Université Paris 1
Panthéon - Sorbonne

Model-based clustering of high-dimensional data: an overview and some recent advances

Charles BOUVEYRON

Laboratoire SAMM, EA 4543
Université Paris 1 Panthéon-Sorbonne

This presentation is based on several works jointly done with S. Girard & G. Celeux

Outline

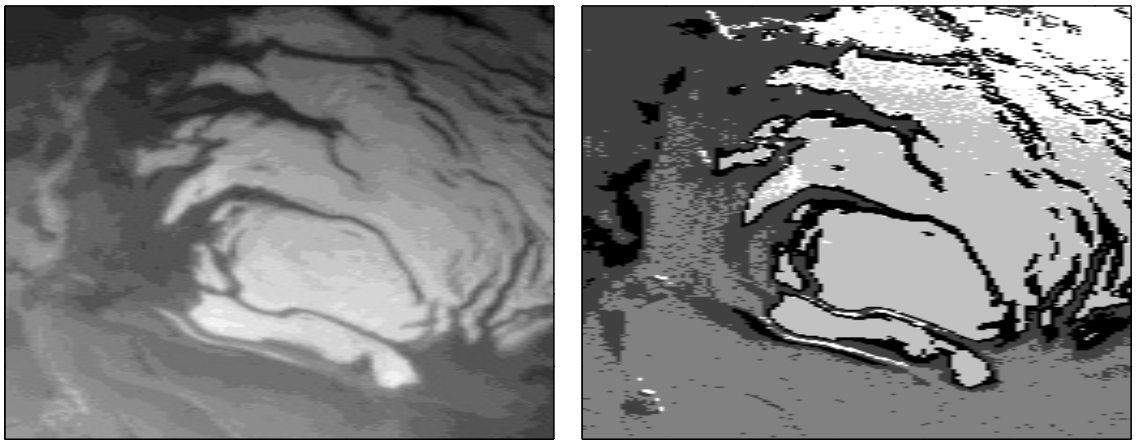
Outline

Introduction

Clustering has become a recurring problem:

- it usually occurs in all applications for which a partition is necessary (interpretation, decision, ...),
- but modern data are very often high-dimensional (p large),
- and the number of observations is sometimes small as well ($n \ll p$).

Example : segmentation of hyper-spectral images

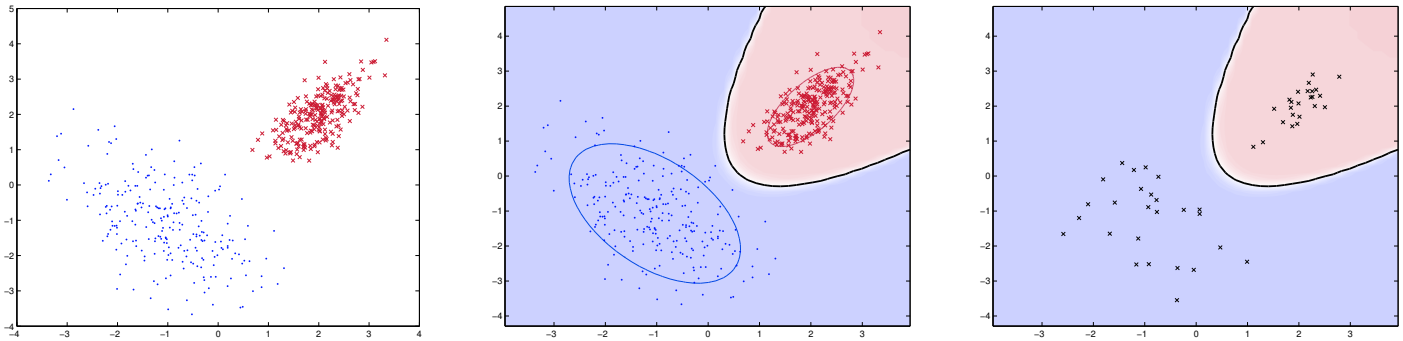


The classification problem

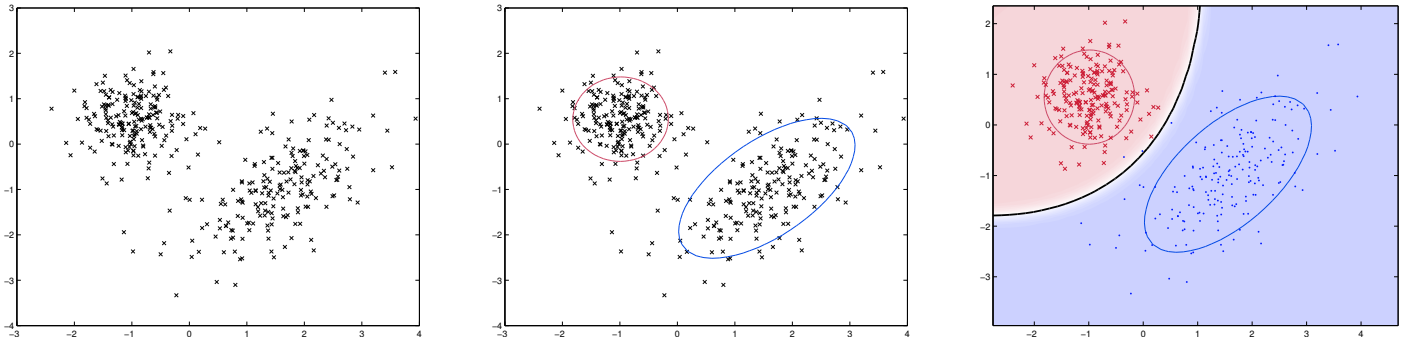
The classification problem consists in:

- organizing the observations $x_1, ..., x_n \in \mathbb{R}^p$ into K classes,
- *i.e.* associating the labels $z_1, ..., z_n \in \{1, ..., K\}$ to the data.

Supervised approach: complete dataset $(x_1, z_1), ..., (x_n, z_n)$



Non-supervised approach : only the observations $x_1, ..., x_n$



Probabilistic classification and the MAP rule

The classical probabilistic framework assumes that:

- the observations $x_1, ..., x_n$ are independant realizations of a random vector $X \in \mathcal{X}^p$,
- the labels $z_1, ..., z_n$ are independant realizations of a random variable $Z \in \{1, ..., K\}$,
- where $z_i = k$ indicates that x_i belongs to the kth class.

The classification aims to build a decision rule δ :

$$\begin{array}{rcl} \delta : \mathcal{X}^p & \rightarrow & \{1, ..., K\}, \\ x & \rightarrow & z. \end{array}$$

The optimal rule δ^* is the one which assigns x to the class with the highest posterior probability (called the MAP rule):

$$\delta^*(x) = \underset{k=1,...,K}{\operatorname{argmax}} P(Z = k | X = x).$$

Probabilistic classification and the MAP rule

The classical probabilistic framework assumes that:

- the observations $x_1, ..., x_n$ are independant realizations of a random vector $X \in \mathcal{X}^p$,
- the labels $z_1, ..., z_n$ are independant realizations of a random variable $Z \in \{1, ..., K\}$,
- where $z_i = k$ indicates that x_i belongs to the kth class.

The classification aims to build a decision rule δ :

$$\begin{aligned} \delta : \mathcal{X}^p &\rightarrow \{1, ..., K\}, \\ x &\rightarrow z. \end{aligned}$$

The optimal rule δ^* is the one which assigns x to the class with the highest posterior probability (called the MAP rule):

$$\delta^*(x) = \underset{k=1,...,K}{\operatorname{argmax}} P(Z = k | X = x).$$

Generative and discriminative approaches

The difference between both approaches:

- the way they estimate the posterior probability $P(Z|X)$
- which is used in the MAP decision rule.

Discriminative methods:

- they directly model the posterior probability $P(Z|X)$,
- by building a boundary between the classes.

Generative methods:

- they first model the joint distribution $P(X, Z)$,
- and then deduce the posterior probability using the Bayes' rule:

$$P(Z|X) = \frac{P(X,Z)}{P(X)} \propto P(Z)P(X|Z).$$

Generative and discriminative approaches

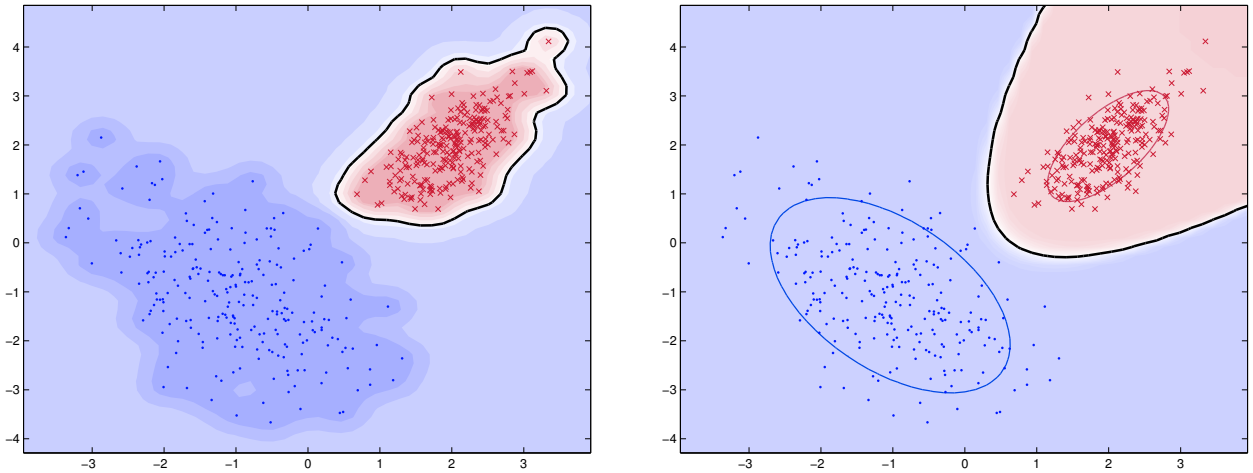


Fig. Discriminative (left) and generative (right) methods.

Discriminative methods:

- logistic regression (it models $log(f_1(x)/f_2(x))$),
- Support Vector Machines (SVM), decision trees, ...

Generative methods:

- mainly, model-based classification methods,
- but it exists also non-parametric methods.

The mixture model

The mixture model:

- the observations $x_1, ..., x_n$ are assumed to be independant realizations of a random vector $X \in \mathcal{X}^p$ with a density:

$$f(x) = \sum_{k=1}^{K} \pi_k f(x, \theta_k),$$

  - K is the number of classes,
  - π_k are the mixture proportions,
  - $f(x, \theta_k)$ is a probability density with its parameters θ_k.

The Gaussian mixture model:

- among all mixture models, the Gaussian mixture model is certainly the most used in the classification context,
- in this case, $f(x, \theta_k)$ is the Gaussian density $\mathcal{N}(\mu_k, \Sigma_k)$ with $\theta_k = \{\mu_k, \Sigma_k\}$.

The mixture model

The MAP decision rule becomes in the mixture model framework:

$$\begin{aligned}\delta^*(x) &= \underset{k=1,\dots,K}{\operatorname{argmax}} P(Z=k|X=x),\\ &= \underset{k=1,\dots,K}{\operatorname{argmax}} P(Z=k)P(X=x|Z=k),\\ &= \underset{k=1,\dots,K}{\operatorname{argmin}} H_k(x),\end{aligned}$$

where H_k is defined by $H_k(x) = -2\log(\pi_k f(x,\theta_k))$.

The building of the decision rule consists in:

1. estimate the parameters θ_k of the mixture model,
2. calculate the value of $H_k(x)$ for each new observation x.

Gaussian mixtures for classification

Gaussian model Full-GMM (QDA in discrimination):

$$H_k(x) = (x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k) + \log(\det \Sigma_k) - 2\log(\pi_k) + C^{st}.$$

Gaussian model Com-GMM which assumes that $\forall k,\ \Sigma_k = \Sigma$ (LDA in discrimination):

$$H_k(x) = \mu_k^t \Sigma^{-1} \mu_k - 2\mu_k^t \Sigma^{-1} x - 2\log(\pi_k) + C^{st}.$$

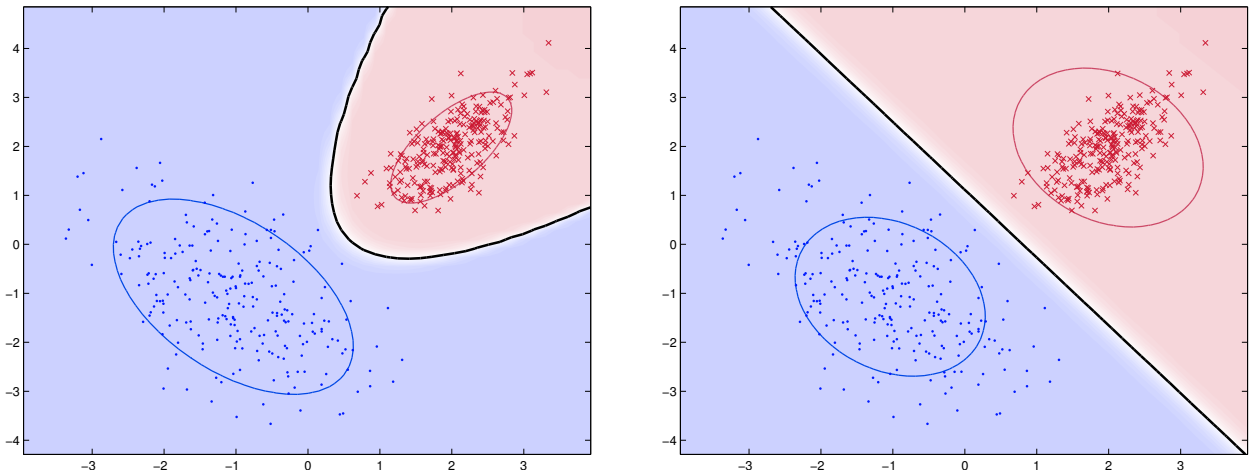


Fig. Decision boundaries for Full-GMM (left) and Com-GMM (right).

The curse of dimensionality

The curse of dimensionality:

- this term was first used by R. Bellman in the introduction of his book "Dynamic programming" in 1957:

  All [problems due to high dimension] *may be subsumed under the heading "the curse of dimensionality". Since this is a curse,* [...], *there is no need to feel discouraged about the possibility of obtaining significant results despite it.*

- he used this term to talk about the difficulties to find an optimum in a high-dimensional space using an exhaustive search,
- in order to promotate dynamic approaches in programming.

The curse of dimensionality

In the mixture model context:

- the building of the data partition mainly depends on:

$$H_k(x) = -2\log(\pi_k f(x, \theta_k)),$$

- model Full-GMM:

$$H_k(x) = (x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k) + \log(\det \Sigma_k) - 2\log(\pi_k) + \gamma.$$

- model Com-GMM which assumes that $\forall k,\ \Sigma_k = \Sigma$:

$$H_k(x) = \mu_k^t \Sigma^{-1}\mu_k - 2\mu_k^t \Sigma^{-1} x - 2\log(\pi_k) + \gamma.$$

Important remarks :

- it is necessary to invert Σ_k or Σ,
- and this will cause big difficulties in certain cases!

The curse of dimensionality

In the mixture model context:

- the number of parameters grows up with p^2,

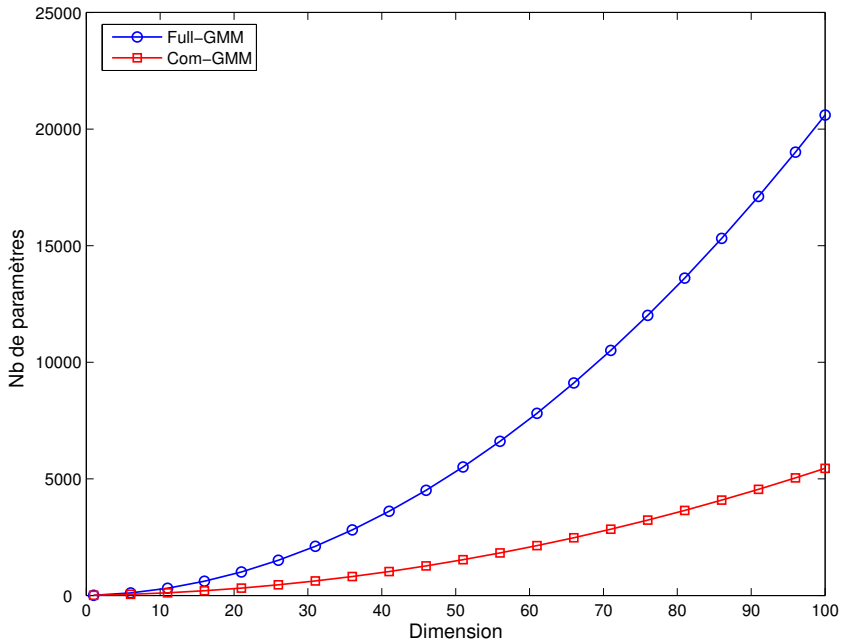


Fig. Number of parameters to estimate for the models Full-GMM and Com-GMM regarding to the dimension and with $k = 4$***.***

- if n is small compared to p^2, the estimates of Σ_k are ill-conditionned or singular,
- it is therefore difficult or impossible to invert Σ_k.

The blessings of dimensionality

As Bellman thought:

- all is not bad in high-dimensional spaces (hopefully!)
- there are interesting things which happen in high-dimensional spaces.

The empty-space phenomenum [Scott83]:

- classical thoughts true in 1, 2 or 3-dimensional spaces are in fact wrong in higher dimensions,
- particularly, high-dimensional spaces are almost empty!

The blessings of dimensionality

First example : the volume of a sphere

$$V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)},$$

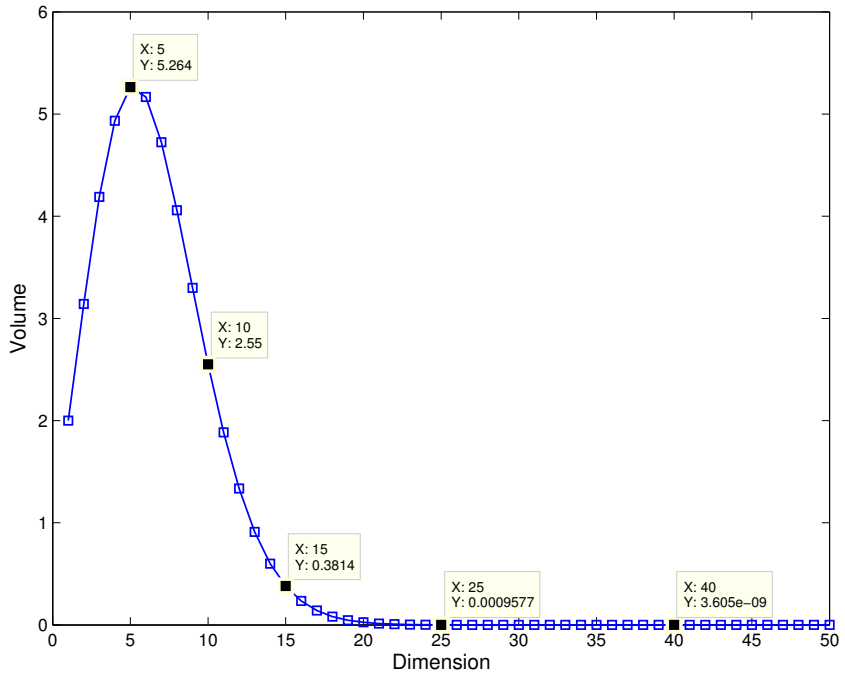


Fig. Volume of a sphere of radius 1 regarding to the dimension p.

The blessings of dimensionality

Second example:

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier.

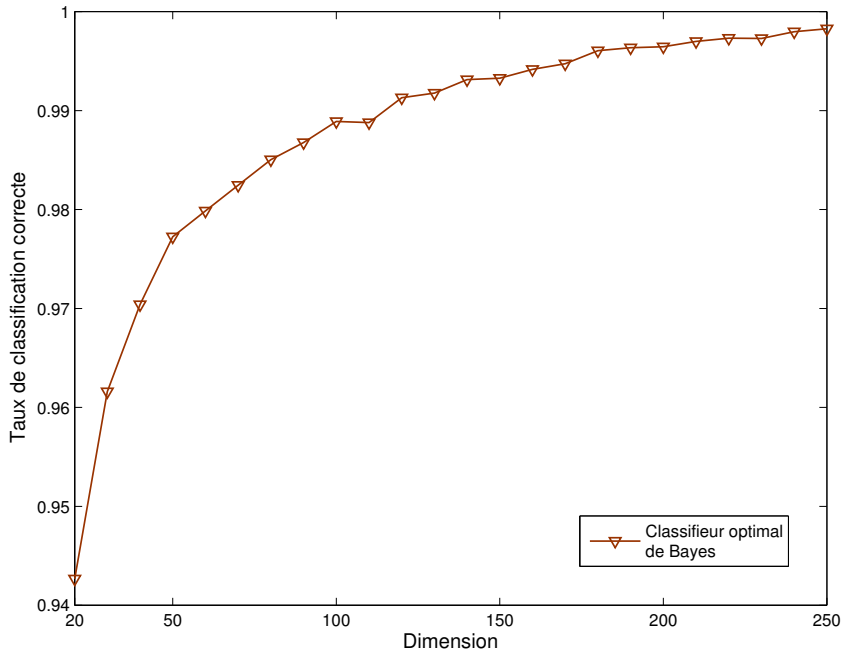


Fig. Correct classification rate of the optimal classifier versus the data dimension on simulated data.

Outline

Classical ways to avoid the curse of dimensionality

Dimension reduction:

- the problem comes from that p is too large,
- therefore, reduce the data dimension to $d \ll p$,
- such that the curse of dimensionality vanishes!

Parsimonious models:

- the problem comes from that the number of parameters to estimate is too large,
- therefore, make additional assumptions to the model,
- such that the number of parameters to estimate becomes more "decent"!

Regularization:

- the problem comes from that parameter estimates are instable,
- therefore, regularize these estimates,
- such that the parameter are correctly estimated!

Dimension reduction

Linear dimension reduction methods:

- feature combination: PCA,
- feature selection: ...

Non linear dimension reduction methods:

- Kohonen algorithms, Self Organising Maps,
- LLE, Isomap, ...
- Kernel PCA, principal curves, ...

Supervised dimension reduction methods:

- the old fashion method: Fisher Discriminant Analysis (FDA),
- many recent works on this topic... but useless in our context!

Parsimonious models

Parsimonious models:

- can be obtained by making additional assumptions on the original model
- in order to adapt the model to the available data.

Parsimonious Gaussian models:

- com-GMM:
  - the assumption: $\Sigma_k = \Sigma$,
  - nb of par. for $K = 4$ and $p = 100$: 5453
- diag-GMM:
  - the assumption: $\Sigma_k = \mathrm{diag}(\sigma_{k1}, ..., \sigma_{kp})$,
  - nb of par. for $K = 4$ and $p = 100$: 803
- sphe-GMM:
  - the assumption: $\Sigma_k = \sigma_k I_p$,
  - nb of par. for $K = 4$ and $p = 100$: 407

Parsimonious models

A family of parsimonious Gaussian models [Banfield93, Celeux95]:

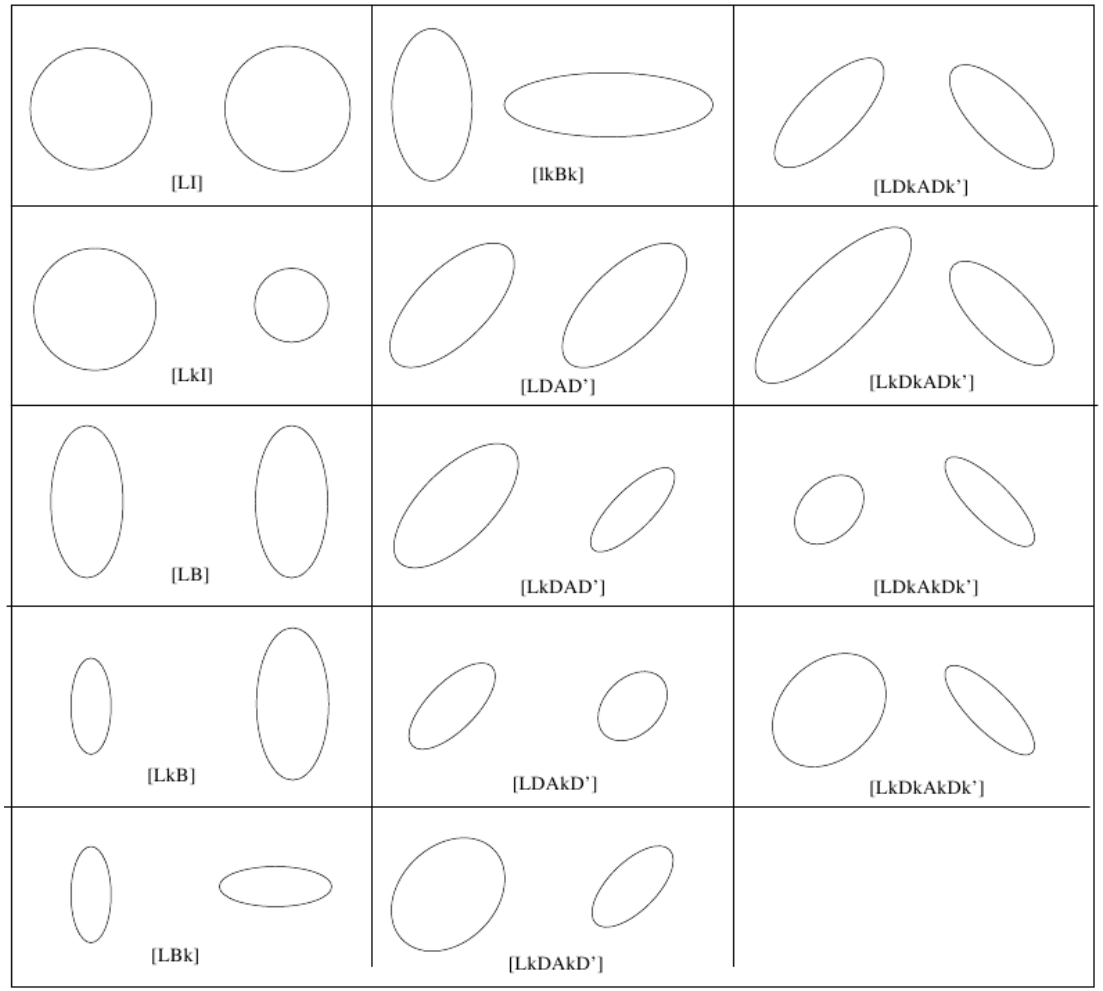


Fig. The family of 14 parsimonious Gaussian models [Celeux95].

Regularization

Regularization of the covariance matrix estimates:

- ridge-like regularization: $\tilde{\Sigma}_k = \hat{\Sigma}_k + \sigma_k I_p$,
- PDA [Hast95] : $\tilde{\Sigma}_k = \hat{\Sigma}_k + \sigma_k \Omega$,
- RDA [Frie89] proposed a regularized classifier which varies between a quadratic and a linear classifier:

$$\tilde{\Sigma}_k(\lambda, \gamma) = (1-\gamma) S_k(\lambda) + \gamma \left(\frac{tr(S_k(\lambda))}{p} \right) I_p$$

where S_k is defined by:

$$S_k(\lambda) = \frac{(n_k - 1)(1-\lambda)\hat{\Sigma}_k + (n-K)\lambda\hat{\Sigma}}{(1-\lambda)(n_k-1) + \lambda(n-K)}.$$

Regularization

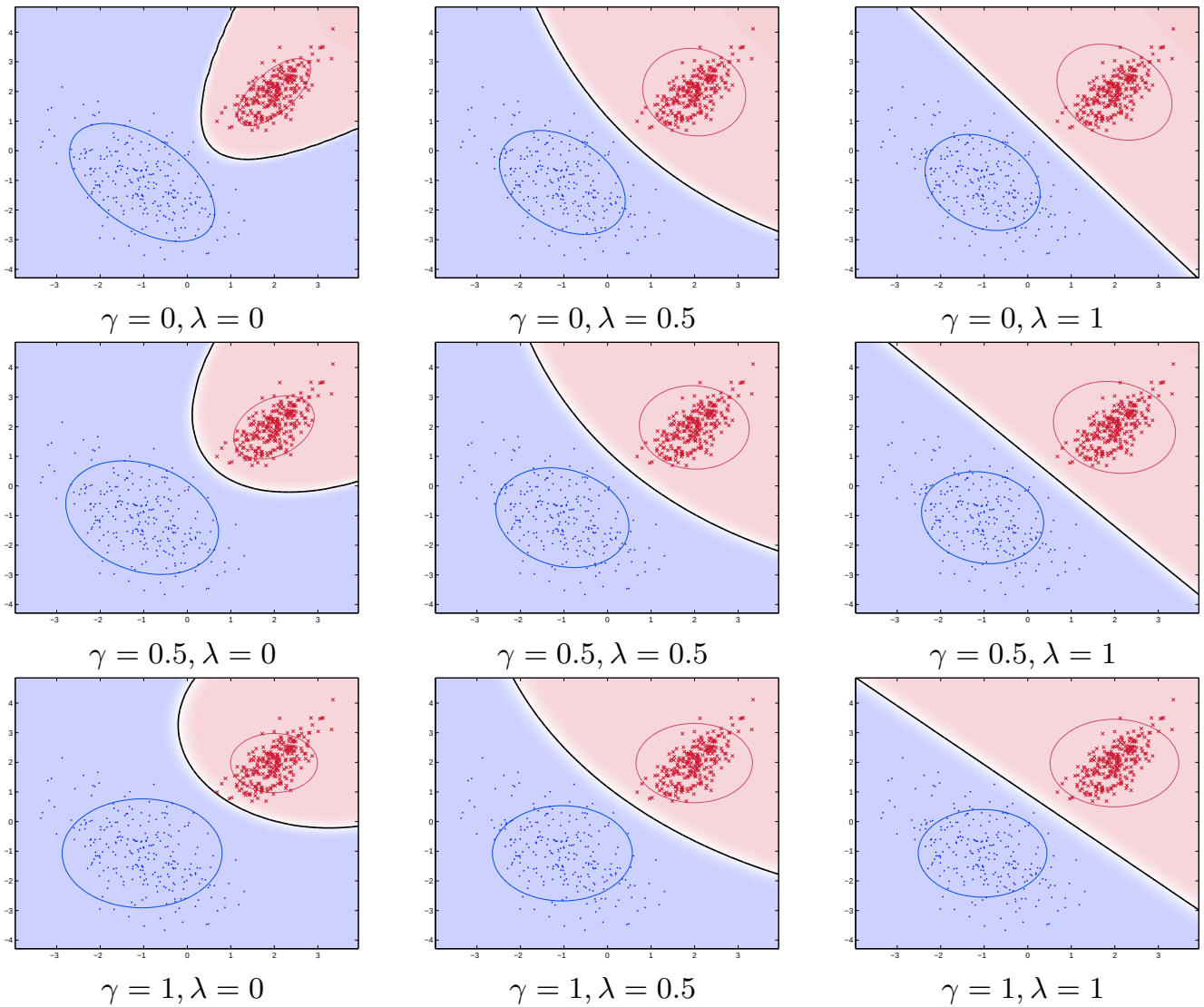


Fig. 4. Influence des paramètres γ et λ sur le classifieur RDA.

Outline

Subspace clustering methods

Recent approaches propose:

- to model the data of each group in specific subspaces,
- to keep all dimensions in order to facilitate the discrimination of the groups.

Several works on this topic in the last years:

- mixture of factor analyzers: Ghahramani *et al.* (1996) and McLachlan *et al.* (2003),
- mixture of probabilistic PCA: Tipping & Bishop (1999) ,
- mixture of HD Gaussian models: Bouveyron & Girard (2007),
- mixture of parsimonious FA: McNicholas and Murphy (2008),
- mixture of common FA: Beak *et al.* (2009).

Subspace clustering methods

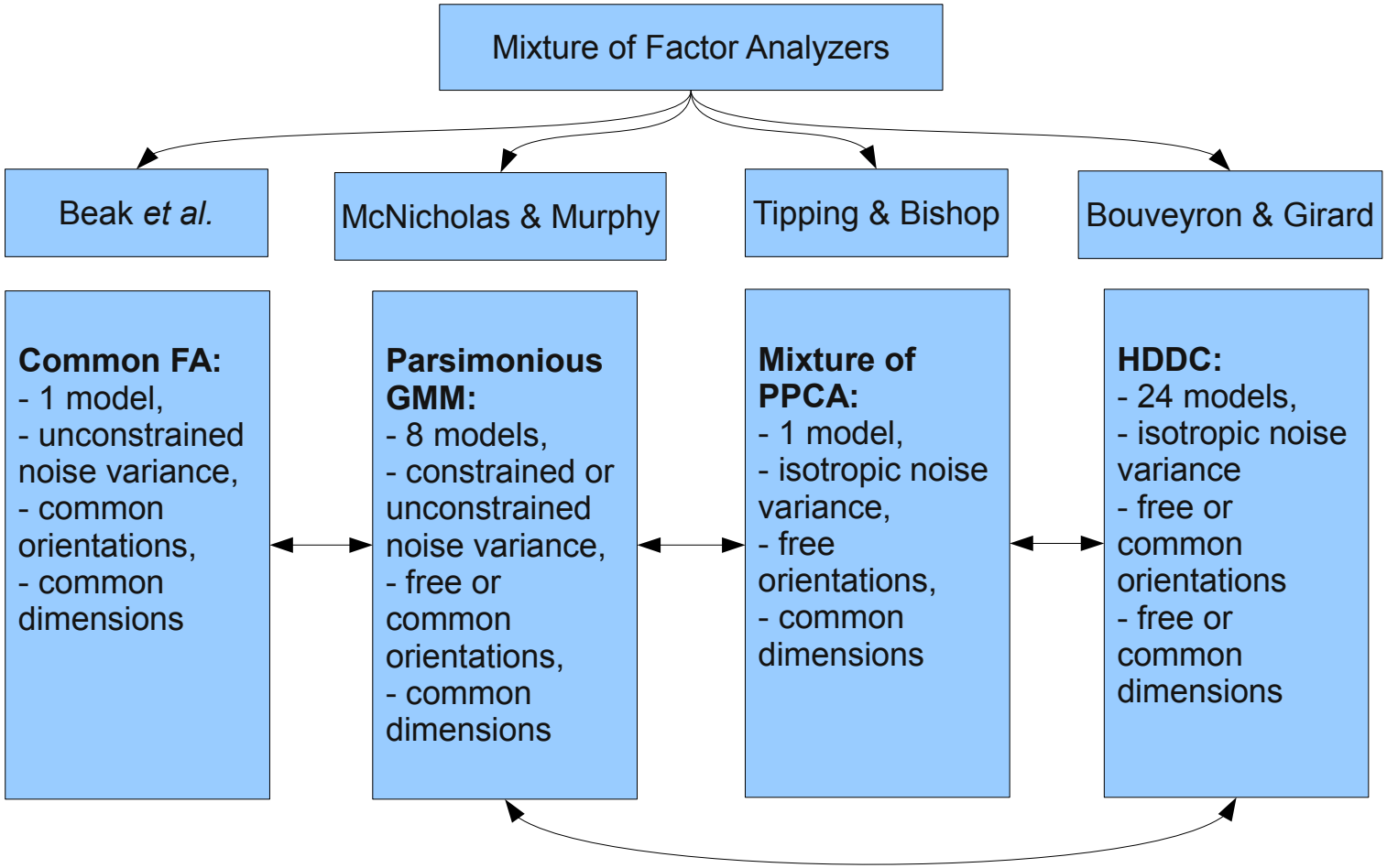


Figure: A tentative family tree of subspace clustering methods.

The model $[a_{kj}b_kQ_kd_k]$

Bouveyron & Girard (2007) proposed to consider the Gaussian mixture model:

$$f(x) = \sum_{k=1}^{K} \pi_k f(x, \theta_k),$$

where $\theta_k = \{\mu_k, \Sigma_k\}$ for each $k = 1, ..., K$.

Based on the spectral decomposition of Σ_k, we can write:

$$\Sigma_k = Q_k \, \Delta_k \, Q_k^t,$$

where:

- Q_k is an orthogonal matrix containing the eigenvectors of Σ_k,
- Δ_k is diagonal matrix containing the eigenvalues of Σ_k.

The model $[a_{kj}b_kQ_kd_k]$

We assume that Δ_k has the following form:

$$\Delta_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kd_k} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (p - d_k) \end{matrix}$$

where:

- $a_{kj} \geq b_k$, for $j = 1, ..., d_k$ and $k = 1, ..., K$,
- and $d_k < p$, for $k = 1, ..., K$.

The model $[a_{kj}b_kQ_kd_k]$

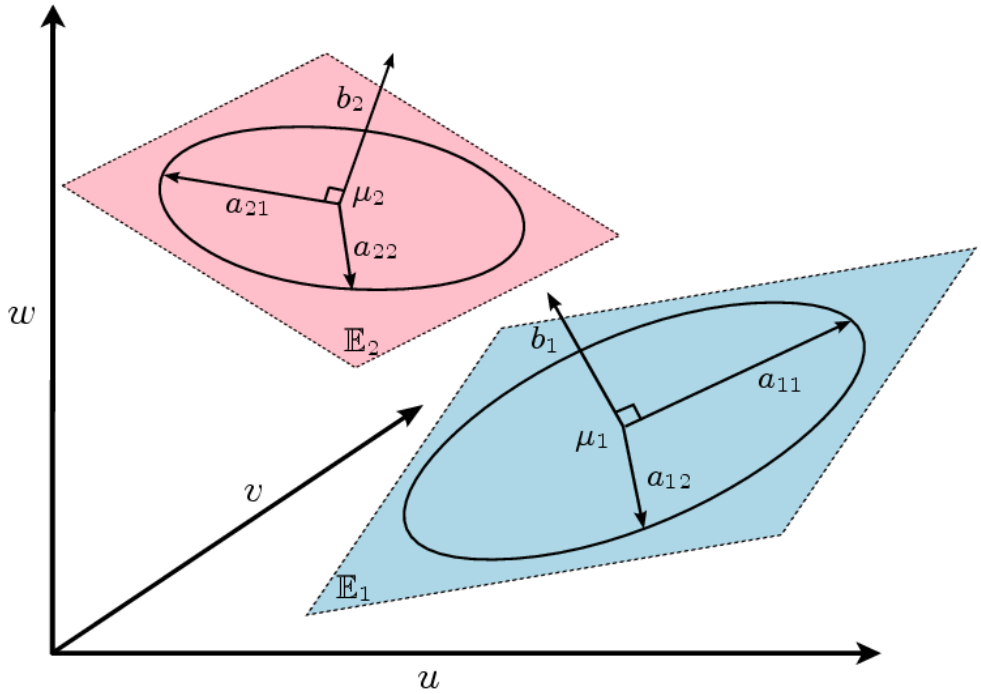


Fig. The subspace $\mathbb{E}_k$ and its supplementary $\mathbb{E}_k^{\perp}$.

We also define:

- the affine space $\mathbb{E}_k$ generated by eigenvectors associated to the eigenvalues a_{kj} and such that $\mu_k \in \mathbb{E}_k$,
- the affine space $\mathbb{E}_k^{\perp}$ such that $\mathbb{E}_k \oplus \mathbb{E}_k^{\perp} = \mathbb{R}^p$ and $\mu_k \in \mathbb{E}_k^{\perp}$,
- the projectors P_k and $P_k^{\perp}$ respectively on $\mathbb{E}_k$ and $\mathbb{E}_k^{\perp}$.

The model $[a_{kj}b_kQ_kd_k]$ and its submodels

We thus obtain a re-parameterization of the Gaussian model:

- which depends on a_{kj}, b_k, Q_k and d_k,
- the model complexity is controlled by the subspace dimensions.

We obtain increasingly regularized models:

- by fixing some parameters to be common within or between the classes,
- from the most complex model to the simplest model.

Our family of GMM contains 28 models and can be splitted into three branches:

- 14 models with free orientations,
- 12 models with common orientations,
- 2 models with common covariance matrices.

The model $[a_{kj}b_kQ_kd_k]$ and its submodels

Model	Number of parameters	Asymptotic order	Nb of prms $k=4$, $d=10$, $p=100$	ML estimation
$[a_{ij}b_iQ_id_i]$	$\rho+\bar{\tau}+2k+D$	kpd	4231	CF
$[a_{ij}bQ_id_i]$	$\rho+\bar{\tau}+k+D+1$	kpd	4228	CF
$[a_ib_iQ_id_i]$	$\rho+\bar{\tau}+3k$	kpd	4195	CF
$[ab_iQ_id_i]$	$\rho+\bar{\tau}+2k+1$	kpd	4192	CF
$[a_ibQ_id_i]$	$\rho+\bar{\tau}+2k+1$	kpd	4192	CF
$[abQ_id_i]$	$\rho+\bar{\tau}+k+2$	kpd	4189	CF
$[a_{ij}b_iQ_id]$	$\rho+k(\tau+d+1)+1$	kpd	4228	CF
$[a_jb_iQ_id]$	$\rho+k(\tau+1)+d+1$	kpd	4198	CF
$[a_{ij}bQ_id]$	$\rho+k(\tau+d)+2$	kpd	4225	CF
$[a_jbQ_id]$	$\rho+k\tau+d+2$	kpd	4195	CF
$[a_ib_iQ_id]$	$\rho+k(\tau+2)+1$	kpd	4192	CF
$[ab_iQ_id]$	$\rho+k(\tau+1)+2$	kpd	4189	CF
$[a_ibQ_id]$	$\rho+k(\tau+1)+2$	kpd	4189	CF
$[abQ_id]$	$\rho+k\tau+3$	kpd	4186	CF
$[a_{ij}b_iQd_i]$	$\rho+\tau+D+2k$	pd	1396	FG
$[a_{ij}bQd_i]$	$\rho+\tau+D+k+1$	pd	1393	FG
$[a_ib_iQd_i]$	$\rho+\tau+3k$	pd	1360	FG
$[a_ibQd_i]$	$\rho+\tau+2k+1$	pd	1357	FG
$[ab_iQd_i]$	$\rho+\tau+2k+1$	pd	1357	FG
$[abQd_i]$	$\rho+\tau+k+2$	pd	1354	FG
$[a_{ij}b_iQd]$	$\rho+\tau+kd+k+1$	pd	1393	FG
$[a_jb_iQd]$	$\rho+\tau+k+d+1$	pd	1363	FG
$[a_{ij}bQd]$	$\rho+\tau+kd+2$	pd	1390	FG
$[a_ib_iQd]$	$\rho+\tau+2k+1$	pd	1357	IP
$[ab_iQd]$	$\rho+\tau+k+2$	pd	1354	IP
$[a_ibQd]$	$\rho+\tau+k+2$	pd	1354	IP
$[a_jbQd]$	$\rho+\tau+d+2$	pd	1360	CF
$[abQd]$	$\rho+\tau+3$	pd	1351	CF
Full-GMM	$\rho+kp(p+1)/2$	$kp^2/2$	20603	CF
Com-GMM	$\rho+p(p+1)/2$	$p^2/2$	5453	CF
Diag-GMM	$\rho+kp$	$2kp$	803	CF
Sphe-GMM	$\rho+k$	kp	407	CF

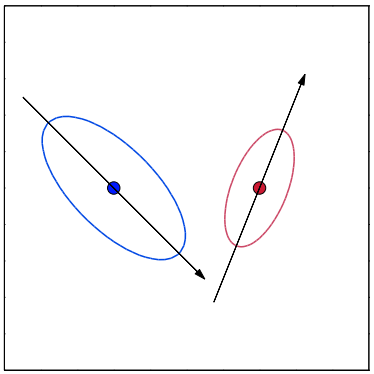
Table: *Properties of the sub-models of* $[a_{kj}b_kQ_kd_k]$

The model $[a_{kj}b_kQ_kd_k]$ and its submodels

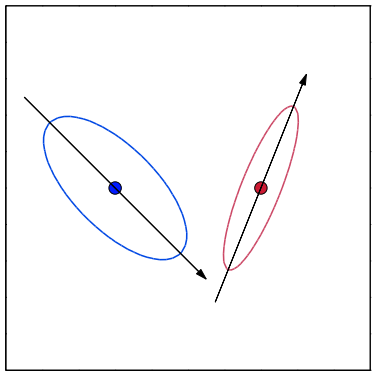
Model	Nb of prms, $K = 4$ $d = 10$, $p = 100$	Classifier type
$[a_{kj}b_kQ_kd_k]$	4231	Quadratic
$[a_{kj}b_kQd_k]$	1396	Quadratic
$[a_jbQd]$	1360	Linear
Full-GMM	20603	Quadratic
Com-GMM	5453	Linear

Table. Properties of the sub-models of $[a_{kj}b_kQ_kd_k]$

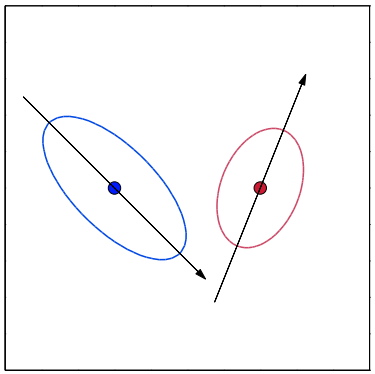
The model $[a_{kj}b_kQ_kd_k]$ and its submodels



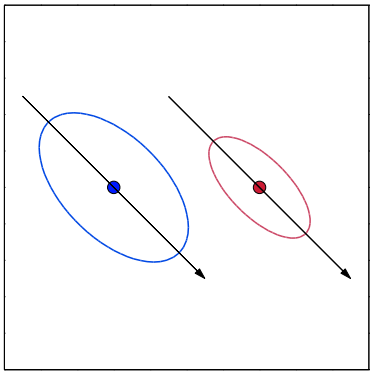
model $[a_kb_kQ_kd]$



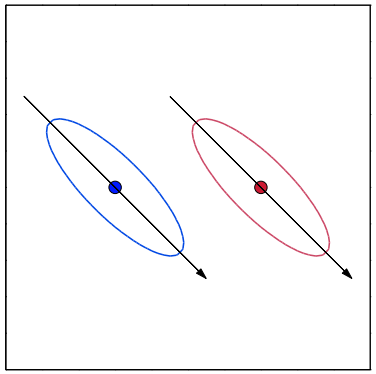
model $[ab_kQ_kd]$



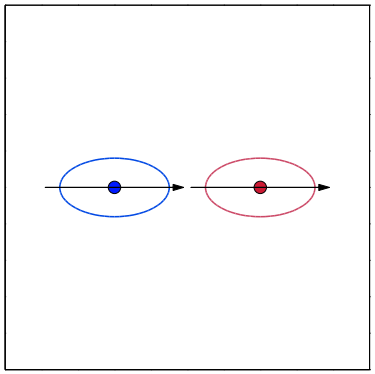
model $[a_kbQ_kd]$



model $[a_kb_kQd]$



model $[abQd]$



model $[abI_2d]$

Fig. Influence of parameters a_k, b_k et Q_k on the densities of 2 classes in dimension 2 with $d_1 = d_2 = 1$.

Construction of the classifier

In the supervised context:

- the classifier has been named HDDA,
- the estimation of parameters is direct since we have complete data,
- parameters are estimated by maximum likelihood.

In the unsupervised context:

- the classifier has been named HDDC,
- the estimation of parameters is not direct since we do not have complete data,
- parameters are estimated through a EM algorithm which iteratively maximizes the likelihood.

HDDC: the E step

In the case of the model $[a_k b_k Q_k d_k]$:

$$H_k(x) = \frac{1}{a_k}\|\mu_k - P_k(x)\|^2 + \frac{1}{b_k}\|x - P_k(x)\|^2 + d_k \log(a_k) + (p - d_k)\log(b_k) - 2\log(\pi_k).$$

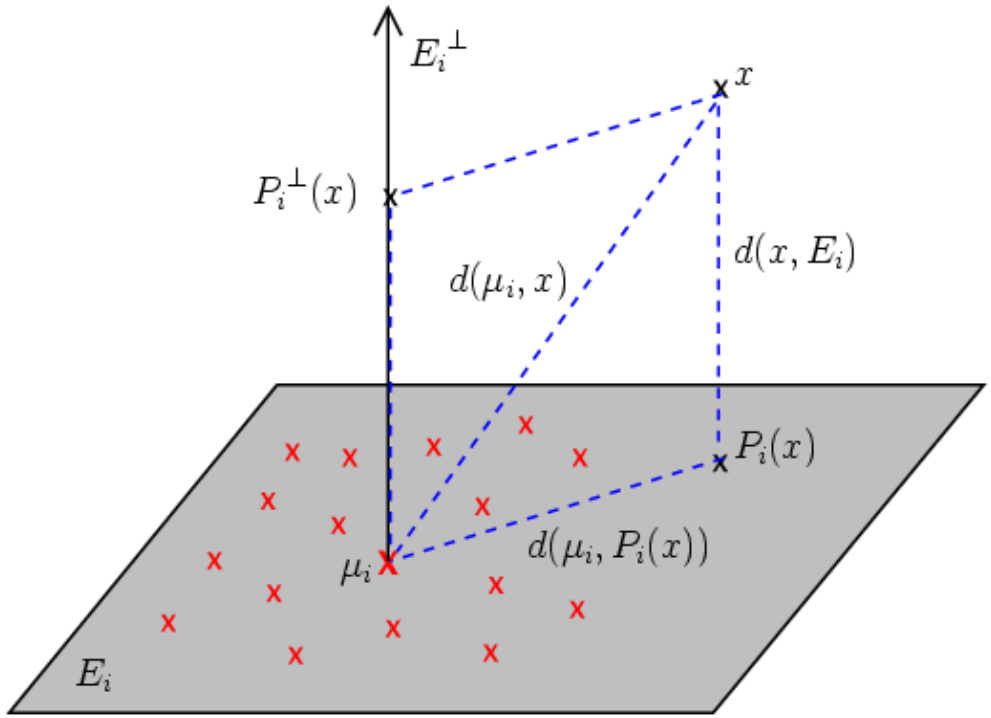


Fig. The subspaces $\mathbb{E}_k$ and $\mathbb{E}_k^{\perp}$ of the kth mixture composant.

HDDC: the M step

The ML estimators for the model $[a_{kj}b_kQ_kd_k]$ are closed forms:

- Subspace $\mathbb{E}_k$: the d_k first columns of Q_k are estimated by the eigenvectors associated to the d_k largest eigenvalues λ_{kj} of the empirical covariance matrix W_k of the kth class.

- Estimator of a_{kj}: the parameters a_{kj} are estimated by the d_k largest eigenvalues λ_{kj} of W_k.

- Estimator of b_k: the parameter of b_k is estimated by:

$$\hat{b}_k = \frac{1}{(p-d_k)}\left(\text{trace}(W_k) - \sum_{j=1}^{d_k}\lambda_{kj}\right).$$

HDDC: hyper-parameter estimation

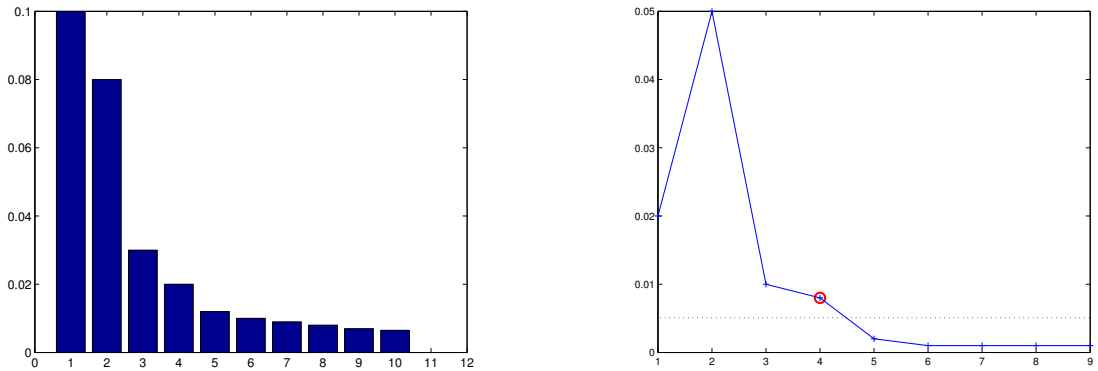


Fig. The scree-test of Cattell based on the eigenvalue scree.

Estimation of the intrinsic dimensions d_k :

- we use the *scree-test* of Cattell [Catt66],
- it allows to estimate the K parameters d_k in a common way.

Estimation of the nomber of groups K :

- in the supervised context, K is known,
- in the unsupervised context, K is chosen using BIC.

Numerical considerations

- Numerical stability : the decision rule of HDDC does not depend on the eigenvectors associated with the smallest eigenvalues of W_k.

- Reduction of computing time : there is no need to compute the last eigenvectors of $W_k \rightarrow$ reduction of computing time with a designed procedure ($\times 60$ for $p = 1000$).

- Particular case $n < p$: from a numerical point of view, it is better to compute the eigenvectors of $\bar{X}_k \bar{X}_k^t$ instead of $W_k = \bar{X}_k^t \bar{X}_k$ ($\times 500$ for $n = 13$ and $p = 1000$).

Effect of the dimensionality

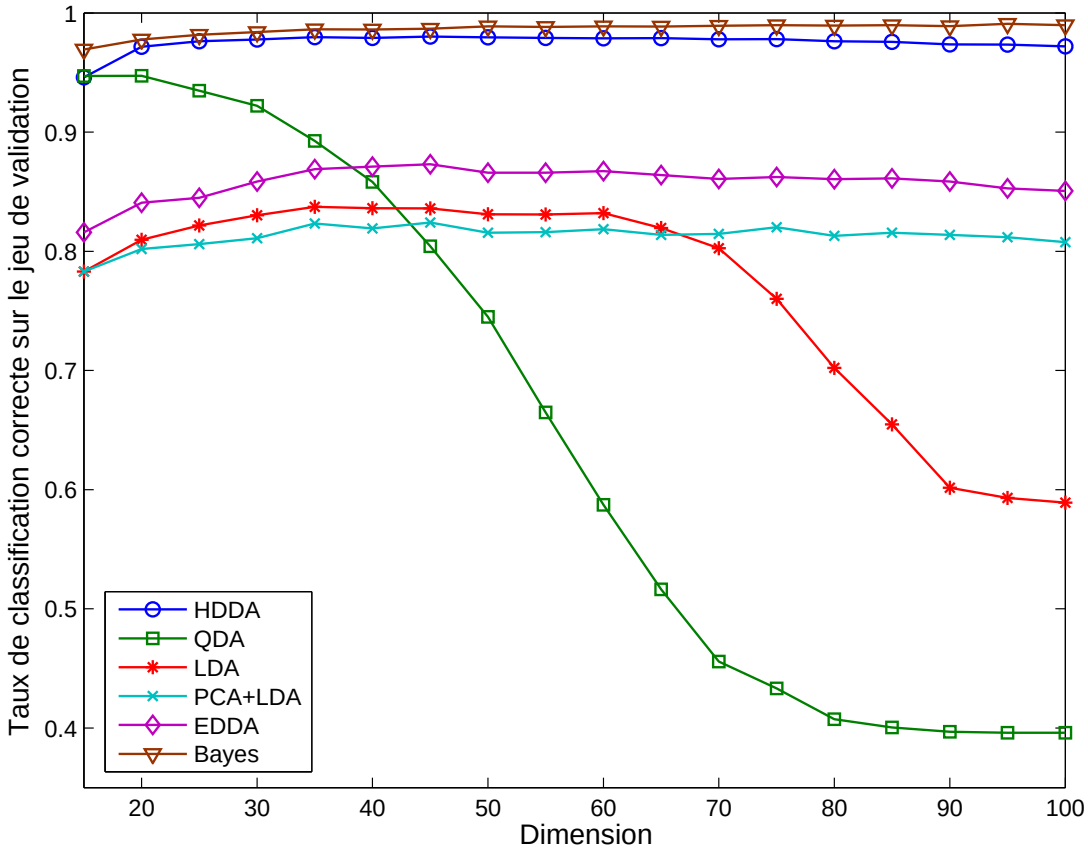


Fig. Correct classification rate versus data dimension (simulated data according to $[a_{ij}b_iQ_id_i]$***).***

Estimation of intrinsic dimensions

Nb of classes k	Chosen threshold s	Dimensions d_i	BIC value
2	0.18	2,16	414
3	0.21	**2,5,10**	407
4	0.25	2,2,5,10	414
5	0.28	2,5,5,10,12	416
6	0.28	2,5,6,10,10,12	424

Table. Selection of discrete parameters using BIC on simulated data where d_i are equal to 2, 5 and 10.

Comparison with variable selection

Model	On original features	With a dimension reduction step (ACP)
Sphe-GMM	0.340	0.340
Diag-GMM	0.355	0.535
Com-GMM	0.625	0.635
Full-GMM	0.640	0.845
VS-GMM [Raft05]	0.925	/
HDDC $[a_i b_i Q_i d_i]$	**0.950**	/

Table. Correct classification rate on a real dataset: Crabs $\in \mathbb{R}^5$***.***

HDDC: an EM-based algorithm

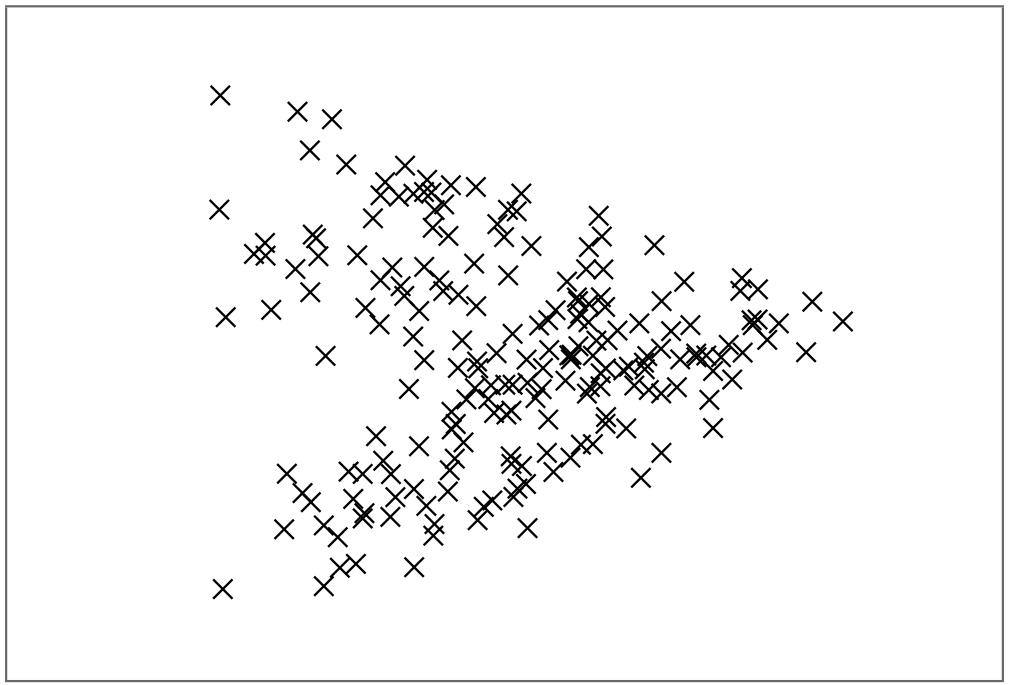


Fig. Projection of the «Crabs» data on the first principal axes.

«Crabs» data:

- 200 observations in a 5-dimensional space (5 morphological features),
- 4 classes: BM, BF, OM and OF.

HDDC: an EM-based algorithm

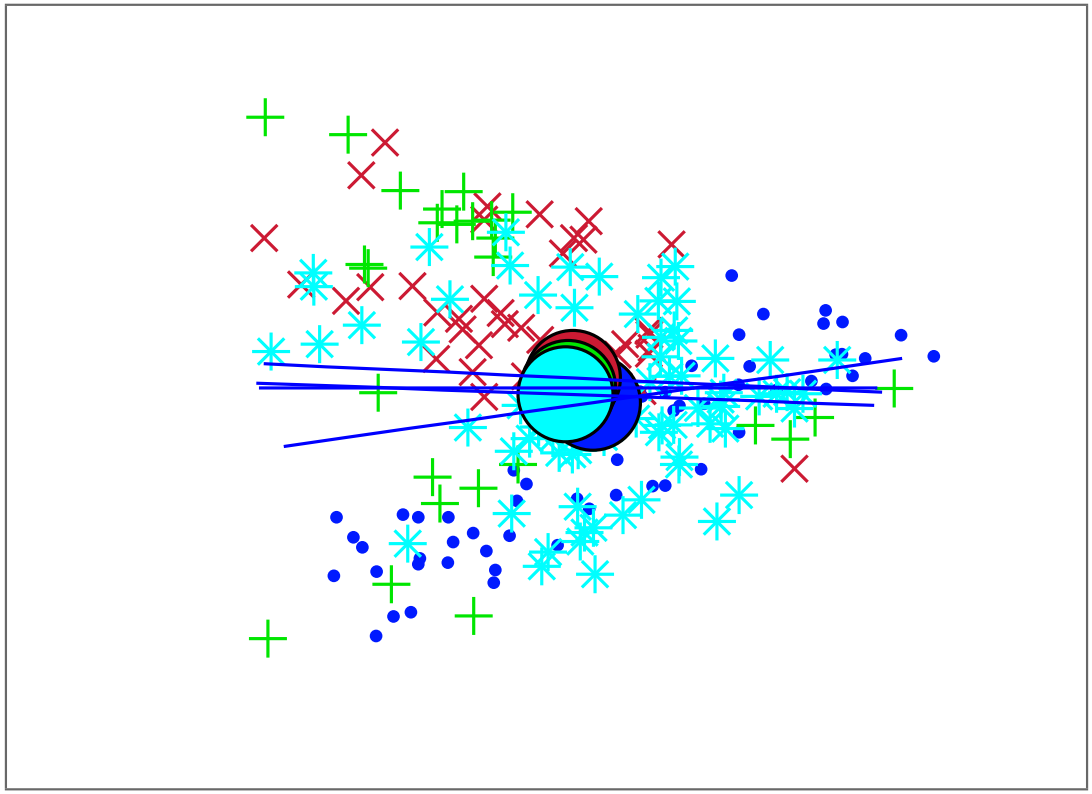


Fig. Step n° 1 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

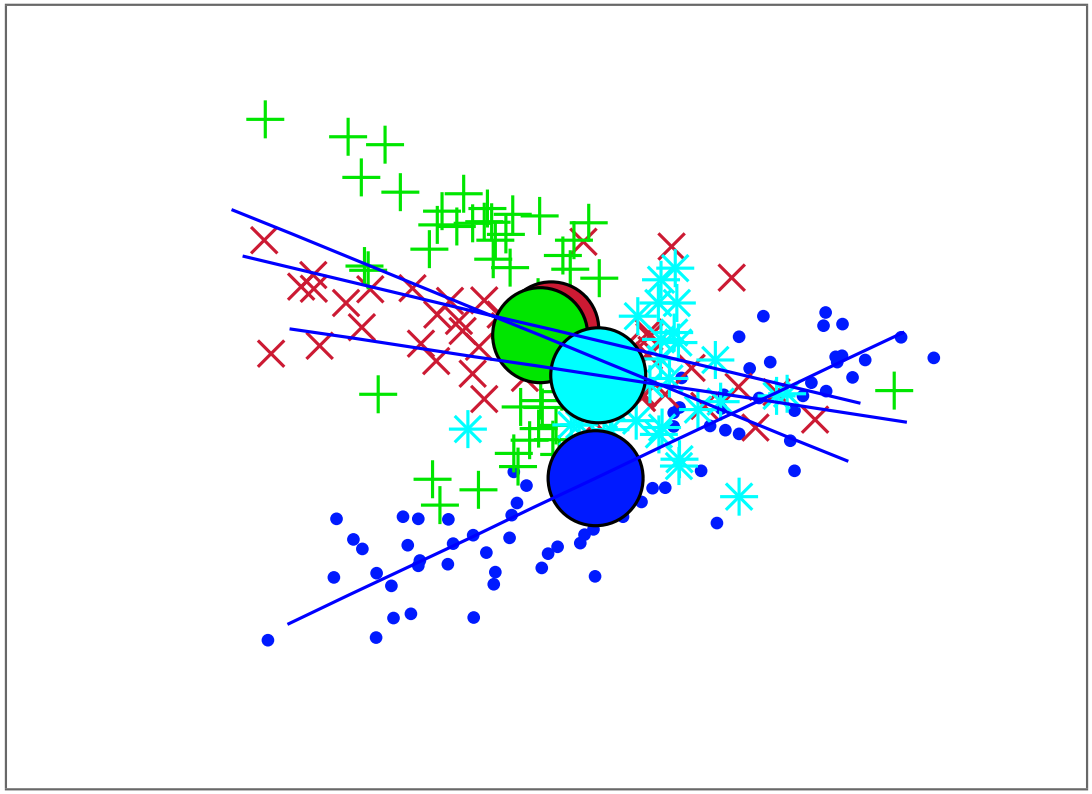


Fig. Step n° 4 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

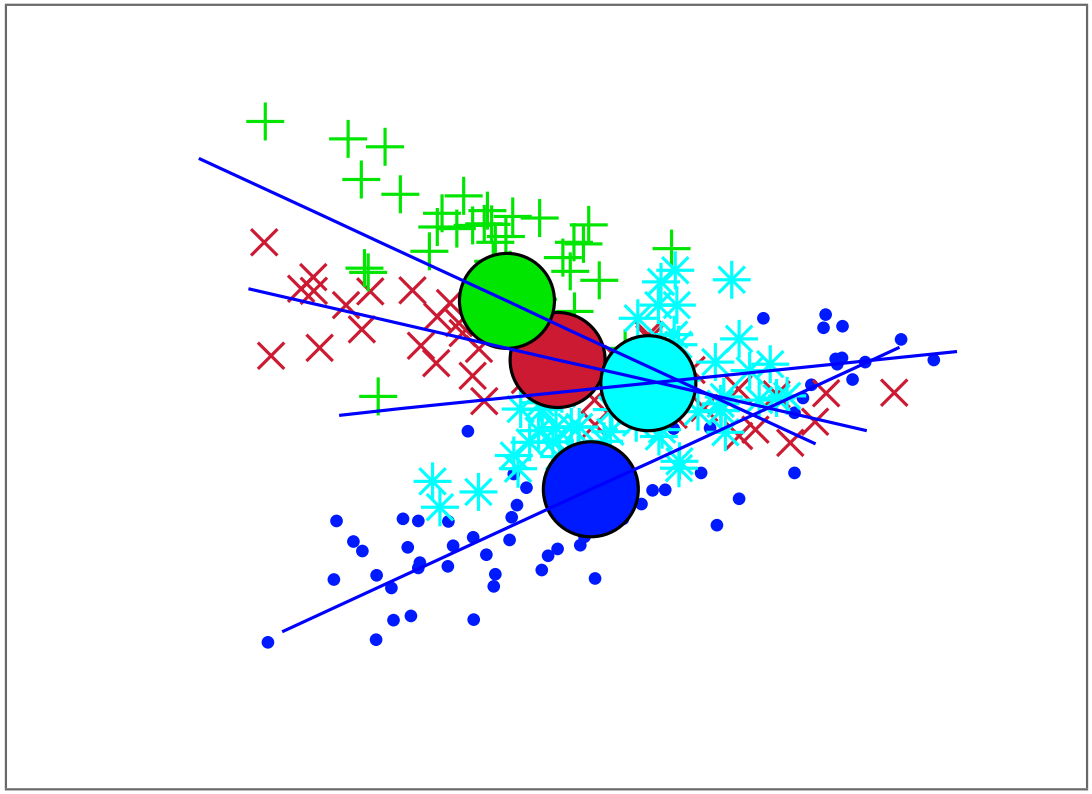


Fig. Step n° 7 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

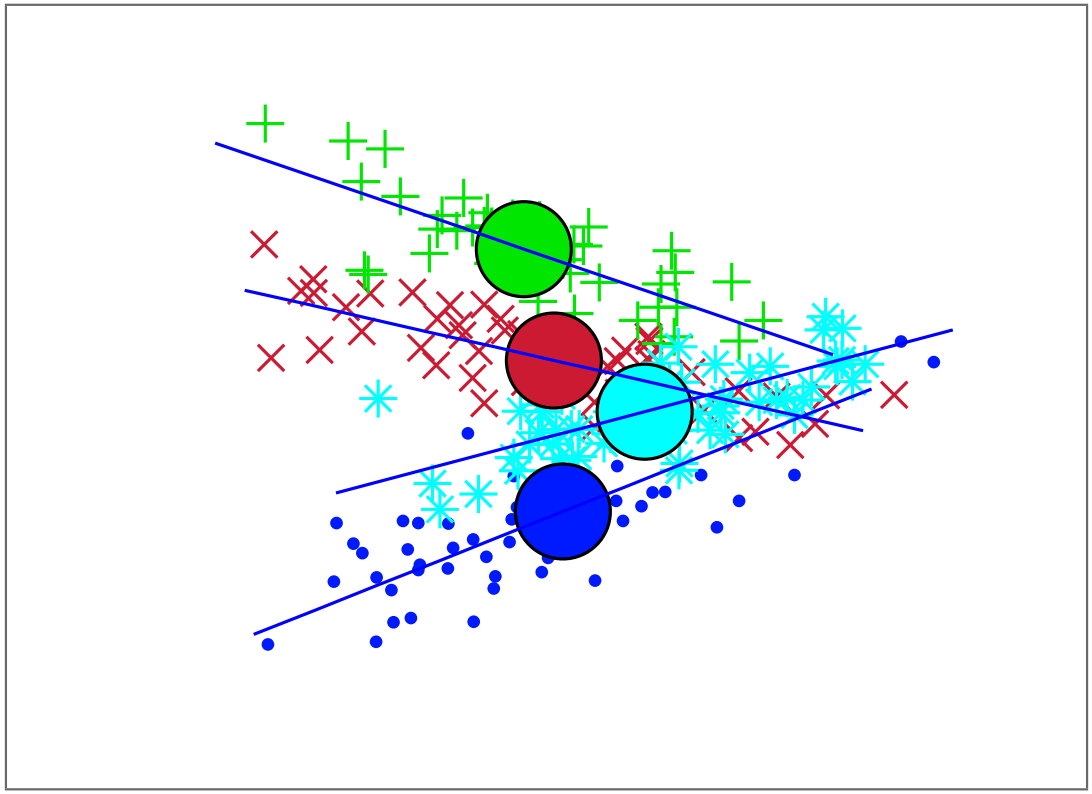


Fig. Step n° 10 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

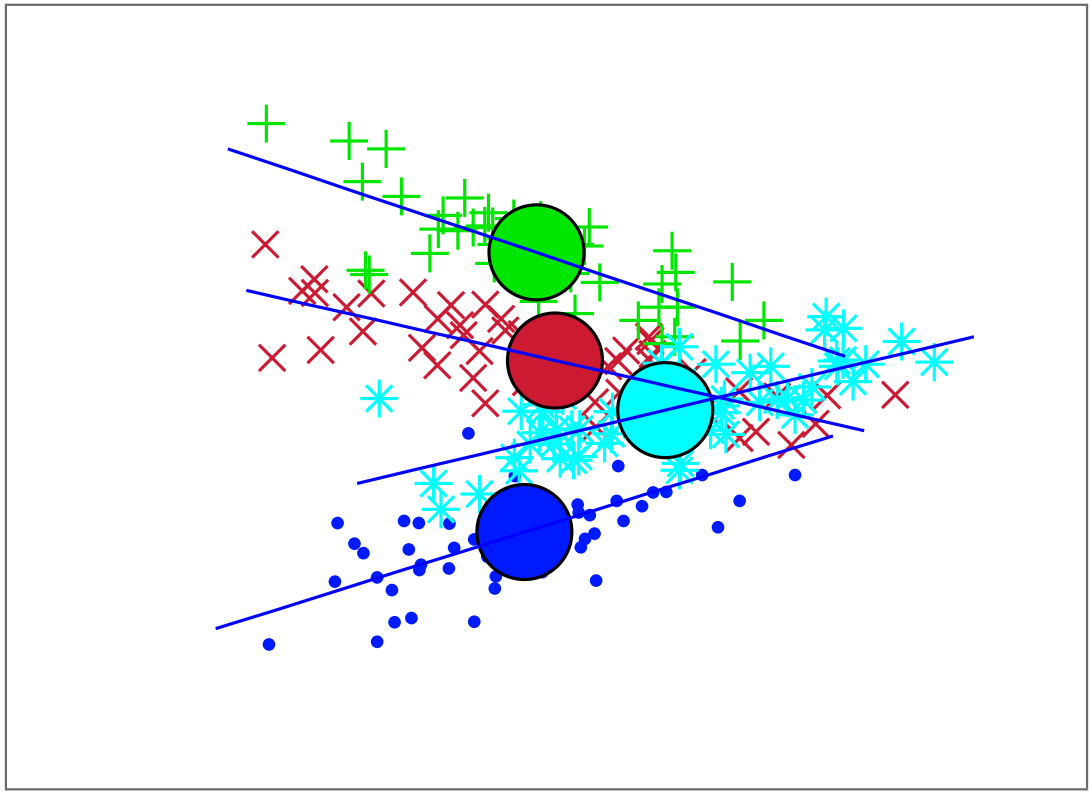


Fig. Step n° 12 of HDDC on the «Crabs» data.

Categorization of the Martian surface

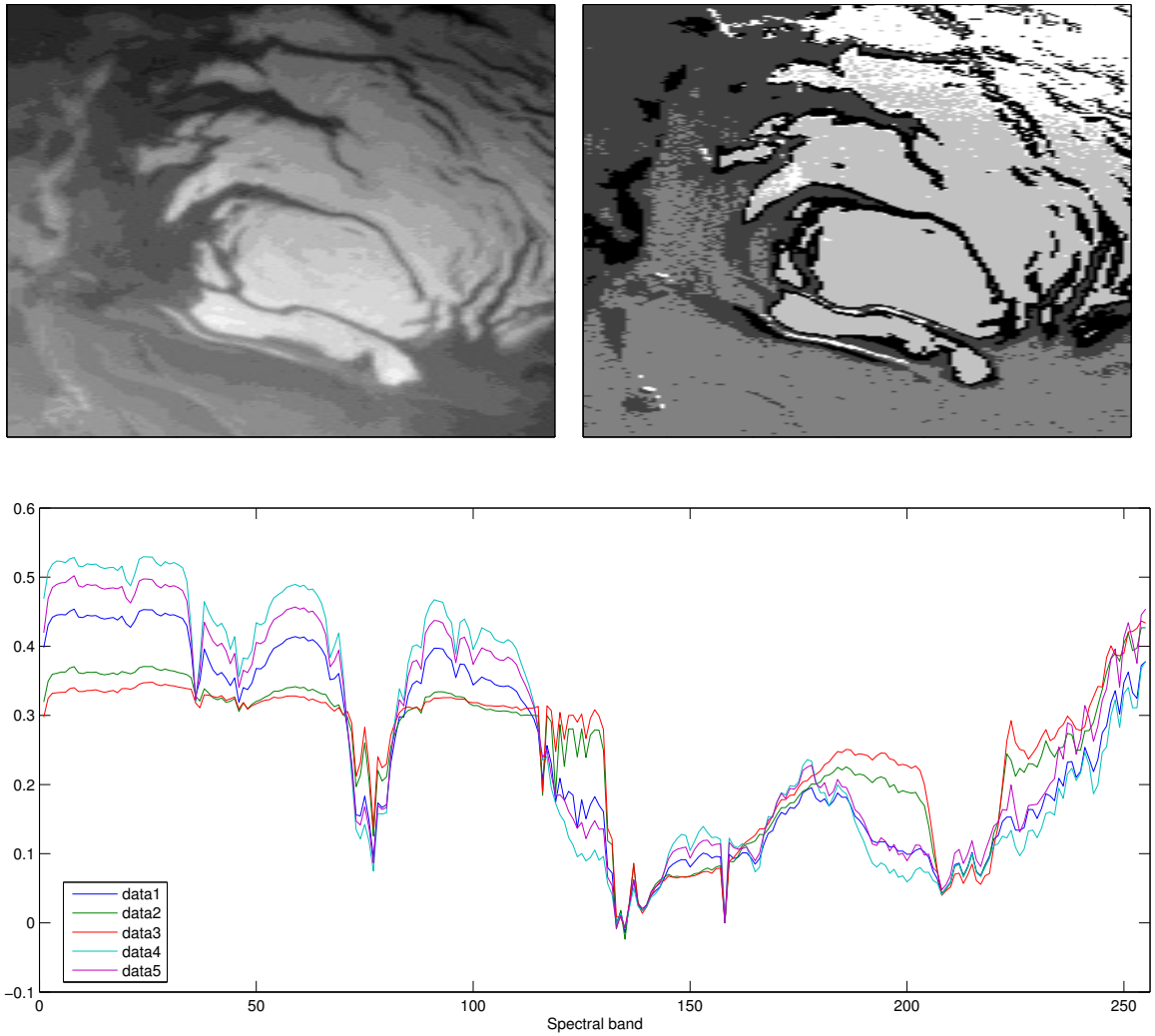


Fig. Categorization of the Martian surface based on HD spectral images.

Object localization in natural images

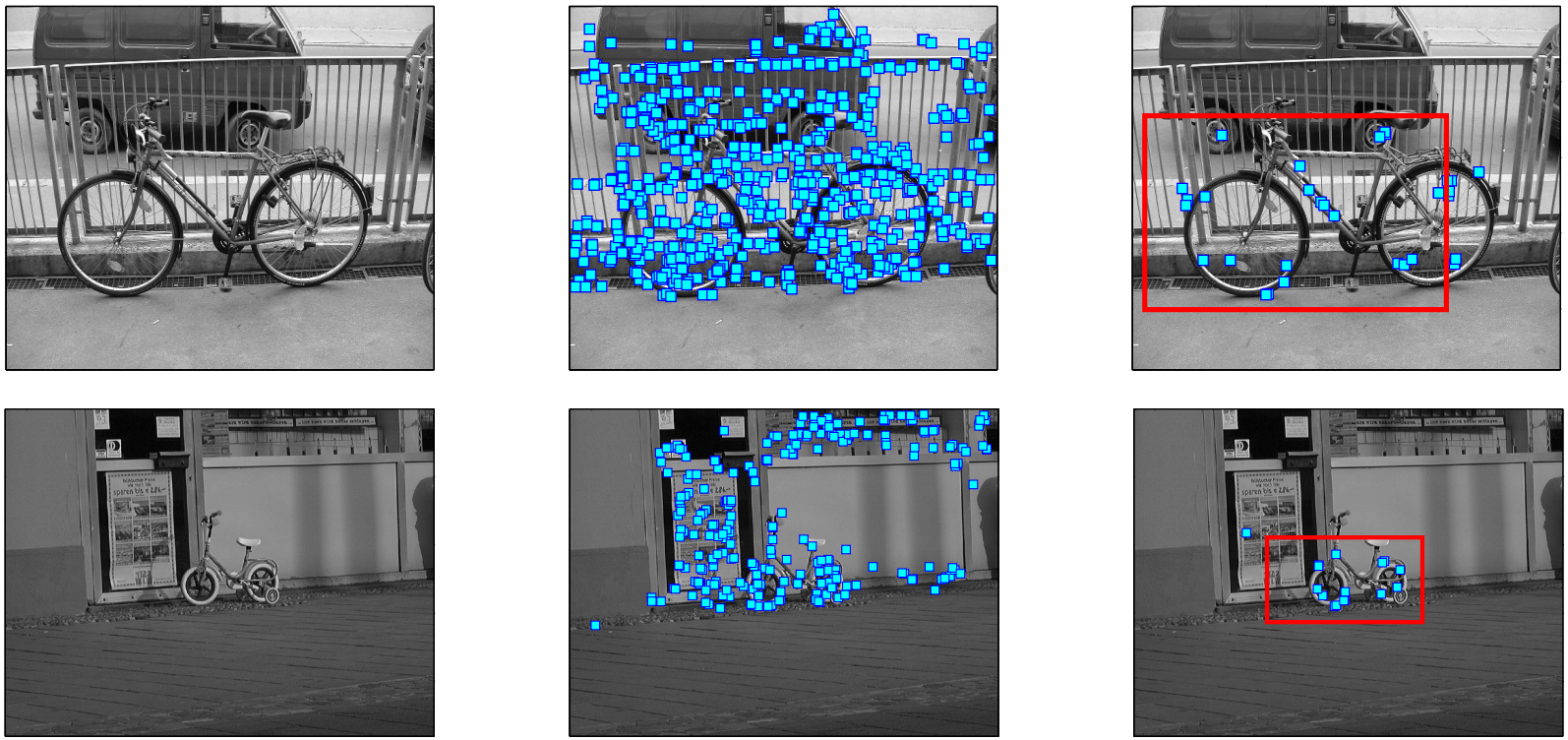


Fig. Object localization of an object "bike" in a natural image.

Texture recognition

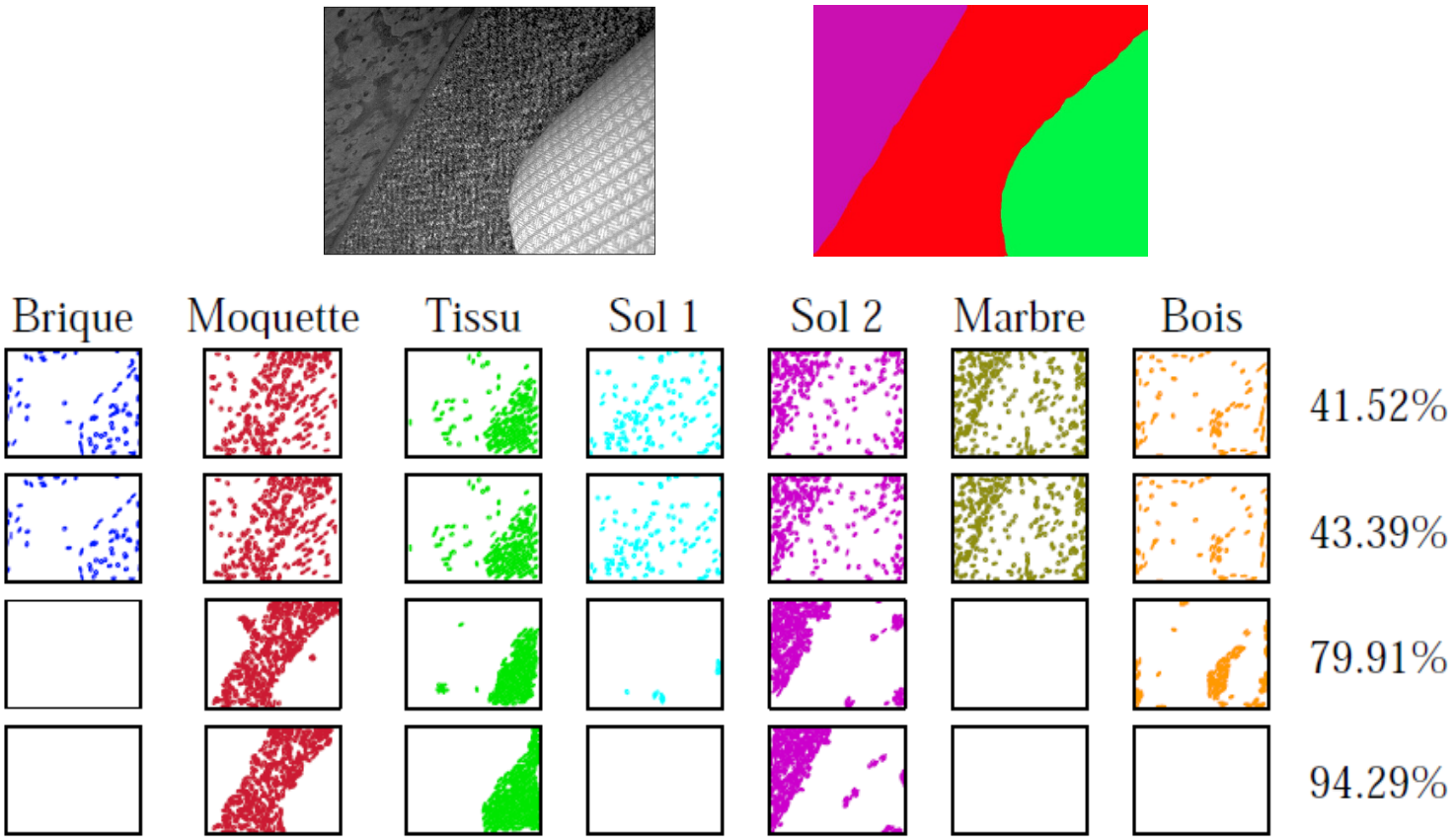


Fig. Segmentation of an image containing several textures: diag-GMM, HD-GMM, diag-GMM with hidden Markov field and HD-GMM with hidden Markov field.

Outline

Intrinsic dimension selection

In subspace clustering:

- the different models are all parametrized by the intrinsic dimension of the subspaces,
- Bouveyron *et al.* have proposed to use the scree-test of Cattell to determine the dimensions d_k,
- this approach works quite well in practice and can be combine to either cross-validation or BIC to select the threshold.

A priori, ML should not be used to determine the d_k:

- since the d_k determine the model complexity and therefore the likelihood increases with d_k,
- *except for the model* $[a_k b_k Q_k d_k]$*!*

The isotropic PPCA model

To simplify, let us define the isotropic PPCA model:

- the observed variable $Y \in \mathbb{R}^p$ and the latent variable $X \in \mathbb{R}^d$ are assumed to be linked:

$$Y = \bar{Q}X + \mu + \varepsilon,$$

- where X and ε have Gaussian distributions such that $\Delta_k = Q^t \Sigma Q$ has the following form:

$$\Delta_k = \begin{pmatrix} \begin{matrix} a & & 0 \\ & \ddots & \\ 0 & & a \end{matrix} & \mathbf{0} \\ \mathbf{0} & \begin{matrix} b & & 0 \\ & \ddots & \\ 0 & & b \end{matrix} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a & & 0 \\ & \ddots & \\ 0 & & a \end{matrix}}\right\} d \\ \left.\vphantom{\begin{matrix} b & & 0 \\ & \ddots & \\ 0 & & b \end{matrix}}\right\} (p-d) \end{matrix}$$

where $a \geq b$ and $d < p$.

ML estimate of d is asymptically consistent

Proposition:

The maximum likelihood estimate of the actal intrinsic dimension d^* is asymptotically unique and consistent.

Sketch of the proof:

At the optimum, the maximization of $\ell(\hat{\theta})$ is equivalent to the minimization of:

$$f_n(d) = d\log(\hat{a}) + (p-d)\log(\hat{b}) + p.$$

1. If $d \leq d^*$: $\hat{a} \to a$ and $\hat{b} \to \frac{1}{p-d}\left[(d^*-d)a + (p-d^*)b\right]$ almost surely when $n \to \infty$ and $f_n \to f$:

$$f(d) = d\log(a) + (p-d)\log\left(\frac{(d^*-d)}{(p-d)}a + \frac{(p-d^*)}{(p-d)}\right),$$

   which has a unique minimum in $d = d^*$.

2. If $d \geq d^*$: $\hat{a} \to \frac{1}{d}\left(d^*a + (d-d^*)b\right)$ and $\hat{b} \to b$ almost surely when $n \to \infty$ and $f_n \to f$:

$$f(d) = d\log\left(\frac{d^*}{d}a + \frac{d-d^*}{d}b\right) + (p-d)\log(b),$$

   which has as well a unique minimum in $d = d^*$.

Experimental setup

To verify the practical interest of the result:

- we define the parameters α and β:

$$\alpha = \frac{n}{p},$$

$$\beta = \frac{d^* a}{(p - d^*)b},$$

- α controls the estimation conditions through the ratio between the number of observations and the observation space dimension,
- β controls the signal to noise ratio through the ratio between the variances in the latent subspace and in its orthogonal subspace.

An introductory simulation

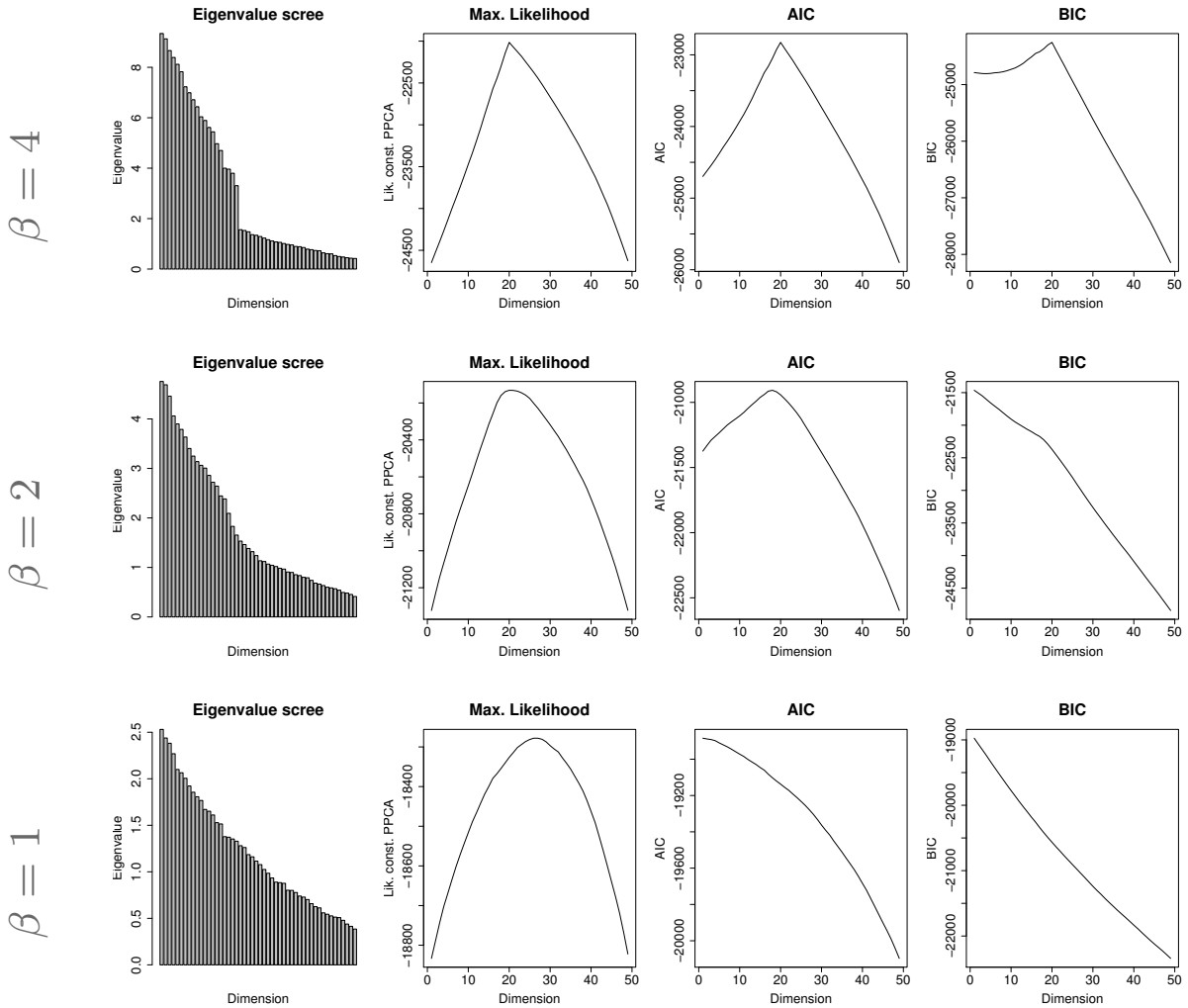


Figure: Intrinsic dimension estimation with $d^* = 20$ and $\alpha = 5$.

Influence of the signal to noise ratio

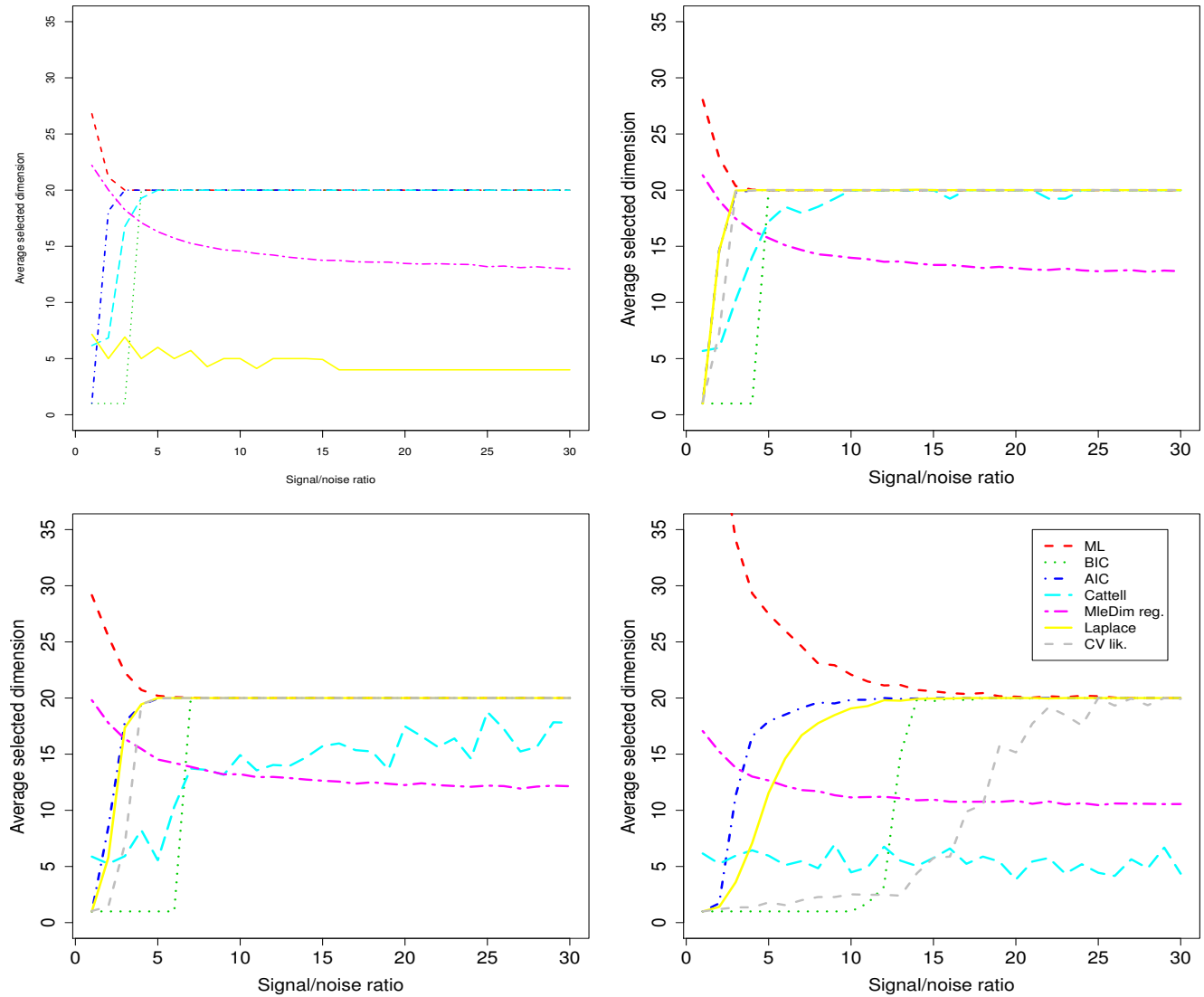


Figure: Average selected dimension according to β for $\alpha = 4, 3, 2$ and 1.

Influence of the n/p ratio

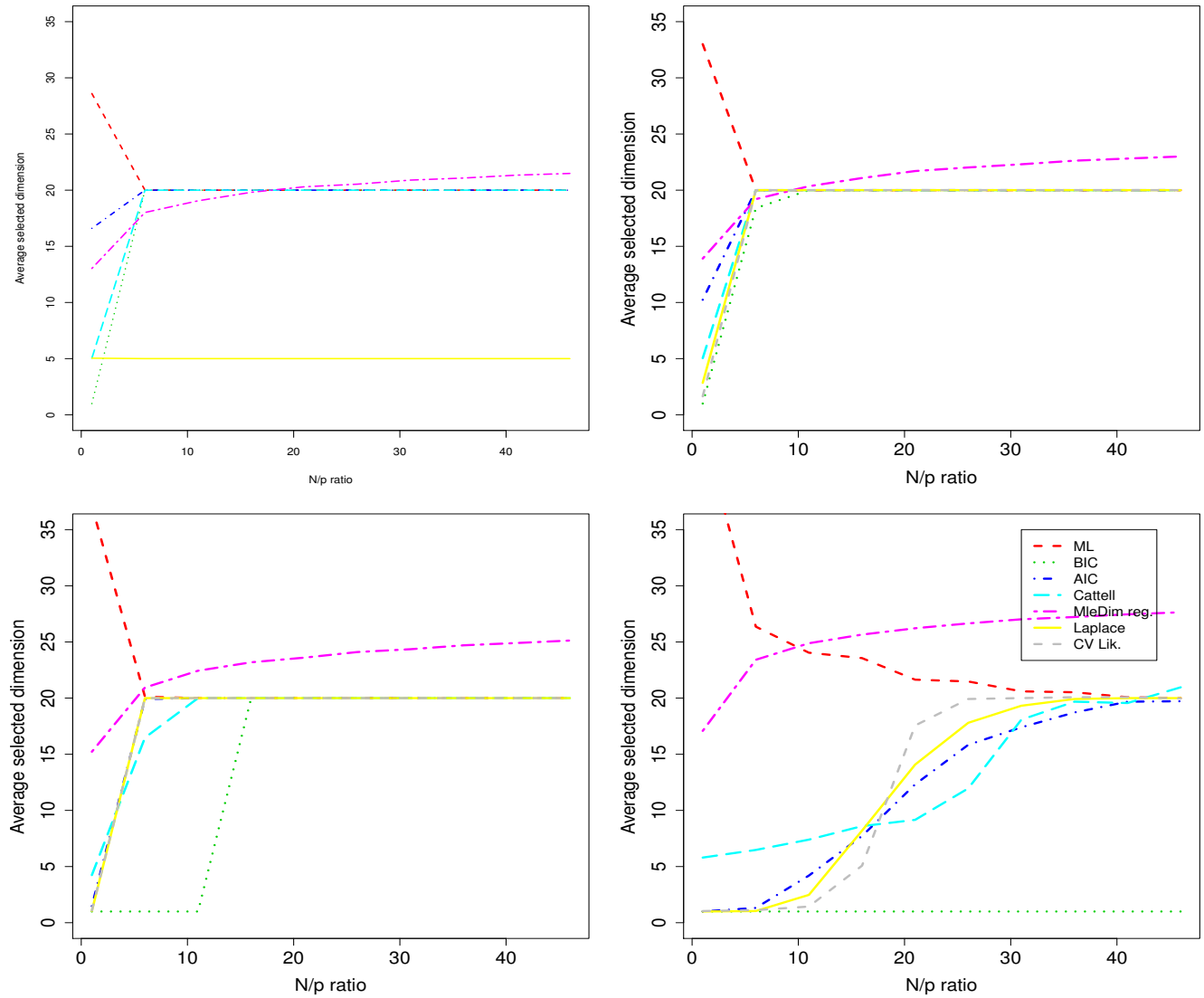


Figure: Average selected dimension according to α for $\beta = 4, 3, 2$ and 1.

A graphical summary

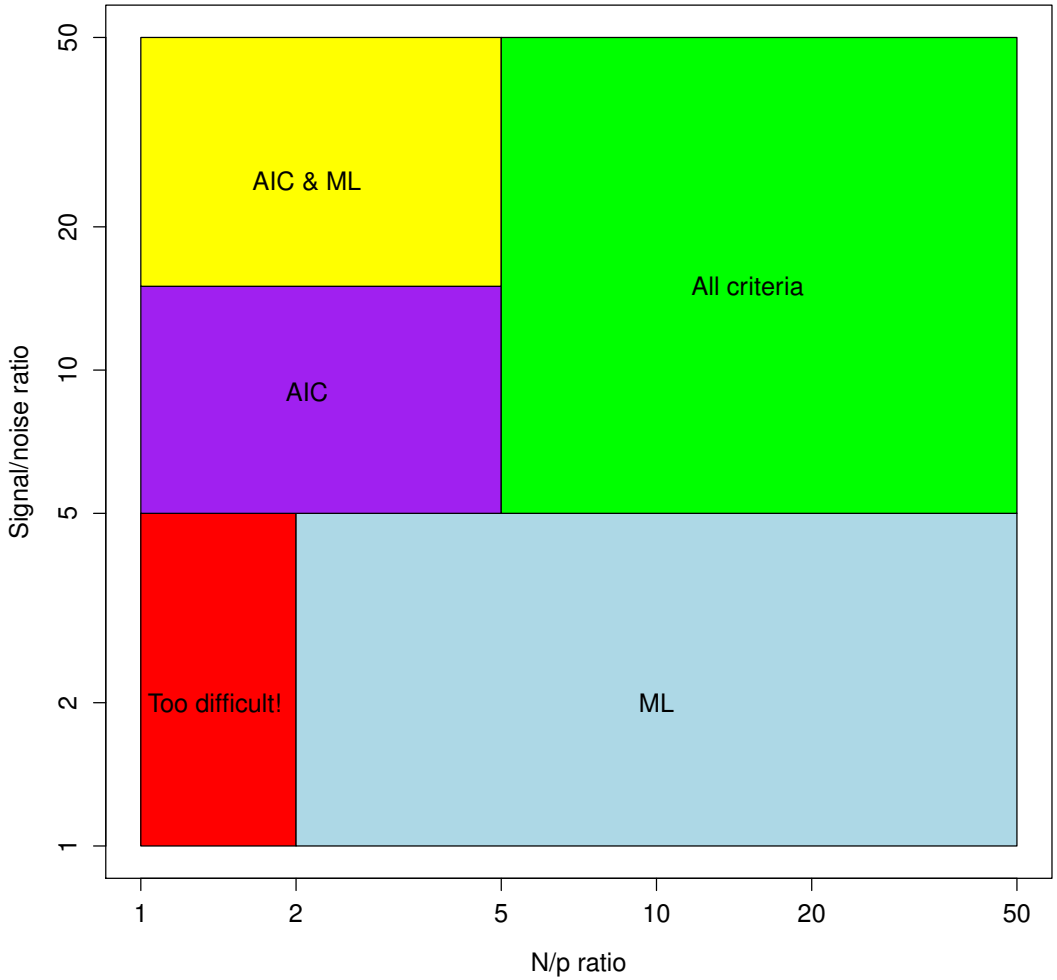


Figure: Recommended criteria for intrinsic dimension selection according to α and β for the isotropic PPCA model.

Outline

Conclusion & further works

Dimension reduction:

- is usefull for visualization purposes,
- but clustering a reduced dataset is suboptimal.

Parsimonious models & regularization:

- allow to adapt the model complexity to the data,
- parsimonious models are usually valid for data with $p<25$,

Subspace clustering:

- adapted for real high dimensional data ($p>25,100,1000,...$),
- even when n is small compared to p,
- the best of dimension reduction and parsimonious models.

Conclusion & further works

Intrinsic dimension selection:

- intrinsic dimension of the subspaces is the key parameter in subspace clustering,
- the old-fashion method of Cattell works quite well in practice,
- BIC, AIC and even ML can also be used in specific contexts.

Further works:

- use ML in HDDA and HDDC to make these methods fully automatic,
- integration of this approach in softwares.

Softwares

HDDA / HDDC:

- Matlab toolboxes are available at:

  http://samm.univ-paris1.fr/-charles-bouveyron-

- 8 models are available in the Mixmod software:

  http://www-math.univ-fcomte.fr/mixmod/

- A R package, nammed **`HDclassif`**, is available for a few weeks on the CRAN servers (thanks to L. Bergé & R. Aidan).

Fisher-EM:

- a R package is planned for next year...