

Clustering of high-dimensional data

Problems, challenges and some recent advances



Charles BOUVEYRON

Professor of Applied Mathematics
Chair of Excellence Inria "Data Science"

Laboratoire LJAD, UMR CNRS 7351
Equipe Epione, Inria Sophia-Antipolis
Université Côte d'Azur

charles.bouveyron@unice.fr
@cbouveyron

Disclaimer

“Essentially, all models are wrong but some are useful”

George E.P. Box

Outline

Introduction

Problems and challenges in clustering

Existing approaches for clustering and their limits

Subspace clustering: HDDC

Discriminative clustering: Fisher-EM

Discriminative variable selection by ℓ_1 penalization

Conclusion

Introduction

Statistical learning is nowadays an unavoidable field:

- it aims to model a phenomenon and predict its future behavior,
- classification is one of the most active topic in this field.

A big challenge is to learn from modern data which are:

- high-dimensional (p large),
- big or as stream (n large),
- evolutive (evolving phenomenon),
- heterogeneous (categorical, functional, networks, ...)

The understanding of the clustering results is essential:

- in many applications, practitioners are very interested in visualizing the clustered data,
- and to have a selection of the relevant original variables for interpretation.

A motivating example: cytology

Cytology:

- it is the study of cells in terms of structure, function and chemistry,
- for the diagnosis of disease (we focused on cervical cancer).

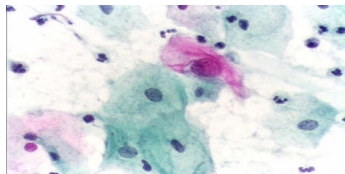
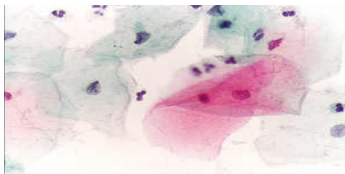


Figure: Normal (left) and abnormal (right) pap smears.

Cervical cancer detection:

- it is an important public health field which is currently treated mostly manually,
- pap smear screening by human experts is complicated by the amount of cells per smear (up to 20 000),
- and by the very small proportion of cancer cells (less than 1%).

A motivating example: cytology

Our data (BC Cancer Agency):

- 20 smears which contains between 4 000 and 10 000 cells,
- each nucleus is described by 111 features (morphological, photometric or texture features),
- only 0.52% of the cells are diseased cells.

Classification is useful in this context:

- for building supervised classifiers which can select the most likely cancer cells,
- for helping experts in labeling the learning data through weakly-supervised classification,
- for selecting discriminative variables which can be used in a semi-automatic process.

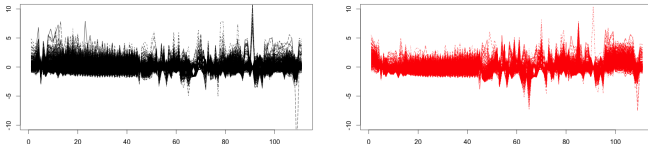
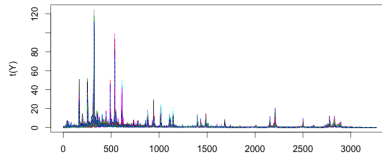


Figure: Control and (cervical) cancer data.

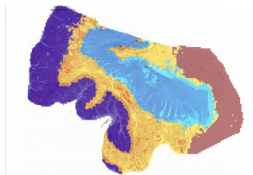
A motivating example: mass spectrometry

Mass spectrometry:

- it is a recent analytical technique that measures the mass-to-charge ratio of charged particles and which aims is to identify the elemental composition of a sample,
- It exist two types of mass spectrometry data:
 - **multi-array data** which aims to analyze serums or tissue fragments



- **MALDI images** which are 2D or 3D MS images of tissues or organs



A motivating example: mass spectrometry

Classification is useful in this context:

- it is used in Medicine for disease diagnostic from blood samples:
 - a supervised classifier is learned from blood samples of healthy and sick patients,
 - the classifier is then used to classify new blood samples.
- a combination of supervised and unsupervised classification can be used to detect errors in the labels

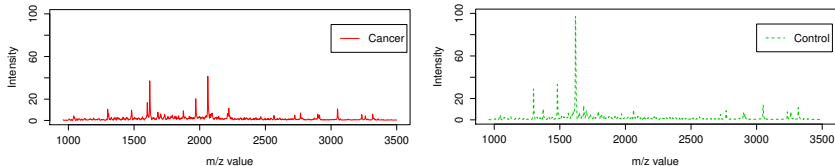


Figure: Control and cancer (colorectal) mass spectrometry spectra.

A motivating example: hyperspectral imaging

Hyperspectral imaging:

- it is an imaging technique which collects information from across the electromagnetic spectrum,
- as a consequence, the result is an image where each pixel is a high-dimensional spectrum,
- among the application fields, we can cite: agriculture, mineralogy, environment, security, astronomy.

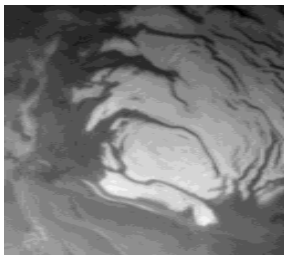


Figure: Image of the studied zone (south pole) of planet Mars.

A motivating example: hyperspectral imaging

The data from IPAG:

- a 300×128 hyperspectral image of the south pole of Mars,
- each “pixel” is described by a 256-dimensional spectrum.

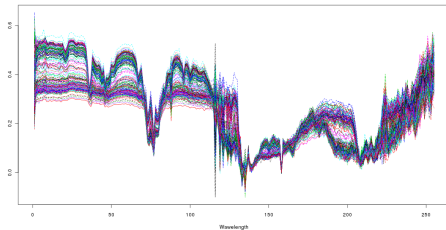


Figure: A few spectra of the studied zone.

Classification is useful in this context:

- for the segmentation of the studied zones -> ground nature classification,
- for selecting the discriminative spectral bands which allows the ground nature determination.

Outline

Introduction

Problems and challenges in clustering

Existing approaches for clustering and their limits

Subspace clustering: HDDC

Discriminative clustering: Fisher-EM

Discriminative variable selection by ℓ_1 penalization

Conclusion

Gaussian mixtures for clustering

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):

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

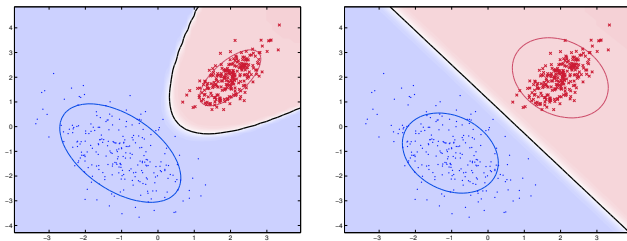


Fig. Decision boundaries for Full-GMM (left) and Com-GMM (right).

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 promote 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**:

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

Consequently:

- it is necessary to invert Σ_k which have a **number of parameters proportional to p^2** ,
- if n is small compared to p^2 , the estimates of Σ_k are **ill-conditioned or singular** and it will be **difficult or impossible to invert Σ_k** .

The curse of dimensionality

From the estimation point of view:

- let us consider the **normalized trace** $\tau(\Sigma) = \text{tr}(\Sigma^{-1})/p$ of the inverse covariance matrix Σ^{-1} of a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$,
- the **estimation of τ** from a sample of n observations $\{x_1, \dots, x_n\}$ conduces to:

$$\tau(\hat{\Sigma}) = \tau(\hat{\Sigma}) = \frac{1}{p} \text{tr}(\hat{\Sigma}^{-1}),$$

$$E[\tau(\hat{\Sigma})] = \left(1 - \frac{p}{n-1}\right)^{-1} \tau(\Sigma).$$

- **consequently**, if the ratio $p/n \rightarrow 0$ when $n \rightarrow +\infty$, then $E[\tau(\hat{\Sigma})] \rightarrow \tau(\Sigma)$,
- **however**, if the dimension p is comparable with n , then $E[\tau(\hat{\Sigma})] \rightarrow c\tau(\Sigma)$ when $n \rightarrow +\infty$, where $c = \lim_{n \rightarrow +\infty} p/n$.

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.

First example: volume of the unit sphere is $V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

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.

First example: volume of the unit sphere is $V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)}$,

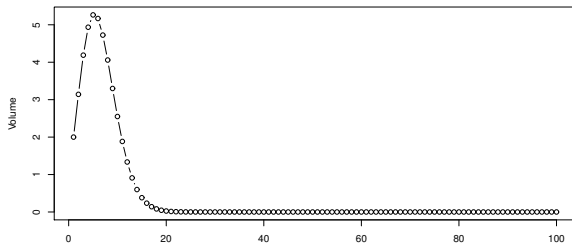


Fig. Volume of a sphere of radius 1 regarding to the dimension p .

The blessings of dimensionality

Second example: probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

$$P(X \in S_{0.9}(p)) = 1 - 0.9^p \xrightarrow{p \rightarrow \infty} 1$$

The blessings of dimensionality

Second example: probability that a uniform variable on the unit sphere belongs to the shell between the spheres of radius 0.9 and 1 is

$$P(X \in S_{0.9}(p)) = 1 - 0.9^p \xrightarrow{p \rightarrow \infty} 1$$

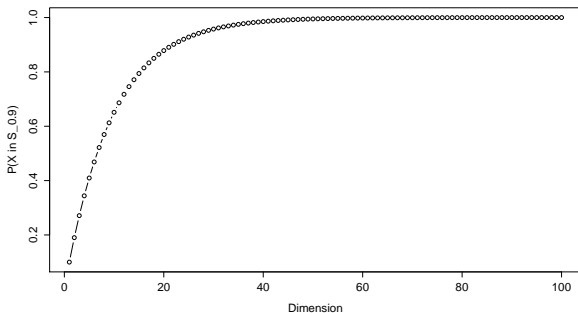


Fig. Probability that X belongs to the shell $S_{0.9}$ regarding to the dimension p .

The blessings of dimensionality

Third example:

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- a way to observe this is to look at the Bayes classifier behaviour.

The blessings of dimensionality

Third example:

- since high-dimensional spaces are almost empty,
- it should be easier to separate groups in high-dimensional space with an adapted classifier,
- a way to observe this is to look at the Bayes classifier behaviour.

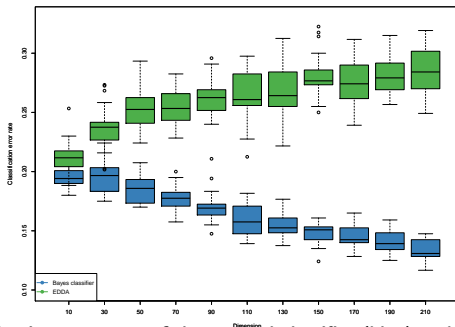


Fig. Classification error rate of the optimal classifier (blue) and EDDA (green) versus the data dimension on simulated data.

Outline

Introduction

Problems and challenges in clustering

Existing approaches for clustering and their limits

Subspace clustering: HDDC

Discriminative clustering: Fisher-EM

Discriminative variable selection by ℓ_1 penalization

Conclusion

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!

Regularization:

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

Parsimonious models:

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

Dimension reduction

A common phantasm about dimension reduction:

- believe that dimension reduction helps for classification,
- **this is not true** because, most of the time, dimension reduction implies an information loss which would be discriminative.

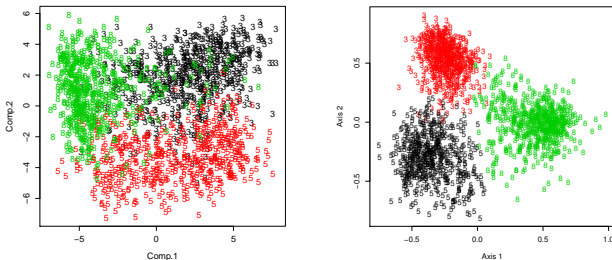


Figure: Projection of the 256-dimensional USPS data with PCA (left, unsupervised) and FDA (right, supervised).

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.

Regularization

Regularization:

- classical regularizations includes ridge and lasso penalties or covariance estimate thresholding,
- although these approaches are often very efficient from the numerical point of view,
- they do not make sense from the modeling point of view and are difficult to parametrize in the unsupervised case.

Regularization of the covariance matrix estimates:

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

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

where S_k is defined by:

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

Parsimonious models

Parsimonious models:

- making restrictive assumptions on the models allows to generate a family of models (Celeux & Govaert, Fraley & Raftery)
- which can fit into several situations while keeping a sense regarding modeling,
- they are however not efficient for truly high-dimensional data ($p > 50$).

Parsimonious Gaussian models:

- com-GMM:
 - the assumption: $\Sigma_k = \Sigma$,
 - nb of par. for $K = 4$ and $p = 100$: 5453
- diag-GMM:
 - the assumption: $\Sigma_k = \text{diag}(\sigma_{k1}, \dots, \sigma_{kp})$,
 - 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

Recent approaches for clustering

In the past decade, several innovative approaches were proposed:

- **subspace clustering:**

- several key works: Tipping & Bishop (Mixt. PPCA), McLachlan *et al.* (MFA), Bouveyron *et al.* (HDDC), McNicholas & Murphy (PGMM), Beak *et al.* (MCFA), Montanari & Viroli (HFMA),
- clustering in low-dimensional subspaces has shown a high efficiency but their result are difficult to interpret,

Recent approaches for clustering

In the past decade, several innovative approaches were proposed:

- **subspace clustering:**

- several key works: Tipping & Bishop (Mixt. PPCA), McLachlan *et al.* (MFA), Bouveyron *et al.* (HDDC), McNicholas & Murphy (PGMM), Beak *et al.* (MCFA), Montanari & Viroli (HFMA),
- clustering in low-dimensional subspaces has shown a high efficiency but their result are difficult to interpret,

- **variable selection for clustering:**

- Dean & Raftery and Maugis *et al.* proposed a Bayesian framework to iteratively select the relevant variables for model-based clustering,
- these approaches successfully identify the relevant variables for the clustering but are time-consuming.

Recent approaches for clustering

In the past decade, several innovative approaches were proposed:

- **subspace clustering:**

- several key works: Tipping & Bishop (Mixt. PPCA), McLachlan *et al.* (MFA), Bouveyron *et al.* (HDDC), McNicholas & Murphy (PGMM), Beak *et al.* (MCFA), Montanari & Viroli (HFMA),
- clustering in low-dimensional subspaces has shown a high efficiency but their result are difficult to interpret,

- **variable selection for clustering:**

- Dean & Raftery and Maugis *et al.* proposed a Bayesian framework to iteratively select the relevant variables for model-based clustering,
- these approaches successfully identify the relevant variables for the clustering but are time-consuming.

- **sparsity:**

- Pan & Shen and Galimberti *et al.* proposed ℓ_1 -penalized maximum likelihood approaches to select the relevant variables,
- Witten & Tibshirani recently proposed a ℓ_1 -penalized approach for k-means and hierarchical clustering,
- these methods are also very efficient but time-consuming and difficult to parametrize.

Outline

Introduction

Problems and challenges in clustering

Existing approaches for clustering and their limits

Subspace clustering: HDDC

Discriminative clustering: Fisher-EM

Discriminative variable selection by ℓ_1 penalization

Conclusion

Objectives of subspace clustering

Our objectives:

- **clustering efficiency**: the methodology should match the performance standard of classical clustering techniques from both the clustering and the computing points of view,
- **modeling**: the methodology should provide a probabilistic modeling of each group and should be able to automatically choose the number of groups,
- **visualization**: the methodology should provide a comprehensive low-dimensional representation of the clustered data,

Objectives of subspace clustering

Our objectives:

- **clustering efficiency**: the methodology should match the performance standard of classical clustering techniques from both the clustering and the computing points of view,
- **modeling**: the methodology should provide a probabilistic modeling of each group and should be able to automatically choose the number of groups,
- **visualization**: the methodology should provide a comprehensive low-dimensional representation of the clustered data,

Our proposal:

- a subspace clustering method which models and clusters the data in low-dimensional subspaces.

The model $[a_k j b_k Q_k d_k]$

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 $k = 1, \dots, K$.

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:

$$\Delta_k = \left(\begin{array}{cc|cc} \boxed{\begin{array}{cc} a_{k1} & 0 \\ & \ddots \\ 0 & a_{kd_k} \end{array}} & & & \\ & \mathbf{0} & & \\ \hline & & \boxed{\begin{array}{cc} b_k & 0 \\ & \ddots \\ 0 & b_k \end{array}} & \\ & \mathbf{0} & & \end{array} \right) \left. \begin{array}{l} \left. \begin{array}{c} \end{array} \right\} d_k \\ \left. \begin{array}{c} \end{array} \right\} (p - d_k) \end{array} \right.$$

where:

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

The model $[a_{kj}b_kQ_kd_k]$

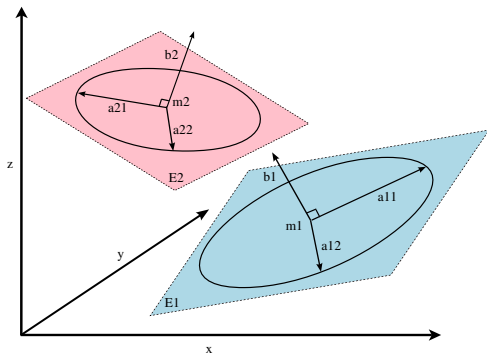


Fig. The subspace \mathbb{E}_k and its supplementary \mathbb{E}_k^\perp .

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	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**.

HDDC: the E step

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

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

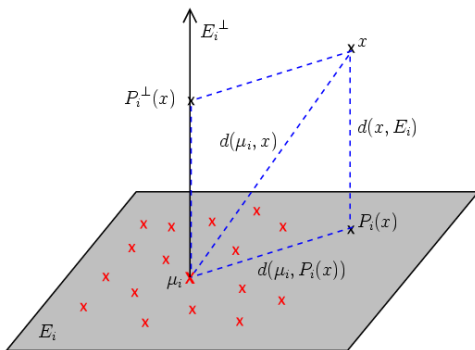


Fig. The subspaces \mathbb{E}_k and \mathbb{E}_k^\perp of the k th mixture component.

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 S_k of the k th class.
- Estimator of a_{kj} : the parameters a_{kj} are estimated by the d_k largest eigenvalues λ_{kj} of S_k .
- Estimator of b_k : the parameter of b_k is estimated by:

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

HDDC: hyper-parameter estimation

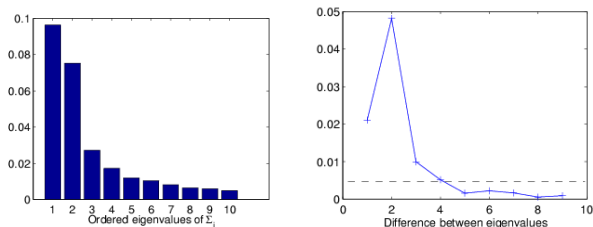


Fig. The scree-test of Cattell based on the eigenvalue scree.

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 **number 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 $S_k \rightarrow$ 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 $S_k = \bar{X}_k^t \bar{X}_k$ ($\times 500$ for $n = 13$ and $p = 1000$).

HDDC: an EM-based algorithm

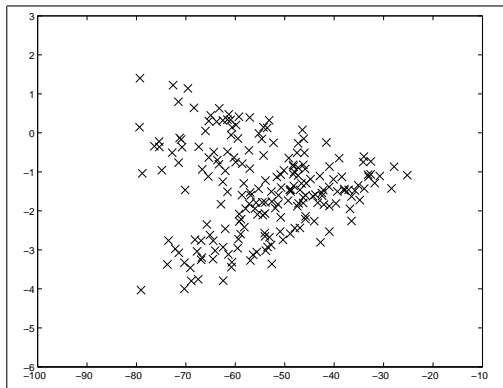


Fig. Projection of the «Crabs» data on the first principal axes.

«Crabs» data:

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

HDDC: an EM-based algorithm

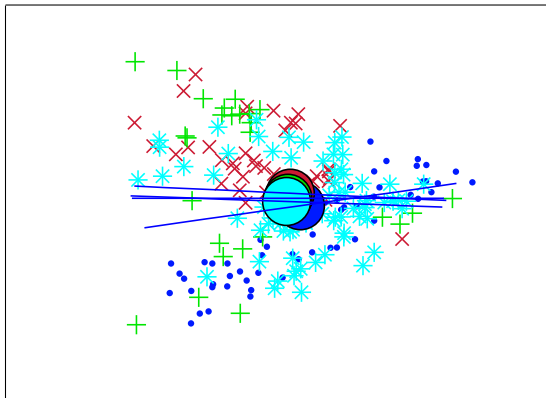


Fig. Step n° 1 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

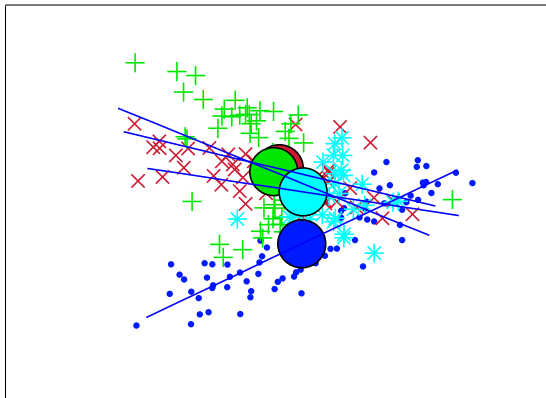


Fig. Step n° 4 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

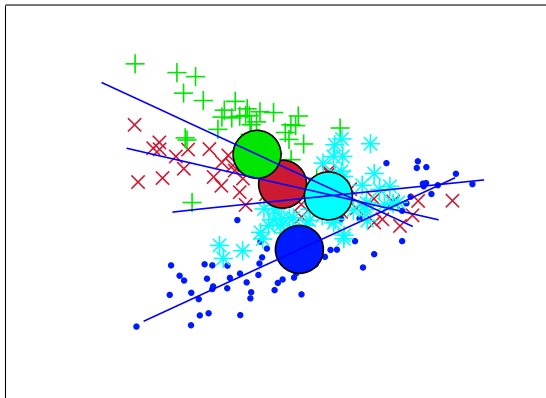


Fig. Step n° 7 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

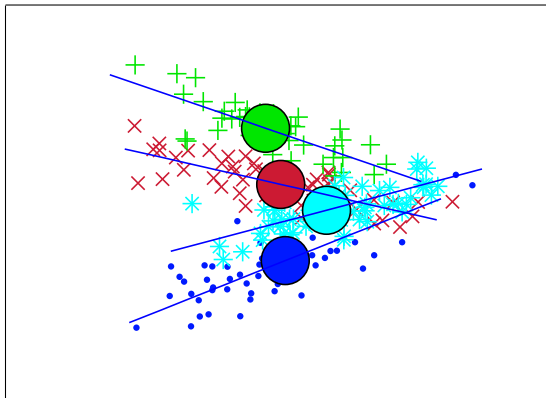


Fig. Step n° 10 of HDDC on the «Crabs» data.

HDDC: an EM-based algorithm

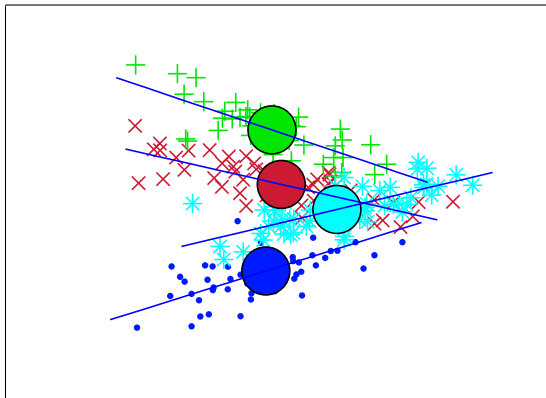


Fig. Step n° 12 of HDDC on the «Crabs» data.

Outline

Introduction

Problems and challenges in clustering

Existing approaches for clustering and their limits

Subspace clustering: HDDC

Discriminative clustering: Fisher-EM

Discriminative variable selection by ℓ_1 penalization

Conclusion

Objectives of discriminative clustering

Our objectives:

- **clustering efficiency**: the methodology should match the performance standard of subspace clustering techniques from both the clustering and the computing points of view,
- **modeling**: the methodology should provide a probabilistic modeling of each group and should be able to automatically choose the number of groups,
- **visualization**: the methodology should provide a **unique** and comprehensive low-dimensional representation of the clustered data,
- **interpretation**: the methodology should allow to select the discriminative variables which may have specific meanings (biology, economics, ...)

Objectives of discriminative clustering

Our objectives:

- **clustering efficiency**: the methodology should match the performance standard of subspace clustering techniques from both the clustering and the computing points of view,
- **modeling**: the methodology should provide a probabilistic modeling of each group and should be able to automatically choose the number of groups,
- **visualization**: the methodology should provide a **unique** and comprehensive low-dimensional representation of the clustered data,
- **interpretation**: the methodology should allow to select the discriminative variables which may have specific meanings (biology, economics, ...)

Our proposal:

- a subspace clustering method which models and clusters the data in a **common** and **discriminative** low-dimensional subspace.

The DLM model... at a glance!

The observed random vector $Y \in \mathbb{R}^p$ is linked to a latent random vector $X \in \mathbb{E}$ (supposed to be the most discriminative) by:

$$Y = UX + \varepsilon,$$

where U is a $p \times d$ orthogonal matrix ($U^T U = I_d$) and $d < p$.

Distribution assumptions, for $k = 1, \dots, K$:

$$\varepsilon \sim \mathcal{N}(\mathbf{0}, \Psi),$$

$$X_{|Z=k} \sim \mathcal{N}(\mu_k, \Sigma_k),$$

The marginal distribution of Y is then:

$$f(y) = \sum_{k=1}^K \pi_k \phi(y; m_k, S_k),$$

where $m_k = U\mu_k$ and $S_k = U\Sigma_k U^T + \Psi_k$.

The DLM model... at a glance!

We finally assume that the noise covariance matrix Ψ_k is such that $\Delta_k = W^T S_k W$ has the following form:

$$\Delta_k = \left(\begin{array}{c|c} \boxed{\Sigma_k} & \mathbf{0} \\ \hline \mathbf{0} & \begin{array}{ccc} \beta_k & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & & \beta_k \end{array} \end{array} \right) \left. \begin{array}{l} \vphantom{\Delta_k} \\ \vphantom{\Delta_k} \end{array} \right\} \begin{array}{l} d \leq K - 1 \\ (p - d) \end{array}$$

where $W = [U, V]$.

This model is referred to by $\text{DLM}_{[\Sigma_k \beta_k]}$ and 11 submodels can be obtained by constraining parameters within or between groups.

The DLM model

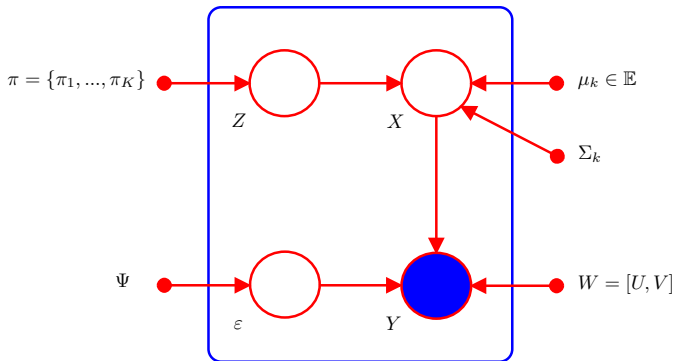


Figure: Graphical summary of the $\text{DLM}_{[\Sigma_k \beta]}$ model

The Fisher-EM algorithm

The inference of mixture models:

- is usually done with the **EM algorithm** since likelihood maximization is intractable,
- **however**, we can not make use of the EM algorithm here since the subspace has to be discriminant.

We therefore proposed **the Fisher-EM algorithm** for inferring the DLM models:

- **a E step** which, roughly speaking, determines the current data partition through the posterior probabilities $t_{ik} = E[z_{ik} = 1|y_i]$,
- **a F step** which determines the orientation matrix U according to the current partition of the data,
- **a M step** which updates the mixture parameters conditionally to U and t_{ik} .

Looking back in the past: Fisher's criterion

We based our F step on the idea of **Fisher's discriminant analysis** (1936):

- knowing a partition of the data, Fisher's objectives were to **find a low-dimensional subspace** such that:
 - the groups are well separated \rightarrow large between-class variance S_B
 - the groups are homogeneous \rightarrow small within-class variance S_W
- since $S = S_W + S_B$, the usual **Fisher criterion** writes as follows:

$$\max_U \quad \text{tr} \left((U^T S U)^{-1} U^T S_B U \right),$$

- the solution of this optimization problem are the **$d = K - 1$ eigenvectors of the matrix $S^{-1} S_B$**

Looking back in the past: Fisher's criterion

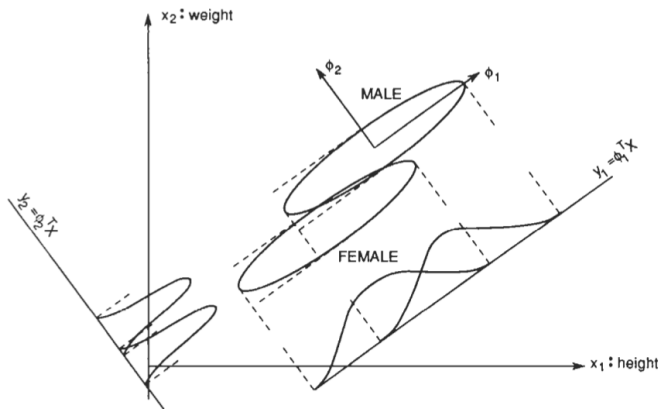


Figure: Discriminative axis vs. principal axis (Fukunaga, 1990)

The F step of the Fisher-EM algorithm

The F step of Fisher-EM:

- determines the orientation matrix U according to the $t_{ik}^{(q)}$ by solving the **unsupervised counterpart of Fisher's criterion**:

$$\begin{cases} \max_U & \text{tr} \left((U^T S U)^{-1} U^T S_B^{(q)} U \right), \\ \text{wrt} & u_j^T u_l = 0, \quad \forall j \neq l \in \{1, \dots, d\}, \end{cases} \quad (1)$$

where:

- $S_B^{(q)} = \frac{1}{n} \sum_{k=1}^K n_k^{(q)} (\hat{m}_k^{(q)} - \bar{y})^T (\hat{m}_k^{(q)} - \bar{y})$,
 - $n_k^{(q)} = \sum_{i=1}^n t_{ik}^{(q)}$, $\hat{m}_k^{(q)} = \frac{1}{n} \sum_{i=1}^n t_{ik}^{(q)} y_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.
- we proposed a **Gramm-Schmidt procedure** to solve this constrained optimization problem.

The Fisher-EM algorithm... at work!

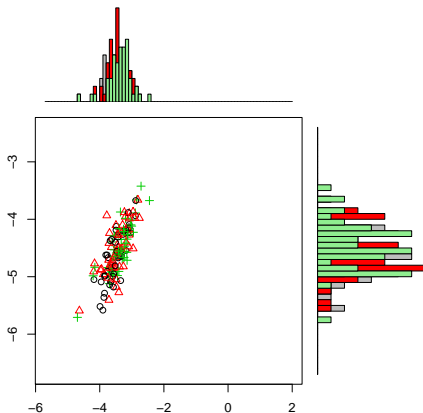


Figure: Initialization of the Fisher-EM algorithm on the Iris data.

The Fisher-EM algorithm... at work!

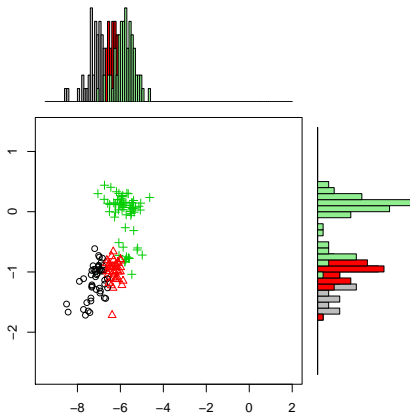


Figure: Step 1 of the Fisher-EM algorithm on the Iris data.

The Fisher-EM algorithm... at work!

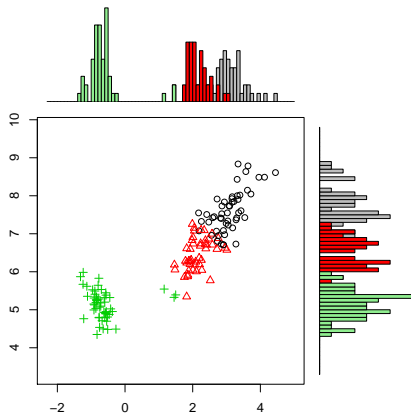


Figure: Step 3 of the Fisher-EM algorithm on the Iris data.

The Fisher-EM algorithm... at work!

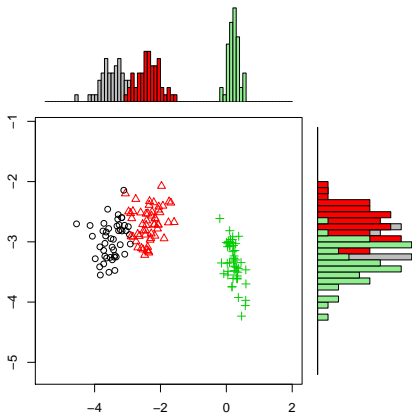


Figure: Step 5 of the Fisher-EM algorithm on the Iris data.

The Fisher-EM algorithm... at work!

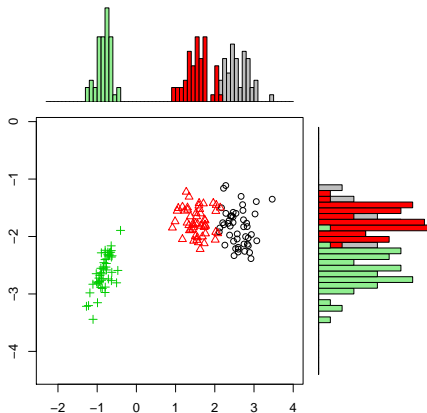


Figure: Step 7 of the Fisher-EM algorithm on the Iris data.

The Fisher-EM algorithm... at work!

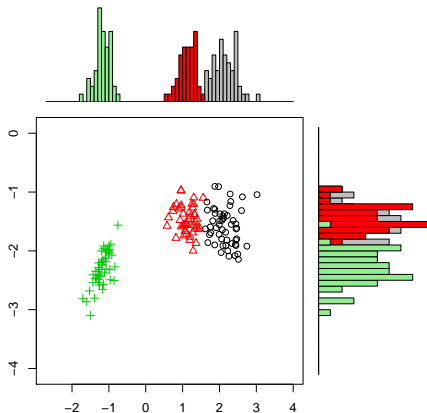


Figure: Step 9 of the Fisher-EM algorithm on the Iris data.

The Fisher-EM algorithm... at work!

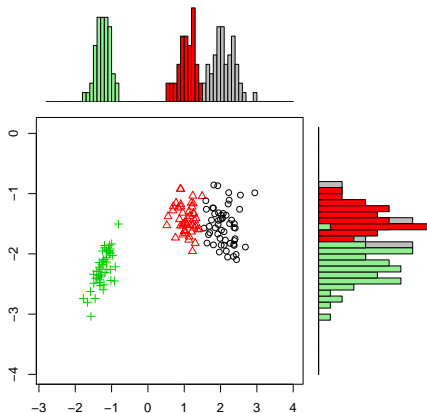


Figure: Step 11 of the Fisher-EM algorithm on the Iris data.

Experimental results: benchmark

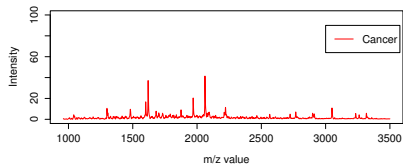
Method	iris	wine	chiro	zoo	glass	satimage	usps358
DLM _[$\Sigma_k \beta_k$]	86.8±7.3†	97.8±0.0*	91.2±6.1	80.1±5.7	48.5±2.6	69.6±0.0*	81.1±5.4*†
DLM _[$\Sigma_k \beta$]	92.6±11	89.3±0.0	98.2±3.4	-	47.9±2.7	64.5±0.0	77.4±9.1
DLM _[$\Sigma \beta_k$]	80.5±3.4	93.8±1.1	94.7±4.2	72.6±5.3	49.4±2.9	65.7±1.3	73.7±7.4
DLM _[$\Sigma \beta$]	79.1±2.9	89.8±0.8	85.2±3.2	79.6±5.6	48.6±3.6	65.5±1.6	76.4±9.9
DLM _[$\alpha_{kj} \beta_k$]	87.8±0.5*	97.2±0.0†	85.0±1.4	71.8±6.6†	49.6±2.6†	70.1±0.0	82.3±4.7
DLM _[$\alpha_{kj} \beta$]	97.8±0.1	95.2±1.6	98.1±5.2	71.4±8.0	51.1±2.1*	61.7±0.2	73.2±9.5
DLM _[$\alpha_k \beta_k$]	92.8±2.1	98.9±0.0	85.5±14*†	71.8±6.9*	48.5±2.2	68.8±0.0	70.9±13.6
DLM _[$\alpha_k \beta$]	95.8±7.3	97.1±0.9	97.8±5.0	71.0±6.4	49.5±2.4	68.8±0.0	68.3±11.2
DLM _[$\alpha_j \beta_k$]	81.6±4.5	91.6±0.5	93.8±4.1	68.5±6.7	49.3±1.8	62.9±0.0†	76.1±11.0
DLM _[$\alpha_j \beta$]	73.6±6.7	89.8±0.9	89.7±4.1	79.1±4.9	47.4±1.2	67.6±2.8	77.4±10.7
DLM _[$\alpha \beta_k$]	80.1±6.9	91.4±3.2	89.3±1.9	70.1±6.5	48.9±1.3	68.7±1.9	80.5±6.0
DLM _[$\alpha \beta$]	66.8±0.0	89.5±1.0	89.2±5.7	80.2±5.3	47.0±1.7	62.1±0.0	69.9±14.2
Full-GMM	79.0±5.7	60.9±7.7	44.8±4.1	-	38.3±2.1	35.9±3.1	-
Com-GMM	57.6±18.3	61.0±14.9	51.9±10.9	59.9±10.3	38.3±3.1	26.1±1.5	38.2±1.1
Mixt-PPCA	89.1±4.2	63.1±7.9	56.3±4.5	50.9±6.5	37.0±2.3	40.6±4.7	53.1±9.6
Diag-GMM	93.5±1.3	94.6±2.8	92.1±4.2	70.9±12.3	39.1±2.4	60.8±5.2	45.9±9.1
Sphe-GMM	89.4±0.4	96.6±0.0	85.9±9.9	69.4±5.4	37.0±2.1	60.2±7.5	78.7±11.2
PCA-EM	66.9±9.9	64.4±5.7	66.1±4.0	61.9±6.2	39.0±1.7	56.2±4.2	67.6±11.2
k-means	88.7±4.0	95.9±4.0	92.9±6.0	68.0±7.4	41.3±2.8	66.6±4.1	74.9±13.9
MCFA ($q = 3$)	80.6±12.6	92.9±8.2	75.4±7.8	-	47.7±6.9	67.9±8.8	54.2±8.7
PGMM	96.7±0.0	97.1±0.0	97.9±0.0	65.3±0.0	41.6±0.0	58.7±0.0	55.5±0.0
Mclust	96.7	97.1	97.9	65.3	41.6	58.7	55.5
Model name	(VEV)	(VVI)	(EEE)	(EII)	(VEV)	(VVV)	(EEE)

Table: Clustering accuracies and their standard deviations on 7 UCI datasets.

Application to mass spectrometry

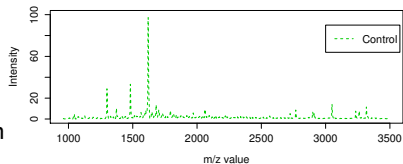
The MS dataset (T. Alexandrov, Univ. of Bremen):

- 112 spectra,
- 6 168 dimensions (m/z ratios),
- 64 patients with a colorectal cancer,
- 48 healthy patients.



His objectives:

- compare the clustering results of Fisher-EM with a known partition from experts,
- to identify potential label errors in the expert partition.



Application to mass spectrometry

PCA-EM			Fisher-EM		
Class	Cluster		Class	Cluster	
	Cancer	Control		Cancer	Control
Cancer	48	16	Cancer	57	7
Control	1	47	Control	3	45
Misclassification rate = 0.15			Misclassification rate = 0.09		

Table: Confusion tables for PCA-EM (left) and Fisher-EM (right).

Application to mass spectrometry

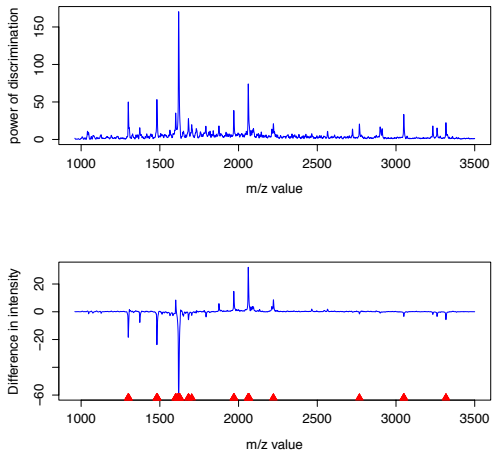


Figure: Interpretation of the loading matrix U .

Outline

Introduction

Problems and challenges in clustering

Existing approaches for clustering and their limits

Subspace clustering: HDDC

Discriminative clustering: Fisher-EM

Discriminative variable selection by ℓ_1 penalization

Conclusion

Discriminative variable selection

Clustering is a data analysis tool and result interpretation is important. Unfortunately, the loading matrix U is usually **difficult to interpret**:

<i>variable</i>	axis 1	axis 2
sepal length	-0.203	-0.062
sepal width	-0.324	-0.697
petal length	0.519	0.404
petal width	0.763	-0.588

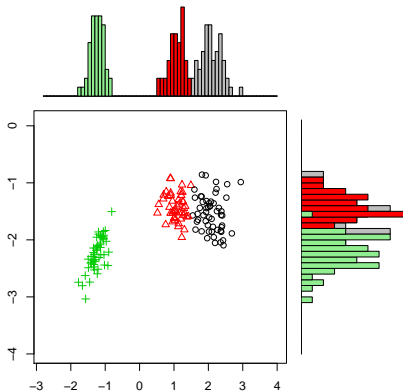
Discriminative variable selection

Clustering is a data analysis tool and result interpretation is important. Unfortunately, the loading matrix U is usually **difficult to interpret**:

<i>variable</i>	axis 1	axis 2
sepal length	-0.203	-0.062
sepal width	-0.324	-0.697
petal length	0.519	0.404
petal width	0.763	-0.588

And we would **prefer**:

<i>variable</i>	axis 1	axis 2
sepal length	0	0
sepal width	0	-1
petal length	0	0
petal width	1	0



Three ways to introduce sparsity

We chose to introduce sparsity within the F step:

- we want to identify the original variables which best discriminate the groups,
- which amounts to estimate the orientation matrix U with, as much as possible, only 0 or ± 1 ,
- a popular way to do that is to use a ℓ_1 penalty (lasso).

We identified three different ways to introduce sparsity:

1. SparseFEM₁: classical F step + sparsity step,
2. SparseFEM₂: F step as a ℓ_1 -penalized regression problem,
3. SparseFEM₃: sparse SVD on the matrix $S^{-1}S_B^{(q)}$.

SparseFEM₂: ℓ_1 -penalized regression problem

Defining the matrices $H_W^{(q)}$ and $H_B^{(q)}$ such that $H_W^{(q)} H_W^{(q)t} = S_W^{(q)}$ and $H_B^{(q)} H_B^{(q)t} = S_B^{(q)}$, we obtained:

Proposition

The best sparse approximation at the level λ of the solution of (1) is the solution \hat{B} of the following penalized regression problem:

$$\min_{A,B} \sum_{k=1}^K \left\| R_W^{(q)-t} H_{B,k}^{(q)} - AB^t H_{B,k}^{(q)} \right\|_F^2 + \rho \sum_{j=1}^d \beta_j^t S_W^{(q)} \beta_j + \lambda \sum_{j=1}^d \|\beta_j\|_1,$$

such that $A^t A = \mathbf{I}_d$ and where $R_W^{(q)} \in \mathbb{R}^{p \times p}$ is such that $S_W^{(q)} = R_W^{(q)t} R_W^{(q)}$, $A = [\alpha_1, \dots, \alpha_d]$, $B = [\beta_1, \dots, \beta_d]$, $H_{B,k}^{(q)}$ is the k th column of $H_B^{(q)}$ and $\rho > 0$ is a ridge-type regularization parameter.

Remark : we proposed an iterative procedure based on the LARS algorithm to solve this problem.

Selection of the sparsity parameter

The selection of the hyper-parameter λ :

- this problem has received very few attention in the unsupervised context,
- a natural way in the model-based clustering context is to use the BIC criterion,
- but, the degree of freedom of the model has to be updated in order to take into account the sparsity!

Selection of the sparsity parameter

The selection of the hyper-parameter λ :

- this problem has received very few attention in the unsupervised context,
- a natural way in the model-based clustering context is to use the BIC criterion,
- but, the degree of freedom of the model has to be updated in order to take into account the sparsity!

Recent works [Zou07, Kachour11] have shown:

- that the number of non zero coefficients is a consistent estimator of the degree of freedom of the model,
- we finally get for the model of SparseFEM:

$$BIC_{pen}(\mathcal{M}) = -2 \log(\mathcal{L}(\hat{\theta})) - \gamma_e \log(n),$$

where $\gamma_e = (K - 1) + Kd + (d[p - (d + 1)/2] - \mathbf{d}_e) + Kd(d + 1)/2 + K$.

Comparison with variable selection

Approaches	iris ($p=4$)	glass ($p=7$)	wine ($p=13$)	zoo ($p=16$)	chiro ($p=17$)	satimage ($p=36$)	usps358 ($p=256$)
Fisher-EM	97.8±0.1	51.1±2.1	98.9±0.0	80.2±5.3	98.2±3.4	70.1±0.0	82.3±4.7
sparseFEM ₁	96.0±0.0 (2.0±0.0)	51.4±1.3 (6.1±0.6)	97.8±0.2 (3.2±2.1)	63.0±9.7 (11±0.6)	78.2±11 (2.0±0.0)	69.6±0.6 (30.1±0.6)	79.3±5.4 (47±4.1)
sparseFEM ₂	88.9±1.4 (4.0±0.0)	51.6±0.9 (6.0±1.6)	98.3±0.0 (3.0±0.0)	75.4±1.9 (13±4.5)	84.1±10 (2.8±0.79)	60.6±3.0 (31±1.2)	78.8±9.1 (82±16)
sparseFEM ₃	97.3±0.0 (3.4±0.9)	53.3±0.7 (7.0±0.0)	97.7±0.0 (2.0±0.0)	72.7±8.1 (14.2±2.5)	81.2±11 (4.9±2.7)	71.7±2.3 (29±2.6)	73.1±7.4 (5.0±1.3)
sparseKmeans	90.7 (4.0)	52.3 (6.0)	94.9 (13.0)	79.2 (16.0)	95.3 (17.0)	71.4 (36.0)	74.7 (213)
ClustVarSel	96.0 (3.0)	48.6 (3.0)	92.7 (5.0)	75.2 (3.0)	71.1 (6.0)	58.7 (19.0)	48.3 (6.0)
SelVarClust	88.7 (3.0)	43.0 (6.0)	94.4 (5.0)	92.1 (5.0)	92.6 (8.0)	56.4 (22.0)	36.7 (5.0)

Table 1: Clustering accuracies and their standard deviations (in percentage) on 7 UCI datasets (iris, wine, chironomus, zoo, glass, satimage, usps358) averaged on 20 trials. The average number of nonzero variables is reported in brackets. No standard deviation is reported for Mclust and Spardcl since their initialization procedure is deterministic and always provides the same initial partition.

Figure: Clustering accuracies and their standard deviations on 7 UCI datasets (averaged on 20 trials, models and λ selected by BIC).

A comparative example: the USPS358 dataset

We first considered the USPS358 dataset:

- which contains 1756 handwritten digits (3, 5 and 8),
- and each 16×16 grayscale image has been transformed as a 256-dimensional vector.

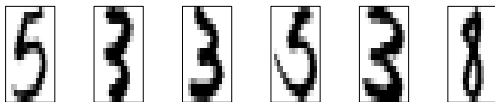


Figure: Sample from the USPS358 dataset.

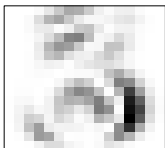
Approaches:	Clustering accuracy	Non-zero variables	Elapsed time in sec.
Fisher-EM	82.3 ± 4.7	256 ± 0.00	218.8 ± 1.5
SparseFEM ₁	82.69 ± 6.82	5.6 ± 0.97	967.8 ± 1.1
SparseFEM ₂	81.42 ± 6.77	16.0 ± 0.00	325.3 ± 1.0
SparseFEM ₃	80.62 ± 8.06	10.1 ± 4.63	58.3 ± 2.6

Table: Clustering accuracies and computing times for the 3 versions of the sparseFEM algorithm on the 256-dimensional dataset USPS358 ($\lambda = 0.1$).

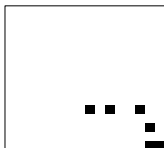
A comparative example: the USPS358 dataset

Method	Computing time	Method	Computing time
SparseFEM ₁	967.8 \pm 1.1 sec.	Sparse k-means	1 783 sec.
SparseFEM ₂	325.3 \pm 1.0 sec.	ClustVarSel	4 602 sec.
SparseFEM ₃	58.3 \pm 2.6 sec.		

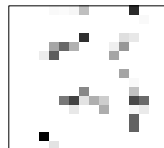
Table: Computing time on the USPS358 dataset.



(a) Sparse k-means



(b) ClustVarSel



(c) SparseFEM2

Figure: Variable selection obtained with the 3 sparse algorithms on the USPS358 dataset.

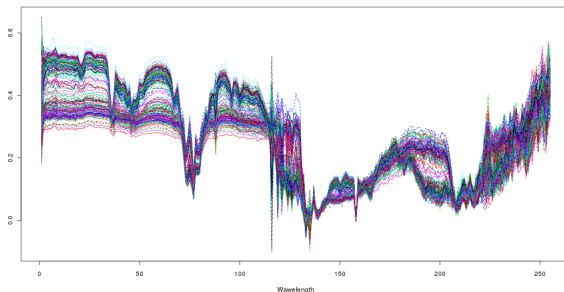
Application to hyper-spectral image analysis

The Mars Express data set:

- hyper-spectral images of the planet Mars taken in 2004,
- we considered the analysis of an image of the south pole of Mars,
- the data are 300×128 pixels described by 256 spectral variables.

We used sparseFEM to analyze this data set:

- the sparsity level λ was fixed to 0.1 to ensure to select a few discriminative variables,
- the whole process took 18 hours on a 2.6 Ghz computer.



Application to hyper-spectral image analysis

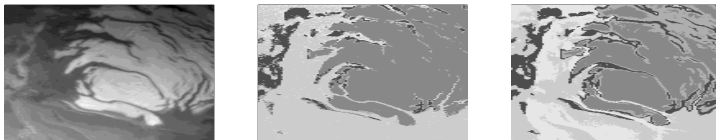


Figure: Segmentation results: original image (left), expert segmentation (center) and sparseFEM segmentation (right).

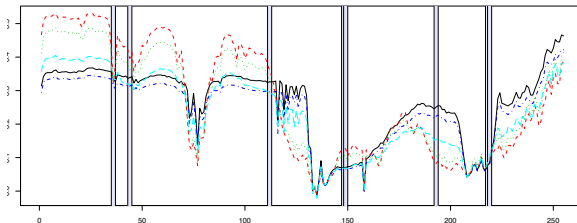


Figure: Selection of the discriminative spectral variables by sparseFEM.

Outline

Introduction

Problems and challenges in clustering

Existing approaches for clustering and their limits

Subspace clustering: HDDC

Discriminative clustering: Fisher-EM

Discriminative variable selection by ℓ_1 penalization

Conclusion

Conclusion

Model-based classification for HD data:

- is an efficient and flexible tool for classification / clustering,
- it provides in addition information about the classification risk.

Our contributions:

- we proposed two models adapted to the classification of HD data and their associated inference algorithms,
- they model and cluster the data in low-dimensional (and discriminative) subspaces,
- they usually performs better than other clustering methods while providing a useful visualizations,
- they allow in addition to identify the original variables which are discriminative.

Software:


- package [HDclassif](#) for the HDDA and HDDC methods,
- package [FisherEM](#) for the Fisher-EM algorithm.


References





C. Bouveyron and C. Brunet, *Model-based clustering of high-dimensional data : A review*, Computational Statistics and Data Analysis, vol. 71, pp. 52-78, 2014.


References


-  C. Bouveyron and C. Brunet, *Model-based clustering of high-dimensional data : A review*, Computational Statistics and Data Analysis, vol. 71, pp. 52-78, 2014.


-  C. Bouveyron, S. Girard and C. Schmid, *High-dimensional data clustering*, Computational Statistics and Data Analysis, vol. 52 (1), pp. 502-519, 2007.

-  C. Bouveyron and C. Brunet, *Simultaneous model-based clustering and visualization in the Fisher discriminative subspace*, Statistics and Computing, vol. 22 (1), pp. 301-324, 2012.

-  C. Bouveyron and C. Brunet, *Discriminative variable selection with the sparse Fisher-EM algorithm*, Computational Statistics, vol. 29(3-4), pp. 489-513, 2014.

-  N. Dean and A. Raftery, *Variable Selection for Model-Based Clustering*, Journal of the American Statistical Association, 101 (473), pp. 168-178, 2006.

-  C. Maugis, G. Celeux and M.-L. Martin-Magniette, *Variable selection in model-based clustering: A general variable role modeling*, Computational Statistics and Data Analysis, 53, pp. 3872-3882, 2009.

-  D. Witten and R. Tibshirani, *A framework for feature selection in clustering*, Journal of the American Statistical Association, 105(490), pp. 713-726, 2010.