Université Paris 1
Panthéon - Sorbonne

HDclassif : an R Package for Model-Based Classification of High-Dimensional Data

Charles BOUVEYRON

Laboratoire SAMM, EA 4543
Université Paris 1 Panthéon-Sorbonne

This joint work with L. Bergé & S. Girard

Disclaimer

"Essentially, all models are wrong but some are useful"

George E.P. Box

Outline

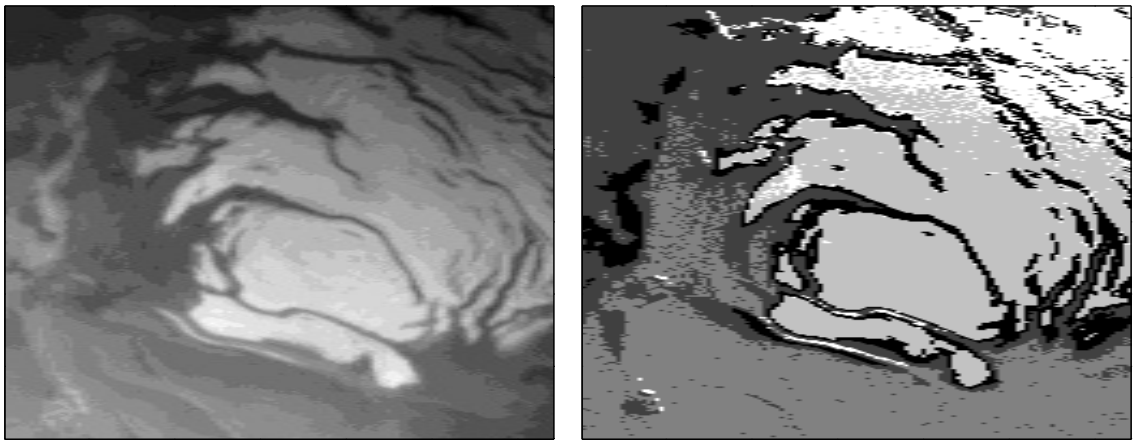
Outline

Introduction

Classification 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

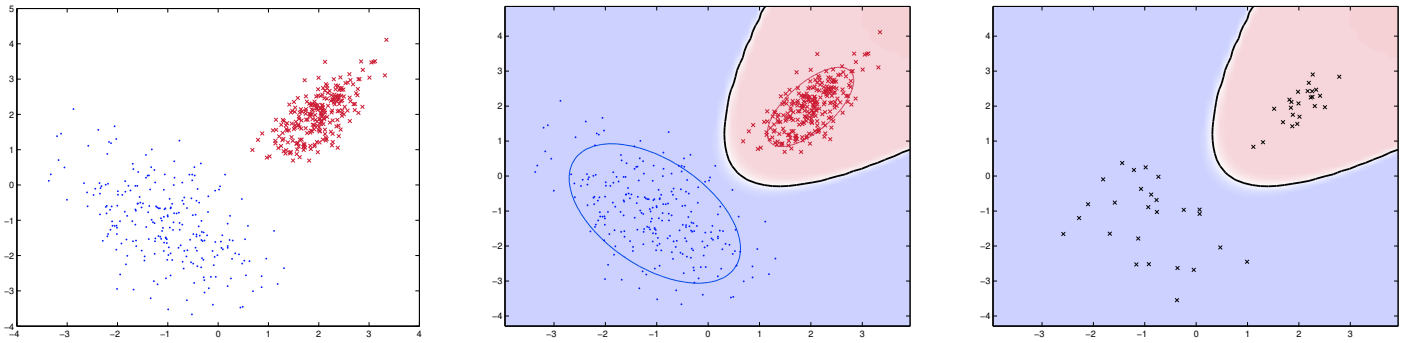


The classification problem

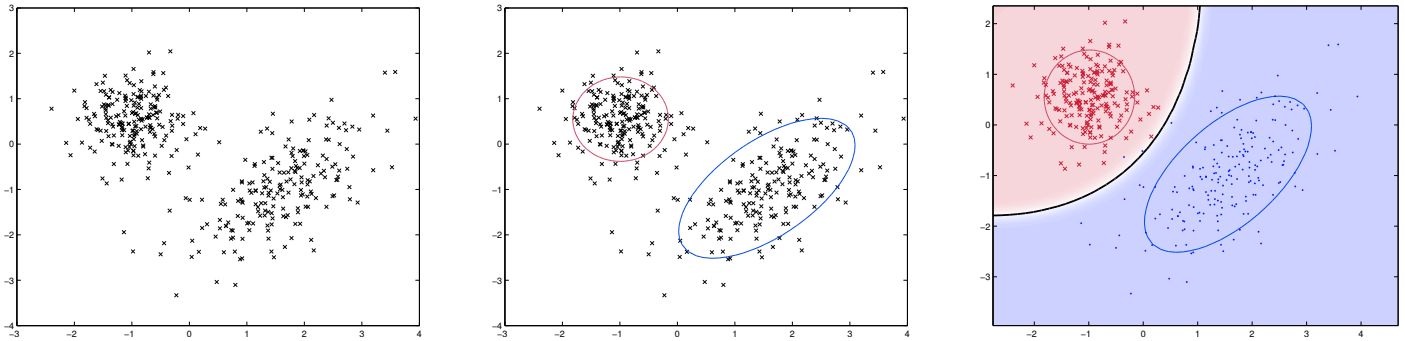
The classification problem consists in:

- 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 $x_1, ..., x_n$



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):

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

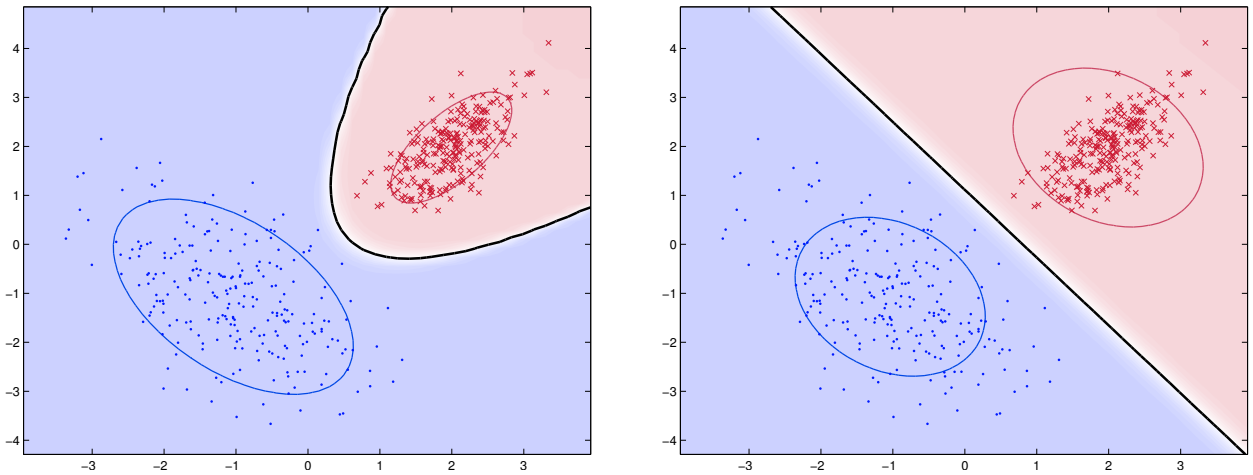


Fig. Decision boundaries for Full-GMM (left) and Com-GMM (right).

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,

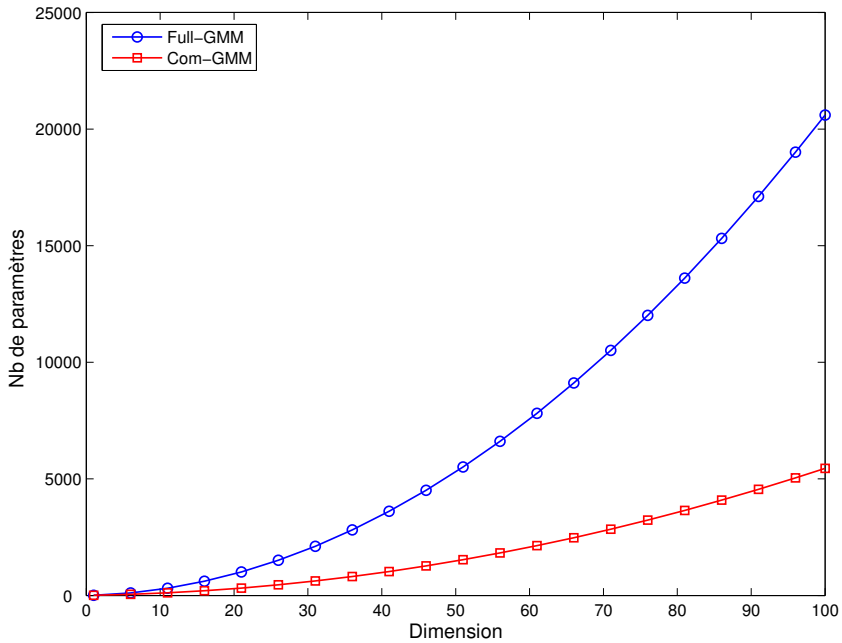


Fig. Number of parameters to estimate for the models Full-GMM and Com-GMM regarding to the dimension and with $k = 4$.

- 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

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

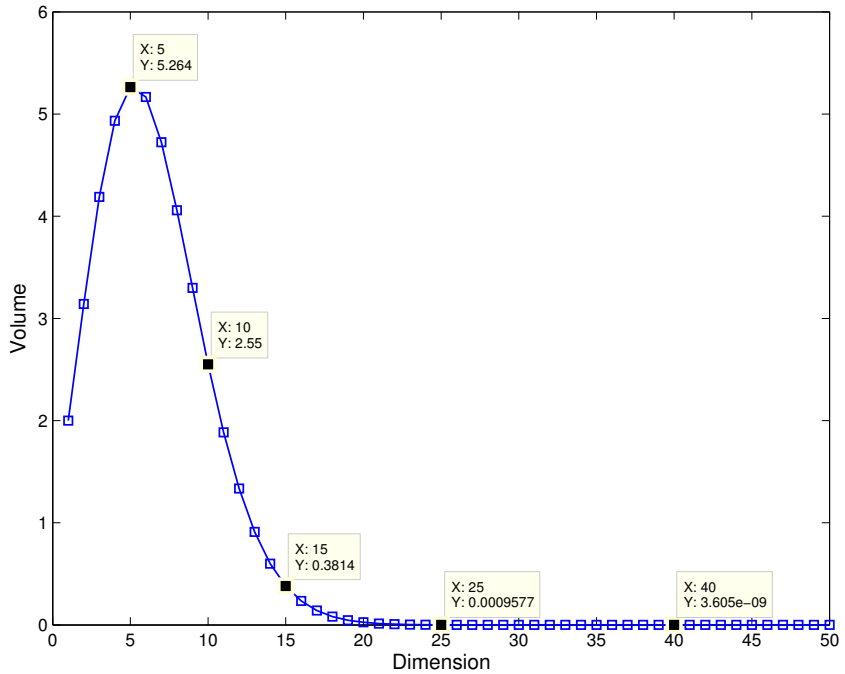


Fig. Volume of a sphere of radius 1 regarding to the dimension p.

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.

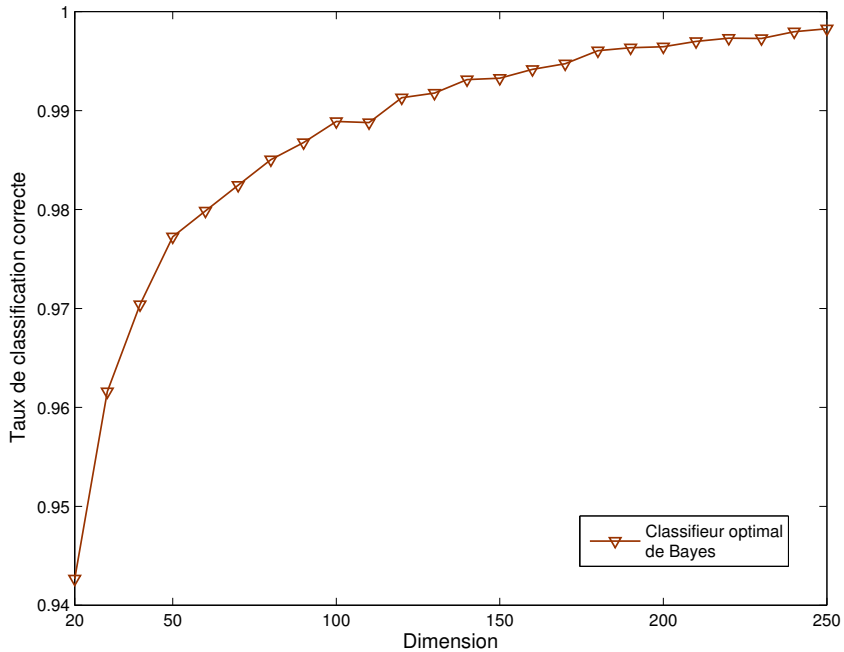


Fig. Correct classification rate of the optimal classifier versus the data dimension on simulated data.

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!

Outline

Subspace classification 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).

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:

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

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 $[a_{kj}b_kQ_kd_k]$

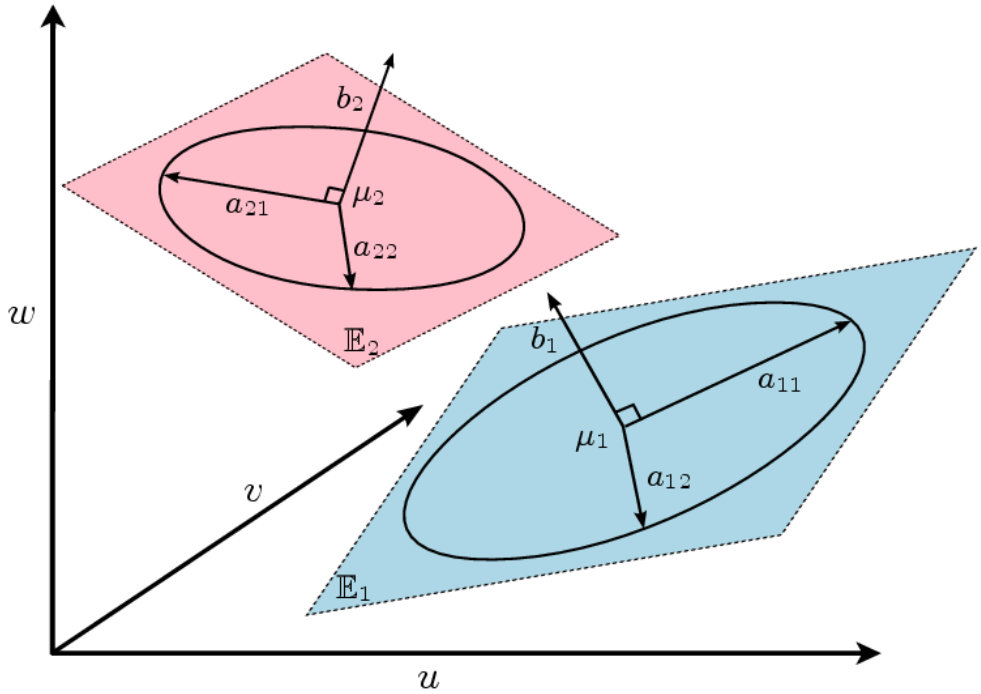


Fig. The subspace $\mathbb{E}_k$ and its supplementary $\mathbb{E}_k^{\perp}$.

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 $[a_{kj}b_kQ_kd_k]$

Construction of the classifiers

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.

Hyper-parameter estimation

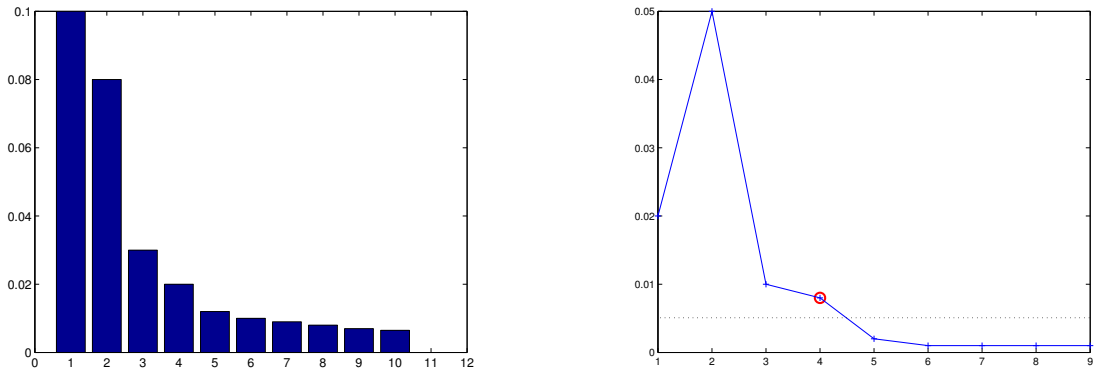


Fig. The scree-test of Cattell based on the eigenvalue scree.

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.

Special case $n \leq p$

In modern data analysis, it is now frequent to consider data sets where the number of observations n is smaller than the dimension p of the observation space.

In this specific case:

- the previous modeling allows to compute the classifiers HDDA and HDDC from the Gram matrices $\bar{X}_k \bar{X}_k^t$ which are $n_k \times n_k$ matrices, $k = 1, ..., K$,
- instead of using the empirical covariance matrices $\bar{X}_k^t \bar{X}_k$ which are $p \times p$ matrices,
- since both matrices share the same eigenvalues and their eigenvectors are linked by $u_j = \bar{X}_k v_j$.

Outline

The package HDclassif: inputs

The **hdda** and **hddc** routines:

- **model**: one of the 14 HD models that we selected for their good behavior in practical situations. The default is "AkjBkQkDk". If "ALL" is specified, then all models are tested and the result of the one with the highest BIC is returned.
- **d**: the way that dimensions d_k are chosen (Cattell, BIC or CV).
- **threshold**: the threshold for the scree-test of Cattell. The default value is 0.2.
- **graph**: if TRUE, several plots are displayed.

The package HDclassif: inputs

The **hddc** routine has in addition:

- **K**: the number of groups to form.
- **algo**: the inference algorithm to use (EM, CEM or SEM).
- **init**: the initialization procedure (random, kmeans, param, mini-em or a personal initialization vector).

The **predict** routine computes the class prediction of a dataset for a parameter set estimated by either **hdda** or **hhdc**.

The **plot** routine allows to visualize the eigenvalue screes and the estimated intrinsic dimensions d_k.

The package HDclassif: outputs

The **hdda** and **hddc** routines:

- **prms**: all estimated parameters (a_{kj}, b_k, Q_k, d_k, ...).
- **bic**: the BIC value.

The **hddc** and **predict** routine has in addition:

- **class**: the clustering partition of the data.
- **posterior**: the $n \times K$ matrix of the posterior probabilities $P(Z = k|X = x_i)$.
- **loglik**: the log-likelihood value at each iteration of the algorithm.

The package HDclassif: HDDA

Live demo ... with all inherent risks !

The package HDclassif: HDDC

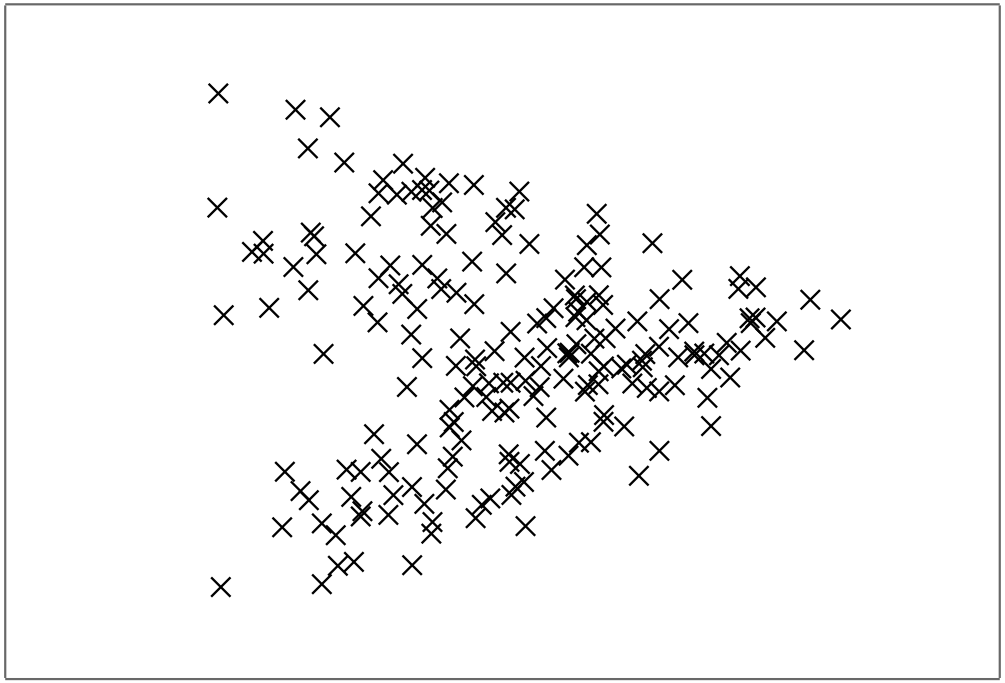


Fig. Projection of the «Crabs» data on the first principal axes.

«Crabs» data:

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

The package HDclassif: HDDC

Live demo ... with all inherent risks !

Outline

Conclusion

Dimension reduction:

- is usefull for visualization purposes,
- but classification in 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 classification:

- 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

Intrinsic dimension selection:

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

Recent extensions:

- the Fisher-EM algorithm models and clusters the data in a discriminative and common latent subspace,
- the methods pgpDA and pgpEM are kernelized versions of HDDA and HDDC which allow to classify data of various types (categorical, functions, networks, mixed data, ...).

Softwares

HDclassif:

- the R package HDclassif is available on the CRAN (thanks to Laurent Bergé),
- the article "HDclassif : an R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data" published in the Journal of Statistical Software (2012) presents the practical use of the package.

Other softwares for HDDA/C:

- 8 models are available in the Mixmod software:

  http://www.mixmod.org

- Matlab toolboxes are available at:

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

Fisher-EM:

- a R package, nammed FisherEM is available on the CRAN as well.