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

Vincent Vandewalle

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 11th, 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)

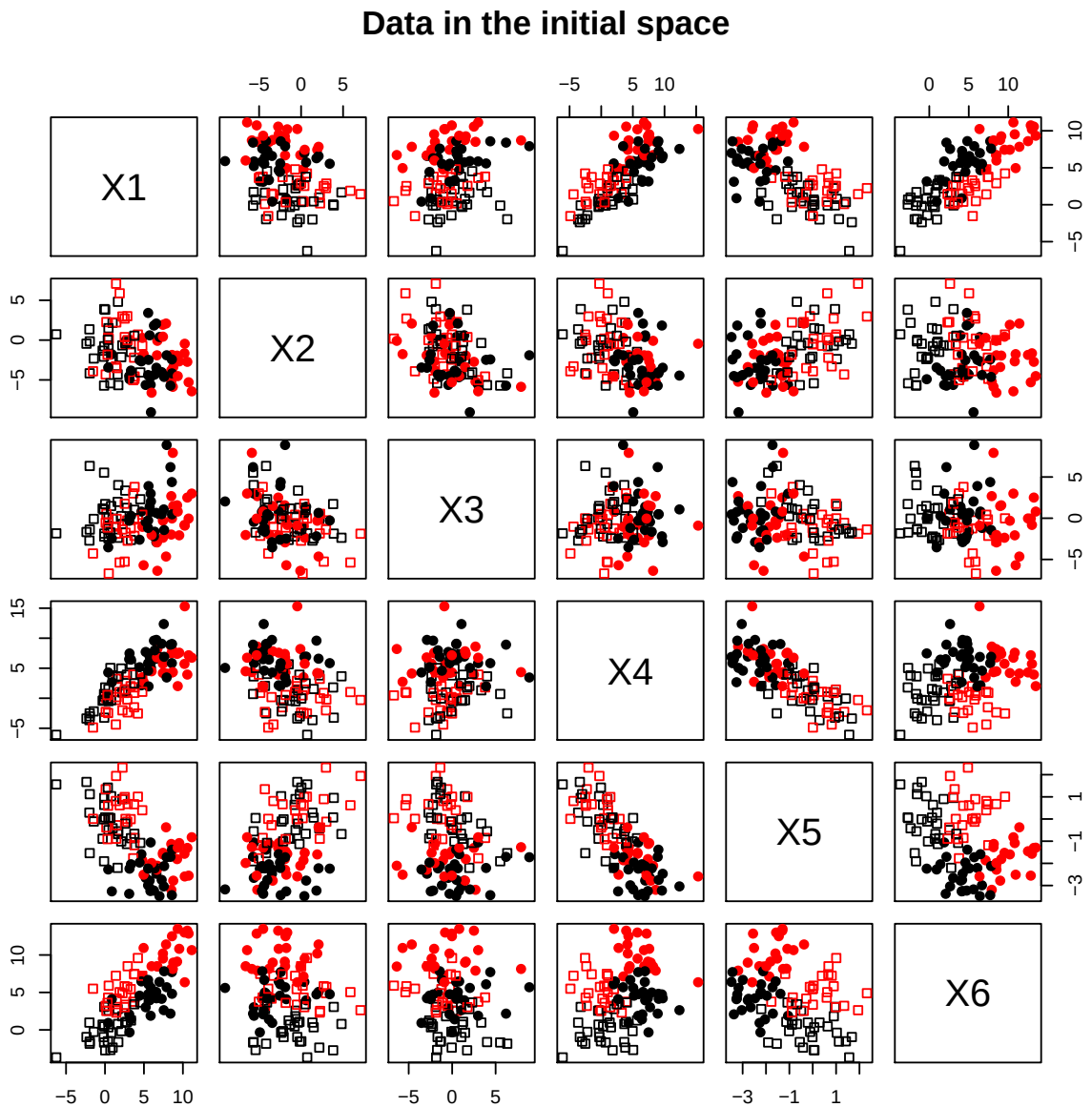
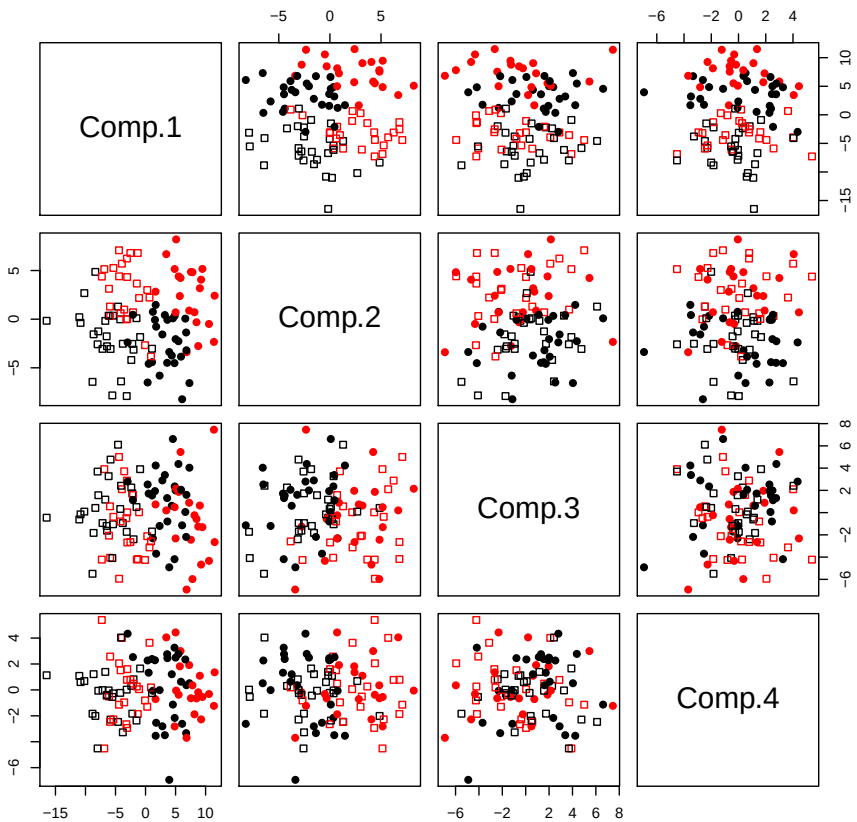
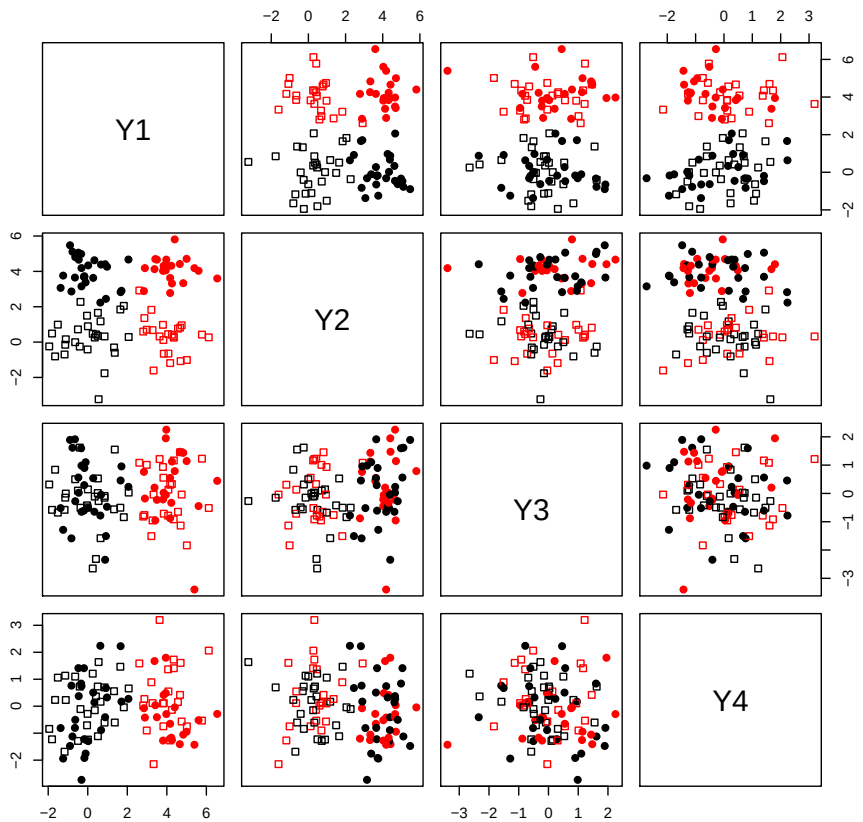


Illustration (2/2)

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 $(\boldsymbol{x}, \boldsymbol{z})$

- $\boldsymbol{z} \sim Mult(\pi_1, \ldots, \pi_K)$ with $\pi_k = p(z_k = 1)$
- $\forall k \in \{1, \ldots, K\}, \quad \boldsymbol{x} | z_k = 1 \sim \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$

Parameters estimation

$\hat{\pi}_k = \frac{n_k}{n}$, $\hat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{x}}_k$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{W}$ (empirical within-class covariance matrix)

Classification

$$p(z_k = 1 | \boldsymbol{x}, \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_k \phi_d(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}})}{\sum_{k'=1}^{K} \hat{\pi}_{k'} \phi_d(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_{k'}, \hat{\boldsymbol{\Sigma}})} \Rightarrow \hat{\boldsymbol{z}} = \underset{k \in \{1,\ldots,K\}}{\text{argmax}} \; \hat{\boldsymbol{\mu}}_k^\top \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_k^\top \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu}_k - \log(\hat{\pi}_k)$$

$\Rightarrow$ 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:

$$\boldsymbol{v}_1 = \arg\max_{\boldsymbol{v}\in\mathbb{R}^d} \frac{\boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{v}}{\boldsymbol{v}^\top(\boldsymbol{W}+\boldsymbol{B})\boldsymbol{v}} = \arg\max_{\boldsymbol{v}\in\mathbb{R}^d} \frac{\boldsymbol{v}^\top \boldsymbol{B}\boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{v}}$$

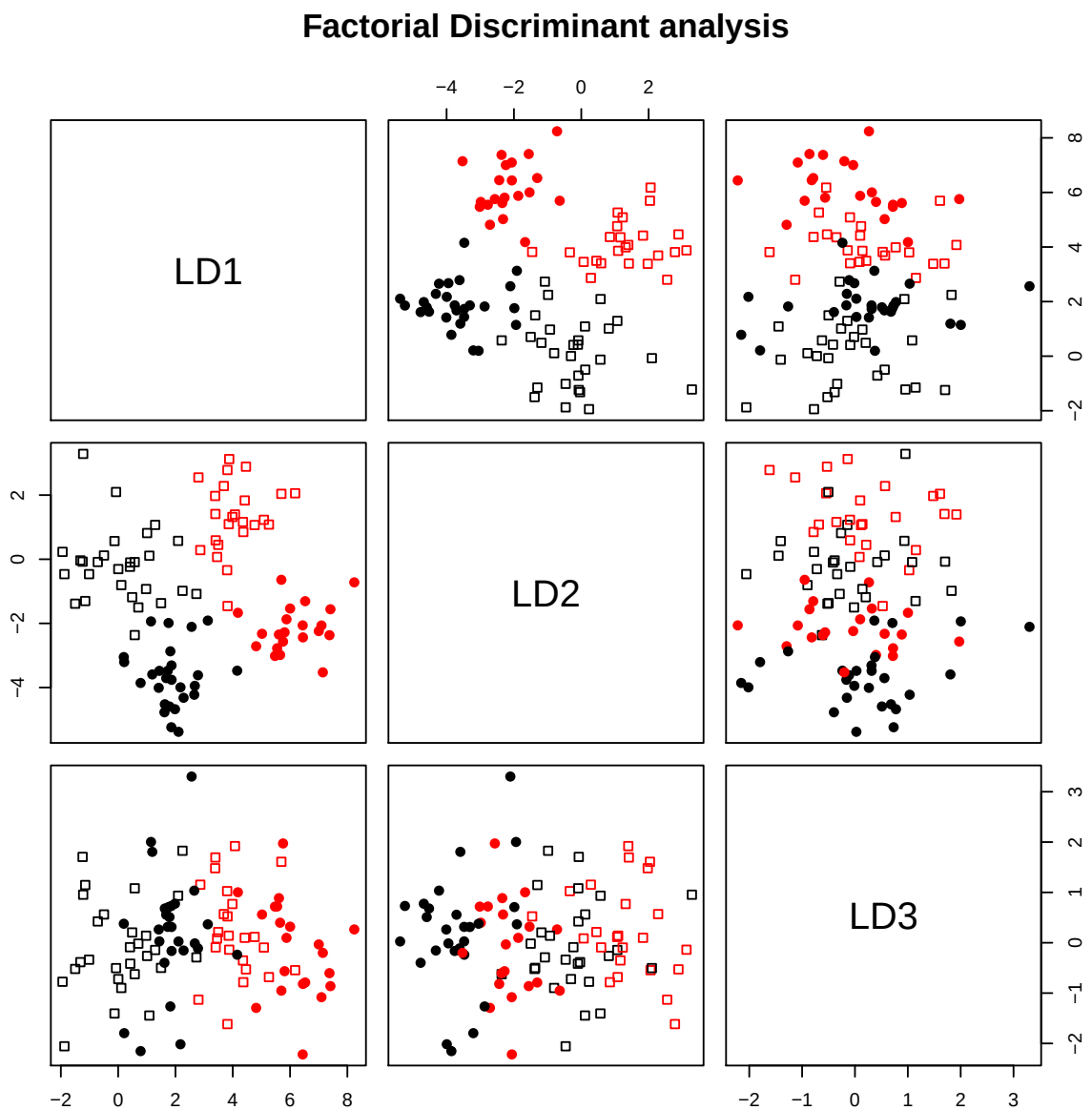
  with $\boldsymbol{B}$ the empirical between class covariance matrix.
- $\boldsymbol{v}_1$ is given by the eigen vector associated with the highest eigen value of $\boldsymbol{W}^{-1}\boldsymbol{B}$
- The remaining discriminant components are obtained through the remaining eigen vectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$

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



Equivalence between LDA and FDA

Campbell (1984)

FDA $\Leftrightarrow$ LDA s.c. rank$(\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}) = p$, where $p \leq K - 1$.

Reparametrization of LDA under constrains

$$\boldsymbol{x}|z_k = 1 \sim \mathcal{N}_d\left(\boldsymbol{A}\begin{pmatrix}\boldsymbol{\nu}_k \\ \boldsymbol{\gamma}\end{pmatrix}, \boldsymbol{A}\boldsymbol{A}^\mathsf{T}\right)$$

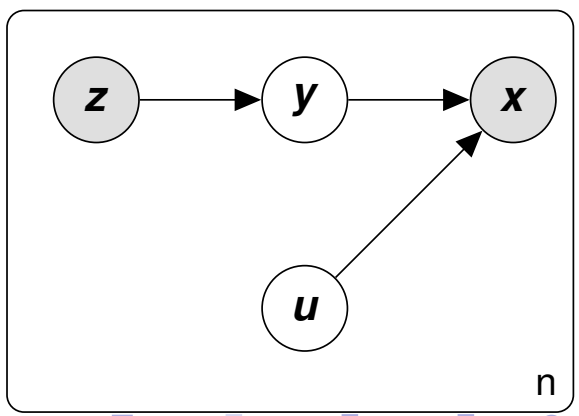
where $\boldsymbol{\nu}_k \in \mathbb{R}^p$, $\boldsymbol{\gamma} \in \mathbb{R}^{d-p}$ and $\boldsymbol{A} \in \mathcal{M}_{d,d}(\mathbb{R})$.

Generative interpretation

Let $\boldsymbol{y} \in \mathbb{R}^p$ and $\boldsymbol{u} \in \mathbb{R}^{d-p}$. The generative model is:

- Draw $\boldsymbol{z}$: $\boldsymbol{z} \sim \mathit{Mult}(1; \pi_1, \ldots, \pi_K)$
- Draw $\boldsymbol{y}|\boldsymbol{z}$: $\boldsymbol{y}|z_k = 1 \sim \mathcal{N}_p(\boldsymbol{\nu}_k, \boldsymbol{I}_p)$
- Draw $\boldsymbol{u}$: $\boldsymbol{u} \sim \mathcal{N}_{d-p}(\boldsymbol{\gamma}, \boldsymbol{I}_{d-p})$
- Compute $\boldsymbol{x}$ based on $\boldsymbol{y}$ and $\boldsymbol{u}$:

$$\boldsymbol{x} = \boldsymbol{A}\begin{pmatrix}\boldsymbol{y} \\ \boldsymbol{u}\end{pmatrix}$$



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 $\begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{R} \end{pmatrix} = \boldsymbol{A}^{-1}$.

$$p(z_k = 1|\boldsymbol{x}) = p(z_k = 1|\boldsymbol{y}, \boldsymbol{u}) = p(z_k = 1|\boldsymbol{y}) = p(z_k = 1|\boldsymbol{V}\boldsymbol{x}).$$

$\boldsymbol{V}\boldsymbol{x}$ conveys all the clustering information.

$$p(z_k = 1|\boldsymbol{x}) = \frac{\pi_k \phi_p(\boldsymbol{V}\boldsymbol{x}; \boldsymbol{\nu}_k, \boldsymbol{I}_p)}{\sum_{k'=1}^{K} \pi_{k'} \phi_p(\boldsymbol{V}\boldsymbol{x}; \boldsymbol{\nu}_{k'}, \boldsymbol{I}_p)}.$$

Parameters estimation

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = n \log \left| \det \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{R} \end{pmatrix} \right| - \overbrace{\frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \| \boldsymbol{V}\boldsymbol{x}_i - \boldsymbol{\nu}_k \|^2}^{\textbf{classifiying}}$$
$$\underbrace{-\frac{1}{2} \sum_{i=1}^{n} \| \boldsymbol{R}\boldsymbol{x}_i - \gamma \|^2}_{\textbf{not classifying}} - \frac{n}{2} \log(2\pi) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k).$$

Maximum likelihood estimation

- $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{R}}$ are obtained through eigen value decomposition of $\boldsymbol{W}^{-1}\boldsymbol{B}$.
- $\hat{\nu}_1, \ldots, \hat{\nu}_K$ and $\hat{\gamma}$ explicit given $\hat{\boldsymbol{V}}$ and $\hat{\boldsymbol{R}}$.

Application in the clustering setting ($\boldsymbol{z}$ unknown)

EM algorithm

Until convergence iterate the following steps:

- E step: Compute $t_{ik}^{(r+1)} = p(\boldsymbol{z}_i | \boldsymbol{x}_i; \boldsymbol{\theta}^{(r)})$
- M step:
  - Compute $t_{ik}^{(r+1)}$-weighted versions of $\boldsymbol{W}$, $\boldsymbol{B}$ and $\bar{\boldsymbol{x}}_k$
  - 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
  - $p(z_k^h = 1) = \pi_k^h$
- H classifying latent variables
  - $\boldsymbol{y}^h \in \mathbb{R}^{p_h}$ with $p_h \leq K_h - 1$
  - $\boldsymbol{y}^h | z_k^h = 1 \sim \mathcal{N}_{p_h}(\boldsymbol{\nu}_k^h, I_{p_h})$
- Not classifying latent variables
  - $\boldsymbol{u} \in \mathbb{R}^{d-p_\bullet}$ with $p_\bullet = \sum_{h=1}^{H} p_h$
  - $\boldsymbol{u} \sim \mathcal{N}_{d-p_\bullet}(\gamma, I_{d-p_\bullet})$
- $\boldsymbol{x}$ defined by $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}^1 \\ \vdots \\ \boldsymbol{y}^H \\ \boldsymbol{u} \end{pmatrix}$.

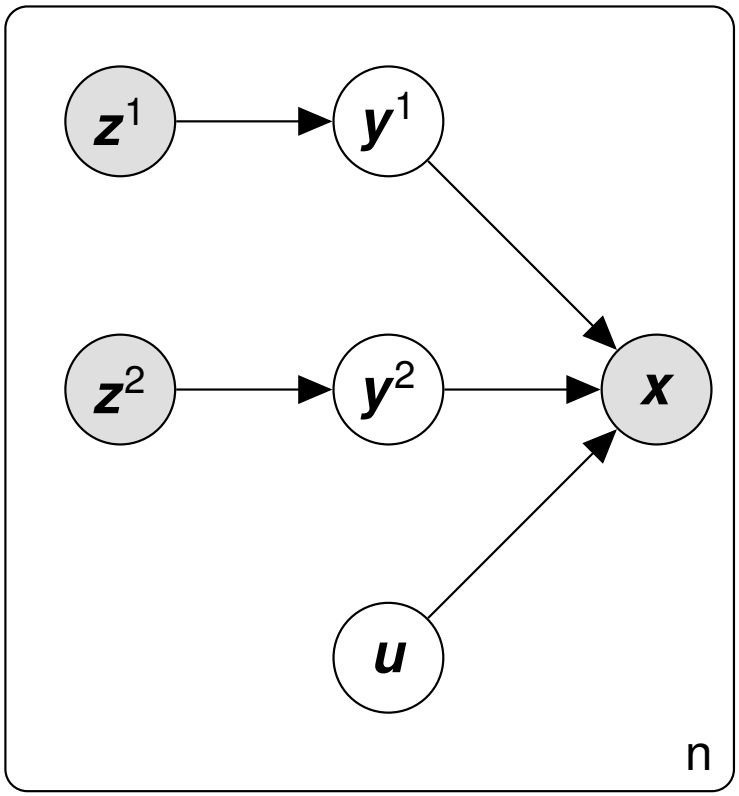


Figure: Illustration for $H = 2$

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

$$\dim(\Theta) = \sum_{h=1}^{H}(K_h - 1) + \sum_{h=1}^{H} \frac{p_h(2K_h + 2d - p_h + 1)}{2} + \frac{(d - p_\bullet)(d + p_\bullet + 3)}{2}$$

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:

$$\begin{pmatrix} \boldsymbol{V}_h^{(r+1)} \\ \boldsymbol{R}^{(r+1)} \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} \boldsymbol{V}_h^{(r)} \\ \boldsymbol{R}^{(r)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}_1 \\ \boldsymbol{M}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_h^{(r)} \\ \boldsymbol{R}^{(r)} \end{pmatrix},$$

- 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 $\begin{pmatrix} \boldsymbol{y}^{h(r)} \\ \boldsymbol{u}^{(r)} \end{pmatrix}$

Estimation in the unsupervised setting

Similar to the supervised setting except that the data at each iteration are now weighted by $t_{ik}^{h\,(r+1)}$ instead of z_{ik}^{h}

EM algorithm

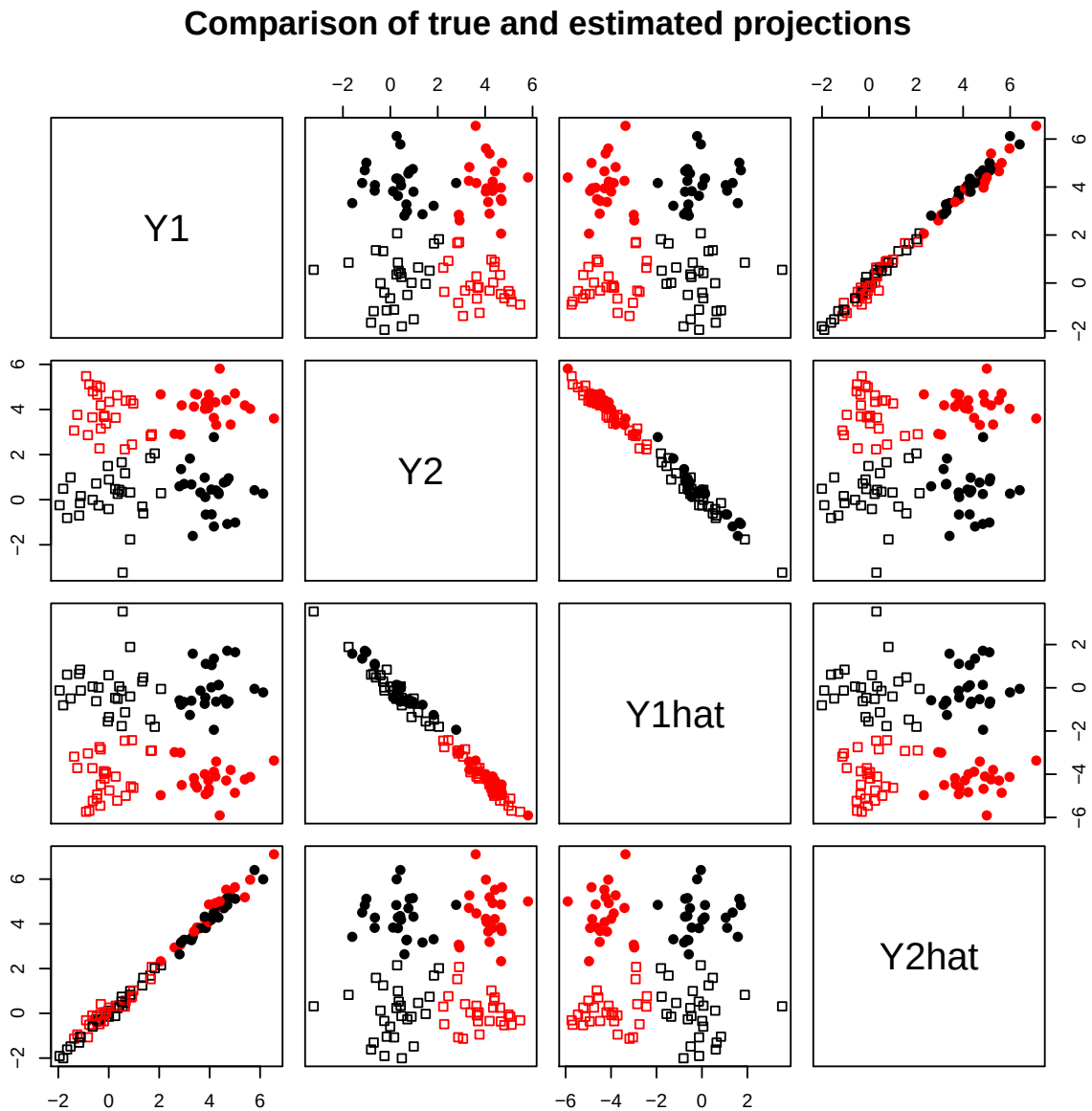
Until convergence, for $h \in \{1, \ldots, H\}$ iterate the following steps:

- E step: compute

$$t_{ik}^{h\,(r+1)} = p(\boldsymbol{z}_i^h | \boldsymbol{x}_i; \boldsymbol{\theta}^{(r)}) = \frac{\pi_k \phi_{p_h}(\boldsymbol{y}_i^{h(r)}; \boldsymbol{\nu}_k^{h(r)}, \boldsymbol{I}_{p_h})}{\sum_{k'=1}^{K} \pi_{k'} \phi_{p_h}(\boldsymbol{y}_i^{h(r)}; \boldsymbol{\nu}_{k'}^{h\,(r)}, \boldsymbol{I}_{p_h})}$$

- 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 γ

Simulated data set: illustration



Simulated data: model choice

- Performed using the BIC criterion

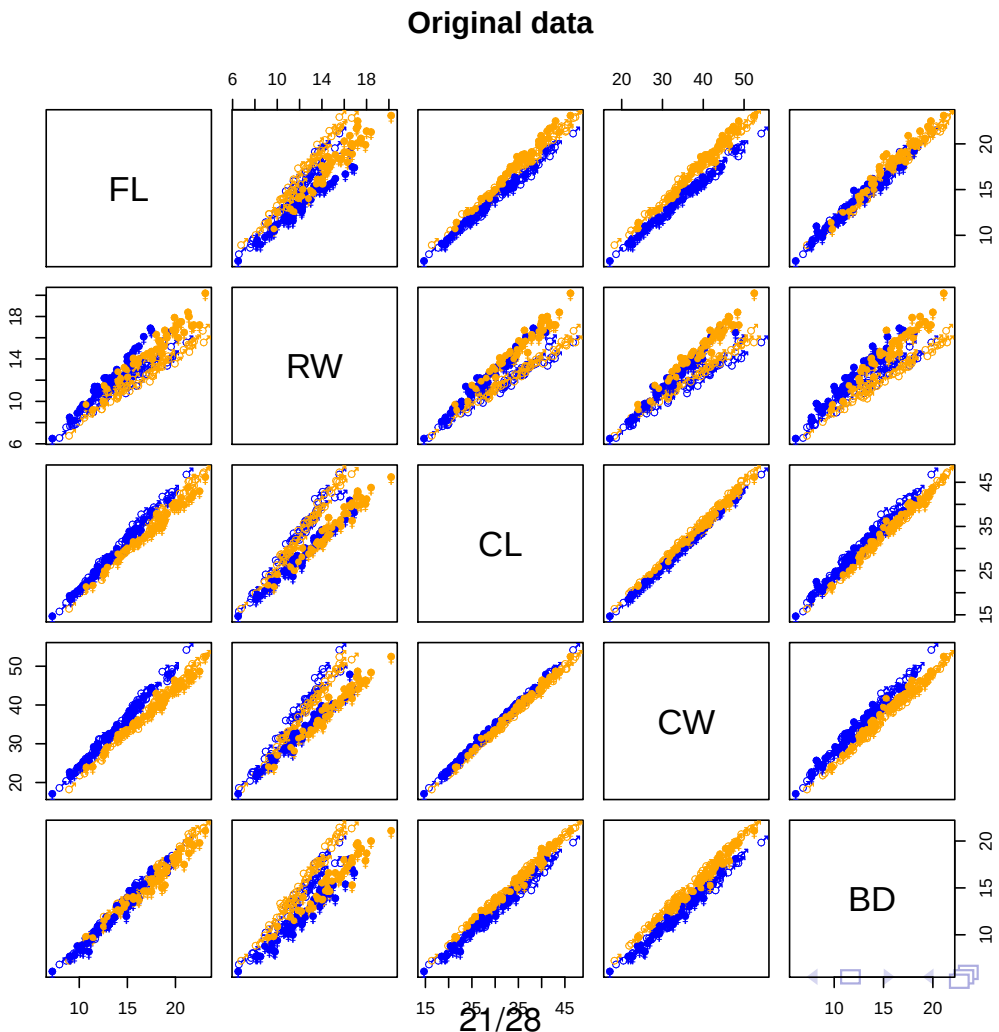
$$BIC = \ell(\boldsymbol{x}; \hat{\boldsymbol{\theta}}) - \frac{\dim(\Theta)}{2} \log n$$

- Model collection: $H = 2$, $K_1 \in \{1, \ldots, 5\}$, $K_2 \in \{1, \ldots, 5\}$, $p_1 = p_2 = 1$

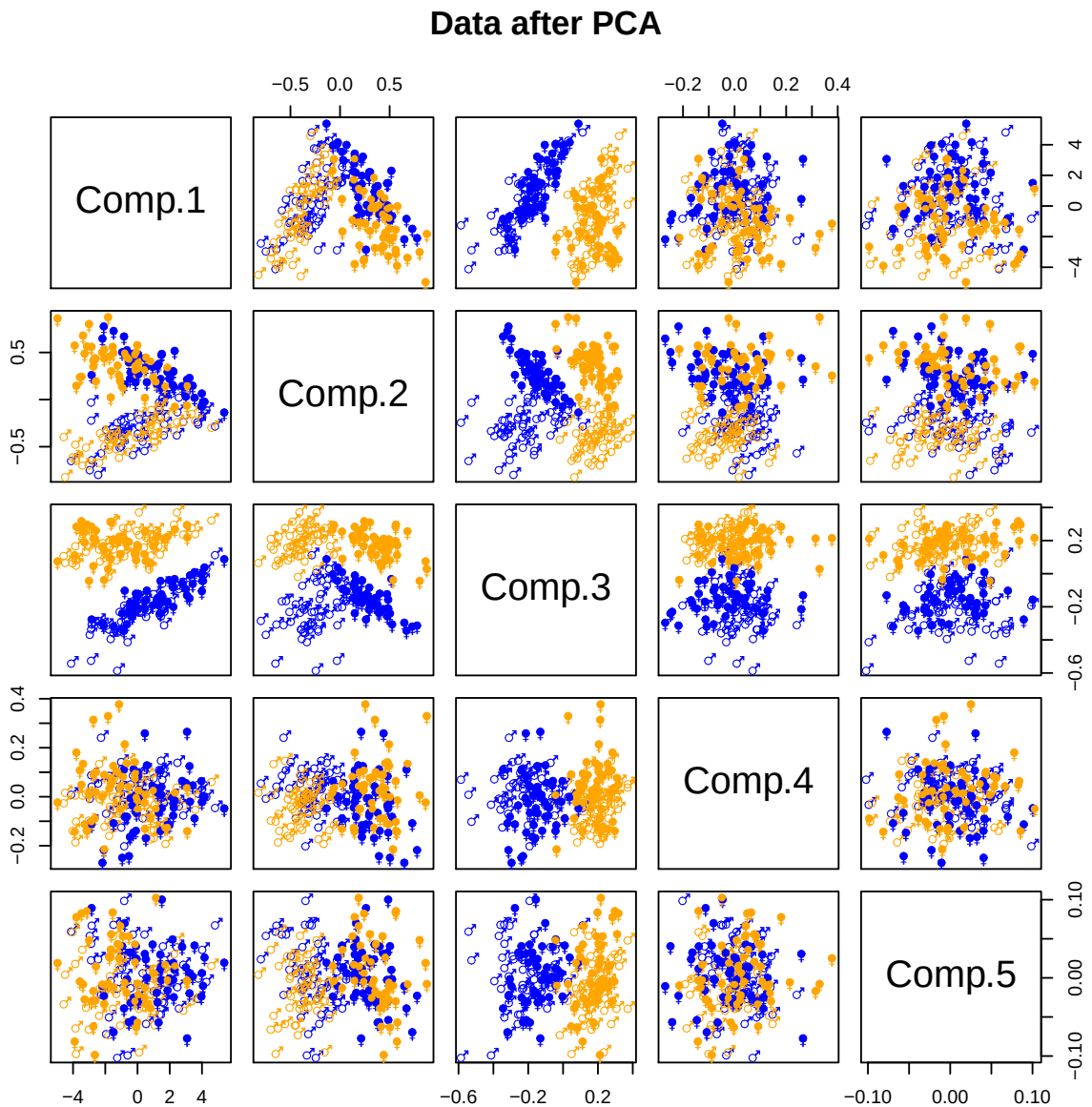
$K_1 \setminus K_2$	1	2	3	4	5
1	-1414.69	-1395.53	-1397.55	-1395.65	-1396.88
2		**-1383.76**	-1384.64	-1397.17	-1393.86
3			-1384.05	-1396.36	-1395.36
4				-1392.30	-1394.01
5					-1389.14

Crabs data set

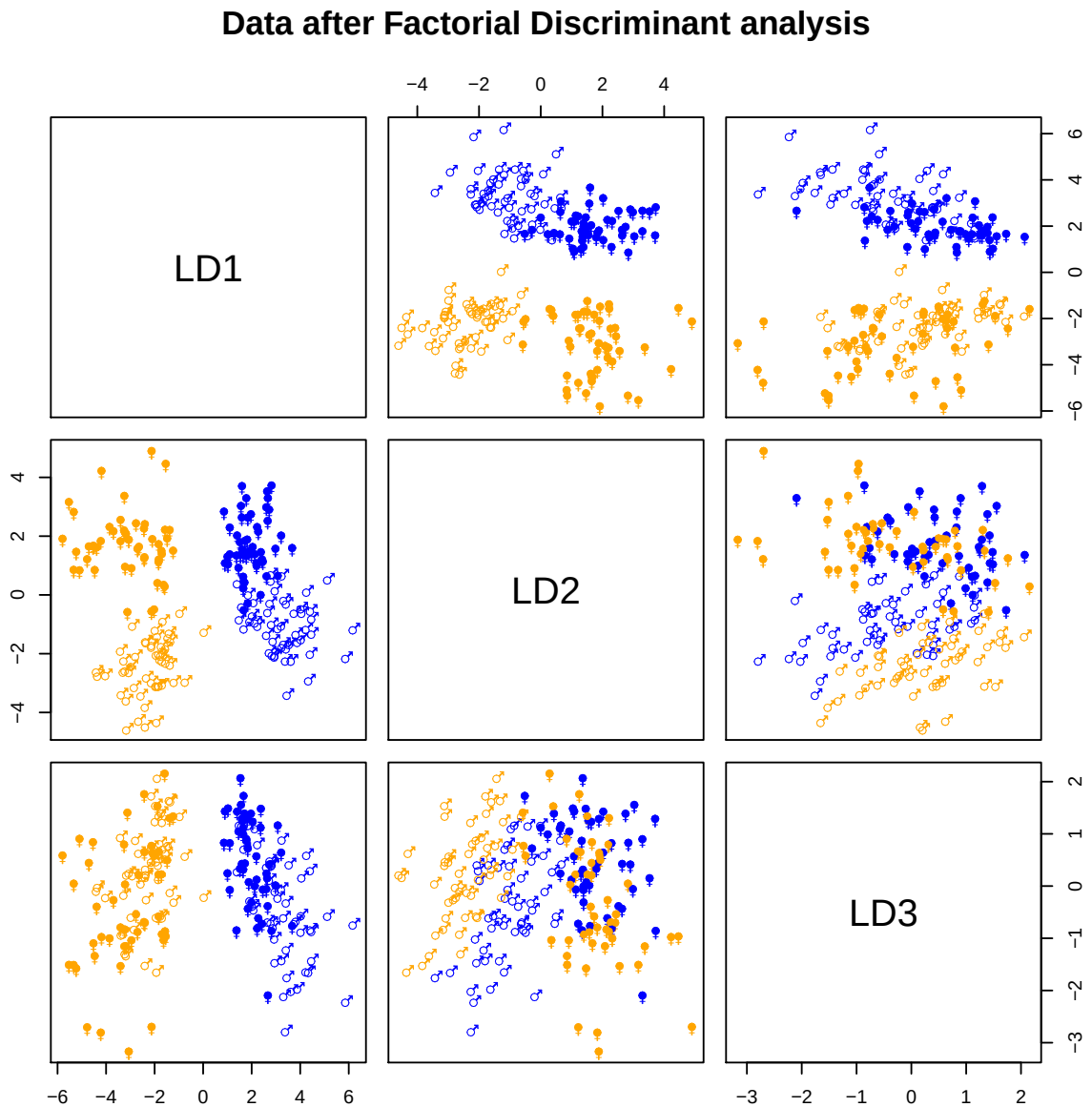
200 crabs morphological data, 5 morphological variables, 50 males orange, 50 males blue, 50 females orange, 50 females blue.



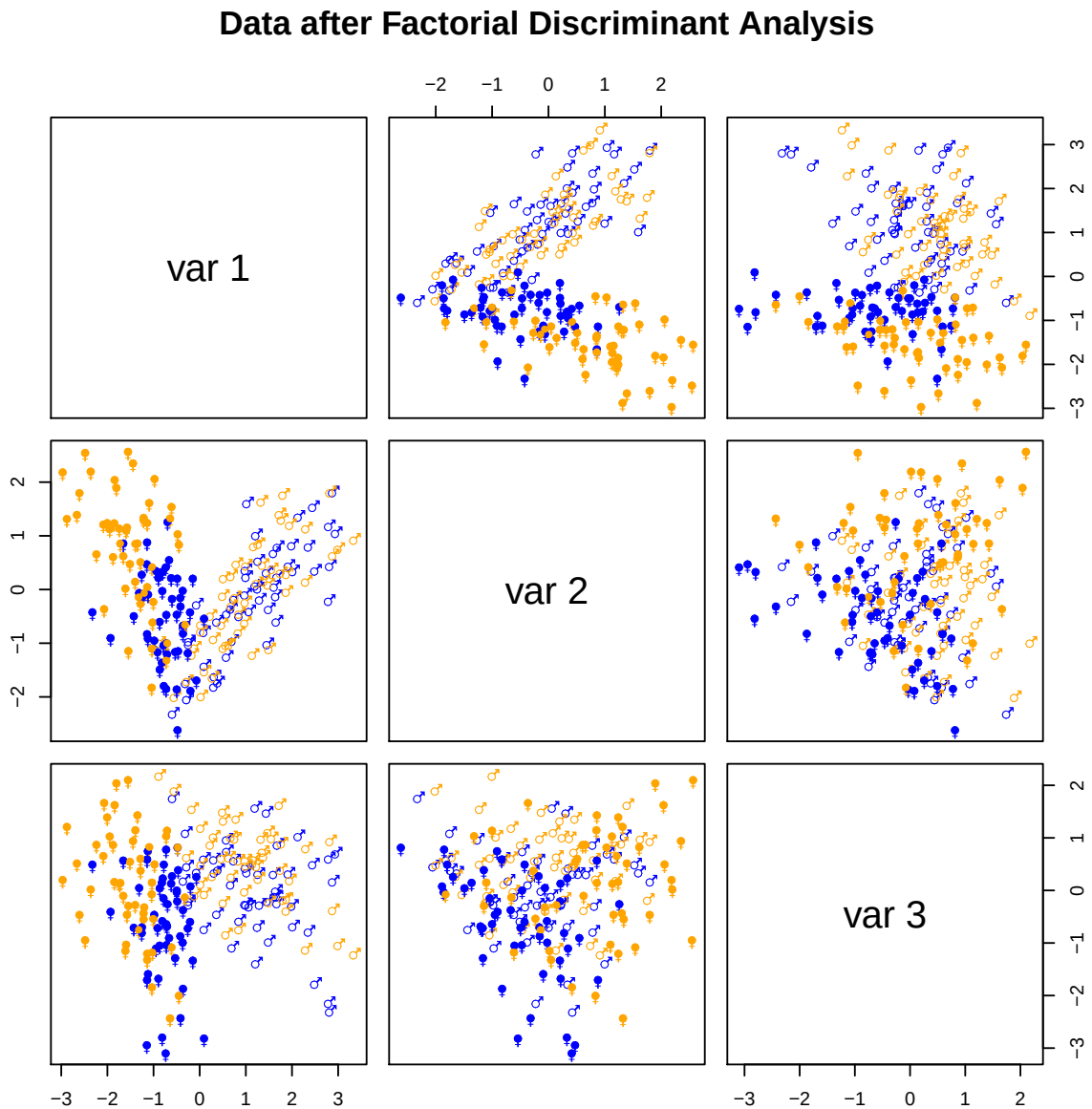
PCA on the crab data set



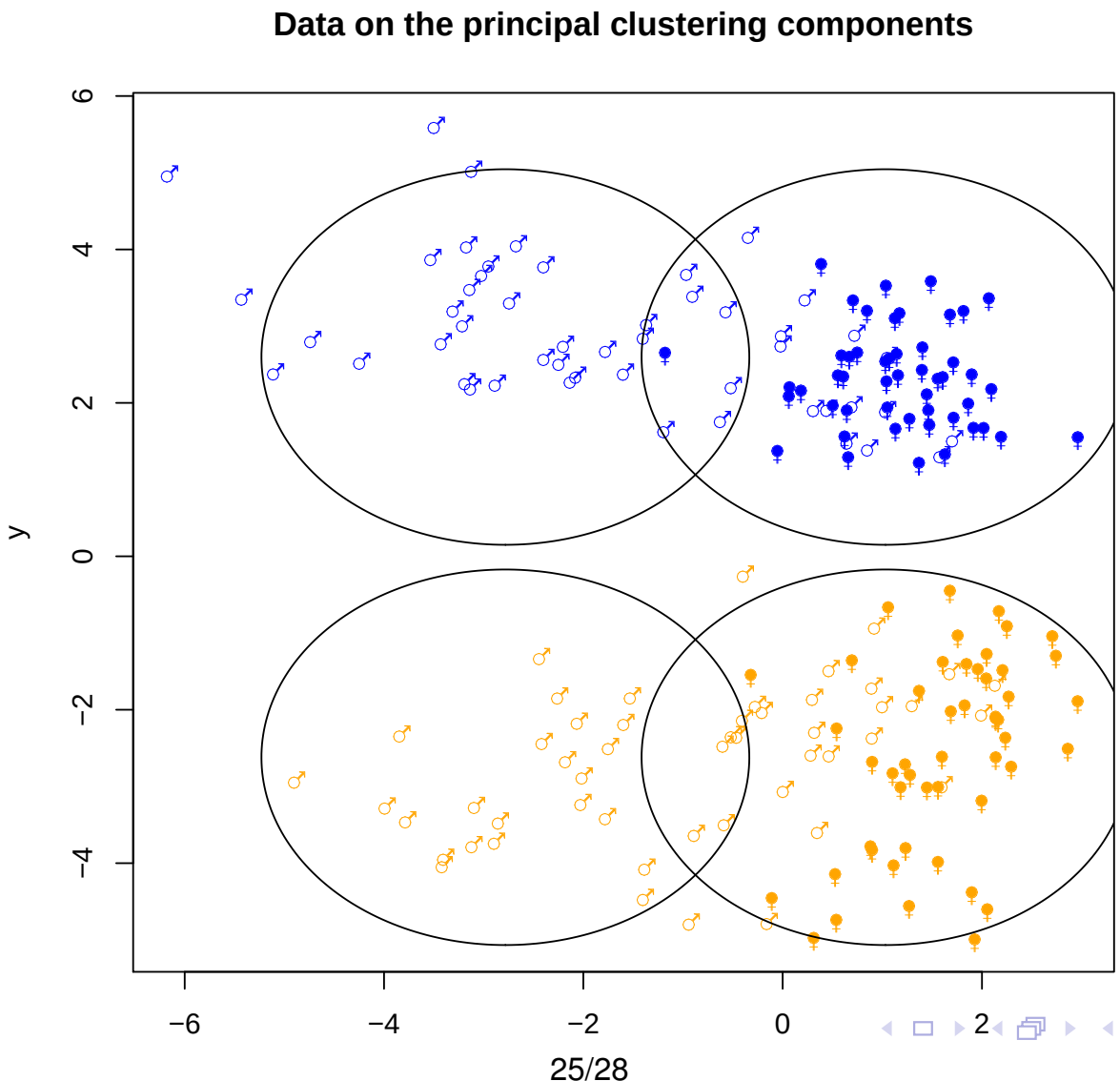
FDA on the crab data set: supervised setting



FDA on the crab data set: unsupervised setting



Principal clustering components: $H = 2$, $K_1 = K_2 = 2$, $p_1 = p_2 = 1$

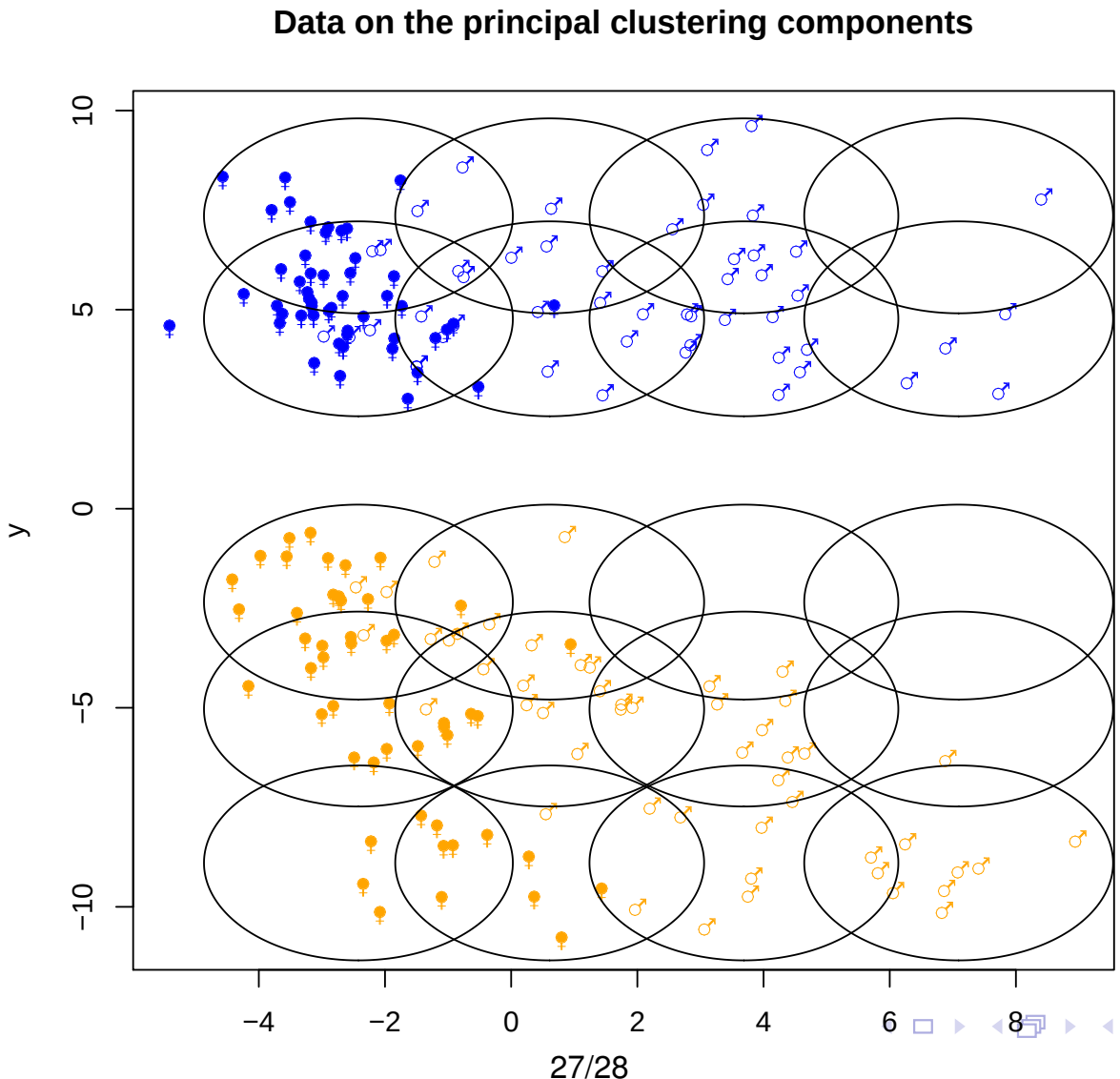


Principal clustering components: model choice

- Performed using the BIC criterion
- Model collection: $H = 2$, $K_1 \in \{1, \ldots, 5\}$, $K_2 \in \{1, \ldots, 5\}$, $p_1 = p_2 = 1$

$K_1 \setminus K_2$	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: $H = 2$, $K_1 = 4$, $K_2 = 5$, $p_1 = p_2 = 1$



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