# Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis 

Vincent Vandewalle

## - To cite this version:

Vincent Vandewalle. Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis. CMStatistics 2016, Dec 2016, Sevilla, Spain. hal-01424965

HAL Id: hal-01424965
https://inria.hal.science/hal-01424965
Submitted on 3 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Simultaneous dimension reduction and multi-objective clustering using probabilistic factorial discriminant analysis 

V. Vandewalle

University Lille 2, EA 2694 \& Inria Lille, Modal team
ERCIM 2016
Sevilla
December $11^{\text {th }}, 2016$

## Dimension reduction and clustering

## Summarizing the data

- Dimension reduction: find some principal components explaining the major variability in the data
- Clustering: find some clusters explaining the major heterogeneity of the data

Link between dimension reduction and clustering
Often when visualizing the data one is interested in visualizing clusters on the visualization space.

In practice
Dimension reduction and clustering are performed separately or sequentially but rarely simultaneously.

## Combining dimension reduction and clustering through probabilistic models

Generative models allow to combine visualization and clustering

- Tibshirani \& Hastie (1996): Reduced rank discriminant analysis
- Kumar \& Andreou (1998): Heteroscedastic discriminant analysis
- Bouveyron \& Brunet (2012): Model-based clustering and visualization in the Fisher discriminative subspace

In a rigorous probabilistic way

- A unique homogeneous criterion to optimize: the likelihood
- Simultaneous selection of the number clusters and of the number of components: BIC
- Missing data naturally taken into account: EM algorithm


## Multi-objective clustering

## Motivation: clustering part

- Usually clustering summarizes the data information by only one latent variable, the clustering variable.
- But we would like to allow for several views of the data with potentially more than one clustering variable.
- Often performances of clustering methods measured based on classification of reference, but many possible references in many settings (sex, species, status, ...).

Motivation: visualisation part

- View each clustering variable on a clustering component
- Heterogeneity-based visualisation rather than inertia-based visualisation


## Illustration (1/2)

Data in the initial space


## Illustration (2/2)

Principal Components Analysis


Principal Cluster Components


## Outline

1. Probabilistic interpretation of the factorial discriminant analysis
2. Multi-objective mixture model
3. Estimation of the parameters and model choice
4. Experiments on simulated and real datasets

## Linear discriminant analysis (LDA)

## Data

- $\boldsymbol{x} \in \mathbb{R}^{d}$ : the observed continuous variables
- $\boldsymbol{z} \in\{1, \ldots, K\}$ : the cluster (latter considered as unknown)

Assumptions
$n$ independent replicates of $(\boldsymbol{x}, \boldsymbol{z})$

- $\boldsymbol{z} \sim \operatorname{Mult}\left(\pi_{1}, \ldots, \pi_{K}\right)$ with $\pi_{k}=p\left(z_{k}=1\right)$
- $\forall k \in\{1, \ldots, K\}, \quad \boldsymbol{x} \mid z_{k}=1 \sim \mathcal{N}_{d}\left(\mu_{k}, \boldsymbol{\Sigma}\right)$

Parameters estimation
$\hat{\pi}_{k}=\frac{n_{k}}{n}, \hat{\boldsymbol{\mu}}_{k}=\overline{\boldsymbol{x}}_{k}$ and $\hat{\boldsymbol{\Sigma}}=\boldsymbol{W}$ (empirical within-class covariance matrix)
Classification

$$
p\left(z_{k}=1 \mid \boldsymbol{x}, \hat{\boldsymbol{\theta}}\right)=\frac{\hat{\pi}_{k} \phi_{d}\left(\boldsymbol{x} ; \hat{\boldsymbol{\mu}}_{k}, \hat{\boldsymbol{\Sigma}}\right)}{\sum_{k^{\prime}=1}^{K} \hat{\pi}_{k^{\prime}} \phi_{d}\left(\boldsymbol{x} ; \hat{\boldsymbol{\mu}}_{k^{\prime}}, \hat{\boldsymbol{\Sigma}}\right)} \Rightarrow \hat{\boldsymbol{z}}=\underset{k \in\{1, \ldots, K\}}{\operatorname{argmax}} \hat{\boldsymbol{\mu}}_{k}^{\top} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x}-\frac{1}{2} \boldsymbol{\mu}_{k}^{\top} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu}_{k}-\log \left(\hat{\pi}_{k}\right)
$$

$\Rightarrow$ linear classification boundary.

## Factorial discriminant analysis (FDA)

FDA principle

- The first component of FDA finds $\boldsymbol{v}_{1} \in \mathbb{R}^{d}$ maximizing the variance explained by the cluster:

$$
\boldsymbol{v}_{1}=\arg \max _{\boldsymbol{v} \in \mathbb{R}^{d}} \frac{\boldsymbol{v}^{\top} \boldsymbol{B} \boldsymbol{v}}{\boldsymbol{v}^{\top}(\boldsymbol{W}+\boldsymbol{B}) \boldsymbol{v}}=\arg \max _{\boldsymbol{v} \in \mathbb{R}^{d}} \frac{\boldsymbol{v}^{\top} \boldsymbol{B} \boldsymbol{v}}{\boldsymbol{v}^{\top} \boldsymbol{W} \boldsymbol{v}}
$$

with $\boldsymbol{B}$ the empirical between class covariance matrix.

- $\boldsymbol{v}_{1}$ is given by the eigen vector associated with the highest eigen value of $\boldsymbol{W}^{-1} \boldsymbol{B}$
- The remaining discriminant components are obtained through the remaining eigen vectors of $\boldsymbol{W}^{-1} \boldsymbol{B}$
Remark: The number of non null eigen-values of $\boldsymbol{W}^{-1} \boldsymbol{B}$ is at most $\min (d, K-1)$.

Classification rule If components within class variance is scaled to one, classification can be performed by nearest class center with the Euclidean distance on the projected data.

## Illustration of FDA

Factorial Discriminant analysis


## Equivalence between LDA and FDA

Campbell (1984)
FDA $\Leftrightarrow \operatorname{LDA}$ s.c. $\operatorname{rank}\left(\left\{\mu_{1}, \ldots, \mu_{K}\right\}\right)=p$, where $p \leq K-1$.
Reparametrization of LDA under constrains

$$
\boldsymbol{x} \left\lvert\, z_{k}=1 \sim \mathcal{N}_{d}\left(\boldsymbol{A}\binom{\nu_{k}}{\gamma}, \boldsymbol{A A}^{\top}\right)\right.
$$

where $\nu_{k} \in \mathbb{R}^{p}, \gamma \in \mathbb{R}^{d-p}$ and $\boldsymbol{A} \in \mathcal{M}_{d, d}(\mathbb{R})$.
Generative interpretation
Let $\boldsymbol{y} \in \mathbb{R}^{p}$ and $\boldsymbol{u} \in \mathbb{R}^{d-p}$. The generative model is:

- Draw $\boldsymbol{z}: \mathbf{z} \sim \operatorname{Mult}\left(1 ; \pi_{1}, \ldots, \pi_{K}\right)$
- Draw $\boldsymbol{y}|\boldsymbol{z}: \boldsymbol{y}| z_{k}=1 \sim \mathcal{N}_{p}\left(\nu_{k}, I_{p}\right)$
- Draw $\boldsymbol{u}: \boldsymbol{u} \sim \mathcal{N}_{d-p}\left(\gamma, \boldsymbol{I}_{\boldsymbol{d}-p}\right)$
- Compute $\boldsymbol{x}$ based on $\boldsymbol{y}$ and $\boldsymbol{u}$ :

$$
\boldsymbol{x}=\boldsymbol{A}\binom{\boldsymbol{y}}{\boldsymbol{u}}
$$



## Posterior class membership probabilities

Posterior probabilities
Let $\boldsymbol{V} \in \mathcal{M}_{p, d}(\mathbb{R})$ and $\boldsymbol{R} \in \mathcal{M}_{d-p, d}(\mathbb{R})$ the matrices such that
$\boldsymbol{y}=\boldsymbol{V} \boldsymbol{x}, \boldsymbol{u}=\boldsymbol{R} \boldsymbol{x}$. Obviously $\binom{\boldsymbol{V}}{\boldsymbol{R}}=\boldsymbol{A}^{-1}$.

$$
p\left(z_{k}=1 \mid \boldsymbol{x}\right)=p\left(z_{k}=1 \mid \boldsymbol{y}, \boldsymbol{u}\right)=p\left(z_{k}=1 \mid \boldsymbol{y}\right)=p\left(z_{k}=1 \mid \boldsymbol{V} \boldsymbol{x}\right) .
$$

$V x$ conveys all the clustering information.

$$
p\left(z_{k}=1 \mid \boldsymbol{x}\right)=\frac{\pi_{k} \phi_{p}\left(\boldsymbol{V} \boldsymbol{X} ; \boldsymbol{\nu}_{k}, \boldsymbol{I}_{p}\right)}{\sum_{k^{\prime}=1}^{K} \pi_{k^{\prime}} \phi_{p}\left(\boldsymbol{V} \boldsymbol{X} ; \boldsymbol{\nu}_{k^{\prime}}, \boldsymbol{I}_{p}\right)} .
$$

## Parameters estimation

$$
\begin{aligned}
\ell(\boldsymbol{\theta} ; \mathbf{x}, \mathbf{z})= & n \log \left|\operatorname{det}\binom{\boldsymbol{V}}{\boldsymbol{R}}\right|-\frac{1}{2} \overbrace{\sum_{i=1}^{n} \sum_{k=1}^{K} z_{i k}\left\|\boldsymbol{V} \boldsymbol{x}_{i}-\boldsymbol{\nu}_{k}\right\|^{2}}^{\text {classifiying }} \\
& -\frac{1}{2} \underbrace{\sum_{i=1}^{n}\left\|\boldsymbol{R} \boldsymbol{x}_{i}-\gamma\right\|^{2}}_{\text {not classifying }}-\frac{n}{2} \log (2 \pi)+\sum_{i=1}^{n} \sum_{k=1}^{K} z_{i k} \log \left(\pi_{k}\right) .
\end{aligned}
$$

Maximum likelihood estimation

- $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{R}}$ are obtained through eigen value decomposition of $\boldsymbol{W}^{-1} \boldsymbol{B}$.
- $\hat{\boldsymbol{\nu}}_{1}, \ldots, \hat{\boldsymbol{\nu}}_{K}$ and $\hat{\gamma}$ explicit given $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{R}}$.


## Application in the clustering setting (z unknown)

EM algorithm
Until convergence iterate the following steps:

- E step: Compute $t_{i k}^{(r+1)}=p\left(\boldsymbol{z}_{i} \mid \boldsymbol{x}_{i} ; \boldsymbol{\theta}^{(r)}\right)$
- M step:
- Compute $t_{i k}^{(r+1)}$-weighted versions of $\boldsymbol{W}, \boldsymbol{B}$ and $\overline{\boldsymbol{x}}_{k}$
- Deduce $\boldsymbol{V}^{(r+1)}$ and $\boldsymbol{R}^{(r+1)}, \boldsymbol{\nu}_{1}^{(r+1)}, \ldots, \boldsymbol{\nu}_{K}^{(r+1)}$ and $\gamma^{(r+1)}$

Remark
When $p<\min (K-1, d)$, not equivalent to perform a standard EM algorithm and then perform FDA at the end of the EM algorithm.

## Presentation of the model (supervised setting)

Model description

- $H$ independent class variables
- $\boldsymbol{z}^{1}, \ldots, \boldsymbol{z}^{H}$ with $K_{1}, \ldots, K_{H}$ modalities
- $p\left(z_{k}^{h}=1\right)=\pi_{k}^{h}$
- $H$ classifying latent variables
- $\boldsymbol{y}^{h} \in \mathbb{R}^{p_{h}}$ with $p_{h} \leq K_{h}-1$
- $\boldsymbol{y}^{h} \mid z_{k}^{h}=1 \sim \mathcal{N}_{p_{h}}\left(\boldsymbol{\nu}_{k}^{h}, I_{p_{n}}\right)$
- Not classifying latent variables
- $\boldsymbol{u} \in \mathbb{R}^{d-p_{\bullet}}$ with $p_{\bullet}=\sum_{h=1}^{H} p_{h}$
- $\boldsymbol{u} \sim \mathcal{N}_{d-p_{\mathbf{e}}}\left(\gamma, l_{d-p_{\mathbf{0}}}\right)$
- $\boldsymbol{x}$ defined by $\boldsymbol{x}=\left(\begin{array}{c}\boldsymbol{V}_{1} \\ \vdots \\ \boldsymbol{V}_{H} \\ \boldsymbol{R}\end{array}\right)^{-1}\left(\begin{array}{c}\boldsymbol{y}^{1} \\ \vdots \\ \boldsymbol{y}^{H} \\ \boldsymbol{u}\end{array}\right) \quad \begin{aligned} & \text { Figure } \\ & H=2\end{aligned}$


## Possible models and number of parameters

Remarks

- Need to choose $H, K_{1}, \ldots, K_{H}$ and $p_{1}, \ldots, p_{H}$ : possible selection by model choice.
- $\boldsymbol{V}_{1}, \ldots, \boldsymbol{V}_{H}$ and $\boldsymbol{R}$ estimated up to isometric transformations.

Number of parameters
$\operatorname{dim}(\Theta)=\sum_{h=1}^{H}\left(K_{h}-1\right)+\sum_{h=1}^{H} \frac{p_{h}\left(2 K_{h}+2 d-p_{h}+1\right)}{2}+\frac{\left(d-p_{\bullet}\right)\left(d+p_{\bullet}+3\right)}{2}$

## Estimation in the supervised setting

- The likelihood cannot be maximized directly $\Rightarrow$ alternate optimisation.
- Constrain $\boldsymbol{V}_{h}^{(r+1)}$ and $\boldsymbol{R}^{(r+1)}$ to be linear combinations of $\boldsymbol{V}_{h}^{(r)}$ and $\boldsymbol{R}^{(r)}$, the parameters of other clustering dimension being fixed:

$$
\binom{\boldsymbol{V}_{h}^{(r+1)}}{\boldsymbol{R}^{(r+1)}}=\boldsymbol{M}\binom{\boldsymbol{V}_{h}^{(r)}}{\boldsymbol{R}^{(r)}}=\binom{\boldsymbol{M}_{1}}{\boldsymbol{M}_{2}}\binom{\boldsymbol{V}_{h}^{(r)}}{\boldsymbol{R}^{(r)}},
$$

- by denoting $\boldsymbol{y}^{h^{(r)}}=\boldsymbol{V}_{h}^{(r)} \boldsymbol{x}$ and $\boldsymbol{u}^{(r)}=\boldsymbol{R}^{(r)} \boldsymbol{x}$,
- finding $\boldsymbol{M}$ and other parameters is equivalent to perform probabilistic FDA on the data $\binom{\boldsymbol{y}^{h^{(r)}}}{\boldsymbol{u}^{(r)}}$


## Estimation in the unsupervised setting

Similar to the supervised setting except that the data at each iteration are now weighted by $t_{i k}^{h(r+1)}$ instead of $z_{i k}^{h}$
EM algorithm
Until convergence, for $h \in\{1, \ldots, H\}$ iterate the following steps:

- E step: compute

$$
t_{i k}^{h(r+1)}=p\left(\boldsymbol{z}_{i}^{h} \mid \boldsymbol{x}_{i} ; \boldsymbol{\theta}^{(r)}\right)=\frac{\pi_{k} \phi_{p_{h}}\left(\boldsymbol{y}_{i}^{h(r)} ; \boldsymbol{\nu}_{k}^{h(r)}, \boldsymbol{I}_{p_{h}}\right)}{\sum_{k^{\prime}=1}^{K} \pi_{k^{\prime}} \phi_{p_{h}}\left(\boldsymbol{y}_{i}^{h(r)} ; \boldsymbol{\nu}_{k^{\prime}}^{h(r)}, \boldsymbol{I}_{p_{h}}\right)}
$$

- M step: update $\boldsymbol{V}_{h}, \boldsymbol{R}, \boldsymbol{\nu}_{1}^{h}, \ldots, \boldsymbol{\nu}_{K_{h}}^{h}, \pi_{1}^{h}, \ldots, \pi_{K_{h}}^{h}$ and $\gamma$


## Simulated data set: illustration

Comparison of true and estimated projections


## Simulated data: model choice

- Performed using the BIC criterion

$$
B I C=\ell(\boldsymbol{x} ; \hat{\boldsymbol{\theta}})-\frac{\operatorname{dim}(\Theta)}{2} \log n
$$

- Model collection: $H=2, K_{1} \in\{1, \ldots, 5\}, K_{2} \in\{1, \ldots, 5\}$, $p_{1}=p_{2}=1$

| $K_{1} \backslash K_{2}$ | 1 | 2 | 3 | 4 | 5 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | -1414.69 | -1395.53 | -1397.55 | -1395.65 | -1396.88 |
| 2 |  | -1383.76 | -1384.64 | -1397.17 | -1393.86 |
| 3 |  |  | -1384.05 | -1396.36 | -1395.36 |
| 4 |  |  |  | -1392.30 | -1394.01 |
| 5 |  |  |  |  | -1389.14 |

## Crabs data set

 200 crabs morphological data, 5 morphological variables, 50 males orange, 50 males blue, 50 females orange, 50 females blue.Original data


## PCA on the crab data set

## Data after PCA



## FDA on the crab data set: supervised setting

Data after Factorial Discriminant analysis


## FDA on the crab data set: unsupervised setting

Data after Factorial Discriminant Analysis


## Principal clustering components: $H=2, K_{1}=K_{2}=2$, <br> $$
p_{1}=p_{2}=1
$$

Data on the principal clustering components


## Principal clustering components: model choice

- Performed using the BIC criterion
- Model collection: $H=2, K_{1} \in\{1, \ldots, 5\}, K_{2} \in\{1, \ldots, 5\}$, $p_{1}=p_{2}=1$

| $K_{1} \backslash K_{2}$ | 1 | 2 | 3 | 4 | 5 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | -313.02 | -252.35 | -244.20 | -250.60 | -249.40 |
| 2 |  | -229.73 | -254.76 | -254.82 | -233.91 |
| 3 |  |  | -248.82 | -255.49 | -233.17 |
| 4 |  |  |  | -238.88 | -231.19 |
| 5 |  |  |  |  | -269.75 |

Principal clustering components: $H=2, K_{1}=4$,

$$
K_{2}=5, p_{1}=p_{2}=1
$$

Data on the principal clustering components


## Conclusion and perspectives

Conclusion

- model combining visualization and clustering with many clustering view points
- possibility to perform model choice

Perspectives

- consider sparse estimates of the $\boldsymbol{R}$ matrix for the high dimensional setting
- extension to the heteroscedastic setting

