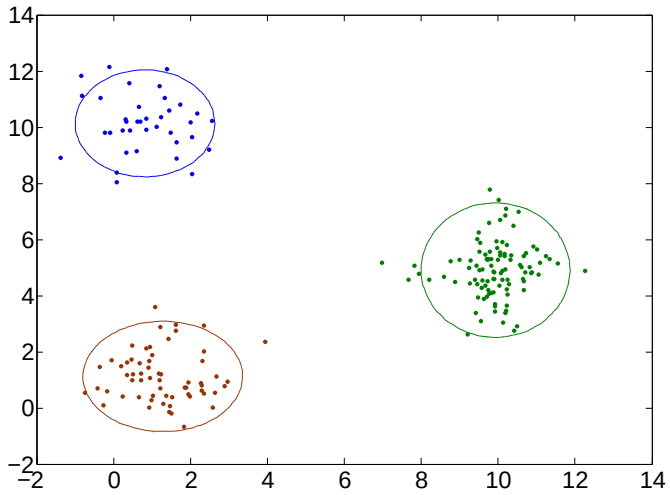
L_{cc}	"Risk" $\times 10^3$
Oracle	3618
ICL	3622
L_{cc}-ICL	3623
SHL_{cc}	3632

"Oracle" number of components : 3

Mixtures of Mixtures

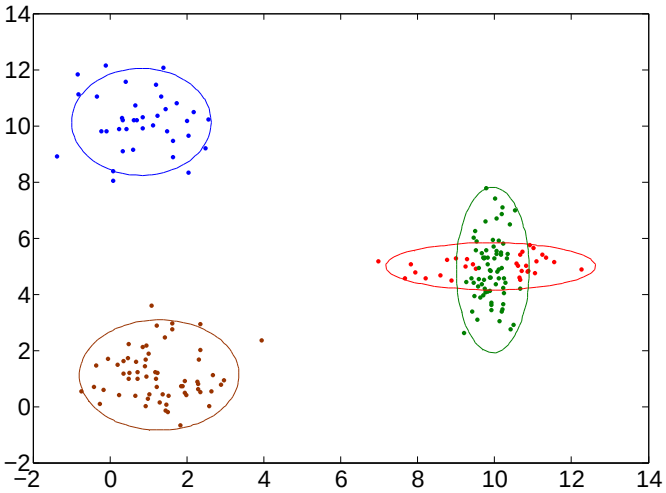


BIC Solution

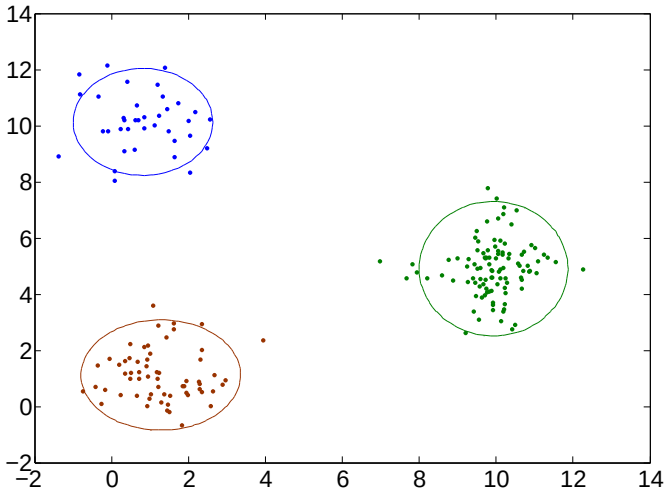


ICL Solution

Mixtures of Mixtures

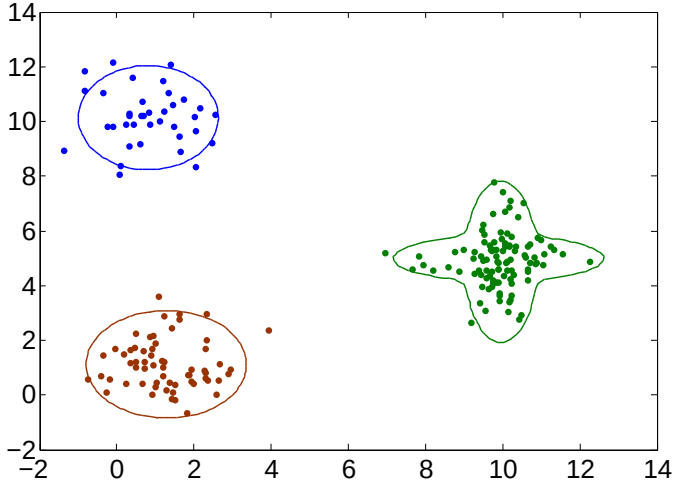


BIC Solution



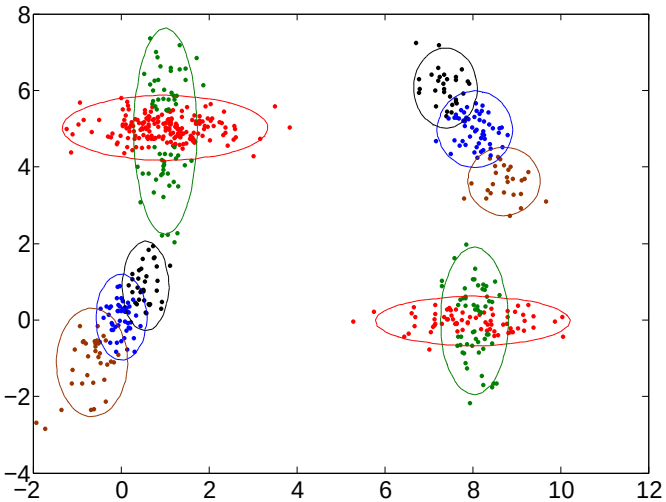
ICL Solution

Combine two components

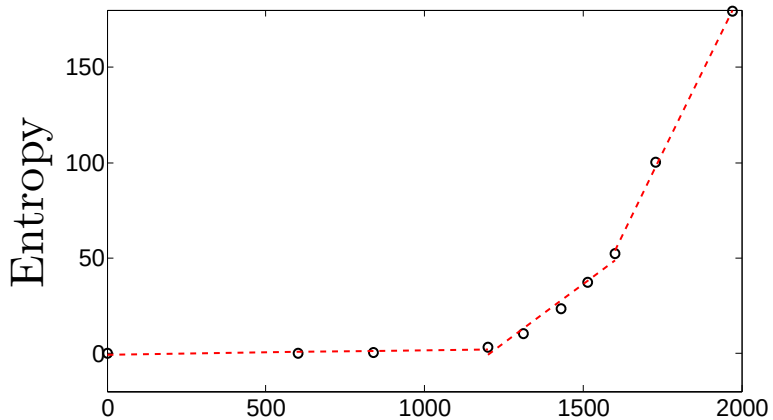


Mixtures of Mixtures

- Hierarchical classes combining:
  - From the BIC solution with $\hat{K}^{\text{BIC}}$ classes;
  - By minimizing the entropy of the combined solution at each step;
  - Until there is only one class left.
- Choosing the number of classes:
  - May be based on substantive ground;
  - The whole hierarchy may be of interest to the user;
  - The plot of the entropy against the number of classes may be helpful for the analysis;
  - Link with the works about penalized criteria?



BIC Solution



Combined solutions entropy

Conclusions and Perspectives

- The theoretical study of ICL led to the definition of a contrast adapted to the clustering objective, and thus to the corresponding estimator and model selection procedures.
- Solutions are proposed to put these into practice. They may also be applied with benefit when computing the usual MLE through the usual EM algorithm.
- A new light is thrown on ICL, viewed as an approximation of L_{cc}-ICL. This is a contribution to the study of the "class" notion in model-based clustering.
- This "class" notion may be further studied.
- The MLE and MLccE estimators on the one hand; the ICL and L_{cc}-ICL model selection criteria on the other hand, may be further compared, notably from a practical point of view.
- The slope heuristics, in this mixture models framework, may be further studied.