Simultaneous dimension reduction and multi-objective clustering

Vincent Vandewalle

Simultaneous dimension reduction and multi-objective clustering

Vincent Vandewalle[a,b]

[a]**Équipe d'Accueil 2694** (Université Lille 2)

[b]**Équipe MODAL** (Inria Lille Nord Europe)

Introduction

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.
- Often when visualizing the data one is interested in visualizing clusters on the visualization space, but 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
  - Trevor Hastie [1996]: Reduced rank discriminant analysis.
  - Kumar and Andreou [1998]: Heteroscedastic discriminant analysis.
  - Bouveyron and 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 several clustering variables.
  - Evaluation of clustering methods based on a gold standard partition, but many possible gold standards 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.

Mixture model for multi-objective clustering

Notations

- n data in $\mathbb{R}^d$: $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$.
- H clustering variables: $\mathbf{z} = \{(\boldsymbol{z}_1^1, \ldots, \boldsymbol{z}_1^H), \ldots, (\boldsymbol{z}_n^1, \ldots, \boldsymbol{z}_n^H)\}$, with $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ composed with K_1, ..., K_H modalities, $z_{ik}^h = 1$ if the clustering variable h takes the modality k for unit i and $z_{ik}^h = 0$.

Generative model

For each $i \in \{1, \ldots, n\}$

1. For all $h \in \{1, \ldots, H\}$

(a) Draw $\boldsymbol{z}_i^h$ according to a multinomial distribution with probabilities $p(z_{ik}^h = 1)$ dented by π_k^h.

(b) Draw $\boldsymbol{y}_i^h \in \mathbb{R}^{p_h}$, the vector of clustering variables related to the class variable $\boldsymbol{z}_i^h$, according to

$$\boldsymbol{y}_i^h | z_{ik}^h = 1 \sim \mathcal{N}_{p_h}(\boldsymbol{\nu}_k^h, I_{p_h})$$

where $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the p-variable Gaussian distribution with expectation $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

2. Draw $\boldsymbol{u}_i$ the vector of non-classifying variables

$$\boldsymbol{u}_i \sim \mathcal{N}_{d-p_\bullet}(\boldsymbol{\gamma}, I_{d-p_\bullet}),$$

with $p_\bullet = \sum_{h=1}^H p_h$.

3. Compute the observed data $\boldsymbol{x}_i$ defined by

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_i^1 \\ \vdots \\ \boldsymbol{y}_i^H \\ \boldsymbol{u}_i \end{pmatrix}.$$

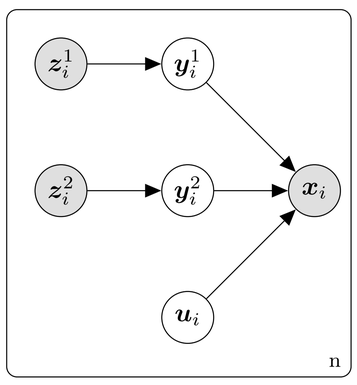

Figure 1: Bayesian network for $H = 2$

Consequences on the posterior membership probabilities

$$p(z_{ik}^h = 1|\boldsymbol{x}_i) = p(z_{ik}^h = 1|\boldsymbol{y}_i^h, \boldsymbol{u}_i) = p(z_{ik} = 1|\boldsymbol{y}_i^h) = p(z_{ik} = 1|\boldsymbol{V}_h\boldsymbol{x}_i) = \frac{\pi_k^h \phi_p(\boldsymbol{V}_h\boldsymbol{x}_i; \boldsymbol{\nu}_k^h, \boldsymbol{I}_{p_h})}{\sum_{k'=1}^{K_h} \pi_{k'}^h \phi_p(\boldsymbol{V}_h\boldsymbol{x}_i; \boldsymbol{\nu}_{k'}^h, \boldsymbol{I}_{p_h})}.$$

with $\phi_p(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ the probability density function of the p-variate Gaussian distribution with expectation $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
Thus only $\boldsymbol{V}_h\boldsymbol{x}_i$ is required to compute the posterior class membership probabilities.

Supervised and unsupervised settings

Supervised setting

- Observed data: $\mathbf{x}$ and $\mathbf{z}$.
- Missing data: $\{(\boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^H)\}_{i=1}^n$ and $\{\boldsymbol{u}_i\}_{i=1}^n$.

Possibility to consider various semi-supervised settings.

Unsupervised setting

- Observed data: $\mathbf{x}$.
- Missing data: $\mathbf{z}$, $\{(\boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^H)\}_{i=1}^n$ and $\{\boldsymbol{u}_i\}_{i=1}^n$.

Identifiability of the model

- $\boldsymbol{\theta} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_H, \boldsymbol{R}, \boldsymbol{\gamma}, \boldsymbol{\nu}_1^1, ..., \boldsymbol{\nu}_{K_H}^H, \pi_1^1, \ldots, \pi_{K_1}^1, \ldots, \pi_{K_H}^H)$.
- Supervised setting: $p_h \leq K_h - 1 \; \forall h \Rightarrow$ model identifiable up to a orthonormal transformation of $\boldsymbol{V}_1$, ..., $\boldsymbol{V}_H$, $\boldsymbol{R}$.
- Unsupervised setting: model identifiable up to a permutation of classifying variables and class labels.

Parameters estimation

Supervised setting

Log-likelihood

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = n \log \left| \det \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix} \right| \underbrace{- \sum_{i=1}^n \sum_{h=1}^H \sum_{k=1}^{K_h} z_{ik}^h \|\boldsymbol{V}_h^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{\nu}_k^h\|^2}_{\text{classifying}} \underbrace{- \sum_{i=1}^n \|\boldsymbol{R}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{\gamma}\|^2}_{\text{non classifying}} + \sum_{i=1}^n \sum_{h=1}^H \sum_{k=1}^{K_h} z_{ik}^h \log(\pi_k^h) - \frac{n}{2}\log(2\pi).$$

- No closed form for the maximum likelihood estimator.
- For only one class variable ($H = 1$), the problem is reduced to linear discriminant analysis with a constraint on the rank of the matrix of centers [Campbell, 1984].

Alternate optimisation

- Until convergence, for each $h \in \{1, \ldots, H\}$
  - All parameters are fixed except $\boldsymbol{V}_h$, $\boldsymbol{R}$, $\boldsymbol{\nu}_1^h$, ..., $\boldsymbol{\nu}_{K_h}^h$, π_1^h, ..., $\pi_{K_h}^h$ and $\boldsymbol{\gamma}$.
  - At iteration $q+1$, $\boldsymbol{V}_h^{(q+1)}$ and $\boldsymbol{R}^{(q+1)}$ are linear combinations of $\boldsymbol{V}_h^{(q)}$ and $\boldsymbol{R}^{(q)}$, i.e.

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

  with $\boldsymbol{M} \in \mathcal{M}_{d-p_\bullet+p_h, d-p_\bullet+p_h}(\mathbb{R})$, $\boldsymbol{M}_1 \in \mathcal{M}_{p_h, d-p_\bullet+p_h}(\mathbb{R})$ and $\boldsymbol{M}_2 \in \mathcal{M}_{d-p_\bullet, d-p_\bullet+p_h}(\mathbb{R})$.
  - Let define $\boldsymbol{y}_i^{h(q)}$ and $\boldsymbol{u}_i^{(q)}$ as $\begin{pmatrix} \boldsymbol{y}_i^{h(q)} \\ \boldsymbol{u}_i^{(q)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_h^{(q)} \\ \boldsymbol{R}^{(q)} \end{pmatrix} \boldsymbol{x}_i$,
  - Then maximisation of the likelihood on $\boldsymbol{M}$, $\boldsymbol{\nu}_1^h$, ..., $\boldsymbol{\nu}_{K_h}^h$, π_1^h, ..., $\pi_{K_h}^h$ and $\boldsymbol{\gamma}$, all the other parameters being fixed, reduced to the reduced rank linear discriminant analysis on $\{(\boldsymbol{y}_i^{h(q)^{\mathrm{T}}}, \boldsymbol{u}_i^{(q)^{\mathrm{T}}})\}_{i=1}^n$.

Unsupervised setting

Principle

- $\mathbf{z}$ unknown $\Rightarrow$ EM algorithm used to "reconstitute" the missing class variables.
- Algorithm similar to the supervised setting but data now weighted by $t_{ik}^{h\,(q+1)} = p(z_{ik}^h = 1|\boldsymbol{x}_i; \boldsymbol{\theta}^{(q)})$ defined below.

EM algorithm

- Until convergence
  - For $h \in \{1, \ldots, H\}$
    - E step: compute

$$t_{ik}^{h\,(q+1)} = \frac{\pi_k^h \phi_{p_h}(\boldsymbol{y}_i^{h(q)}; \boldsymbol{\nu}_k^{h(q)}, \boldsymbol{I}_{p_h})}{\sum_{k'=1}^K \pi_{k'}^h \phi_{p_h}(\boldsymbol{y}_i^{h(q)}; \boldsymbol{\nu}_{k'}^{h(q)}, \boldsymbol{I}_{p_h})}.$$

    - **M step:** compute $\pi_1^{h(q+1)}, \ldots, \pi_1^{H(q+1)}$, $\boldsymbol{V}_h^{(q+1)}$, $\boldsymbol{R}^{(q+1)}$, $\boldsymbol{\gamma}^{(q+1)}$ and $\boldsymbol{\nu}_1^{h(q+1)}, \ldots, \boldsymbol{\nu}_{K_h}^{h\,(q+1)}$ similarly to the supervised setting using the weights $t_{ik}^{h\,(q+1)}$.

Remarks: Properly speaking the presented algorithm is a generalized EM algorithm (GEM) since at each iteration the expectation is not maximized but only increased. As all EM algorithms it is sensitive to starting values. The algorithm can be started with different values based for instance on random projections and then performing a clustering on each projection.

Illustration on cabs dataset

Data from Campbell and Mahon [1974]

- Five morphological variables on 200 crabs : Frontal lobe size, Rear width, Carapace length, Carapace width, Body depth.
- Two categorical variables : sex (male or female) and subspecies (blue or orange), 50 individuals for each variables crossing.
- Data are presented on the three first components of the normalized PCA in Figure 2. The first component does not allow to well discriminate the clusters, whereas components 2 and 3 allow respectively to well discriminate the sex and the subspecies.
- In unsupervised classification one class variable representing the sex and one class variable representing the subspecies would be expected.

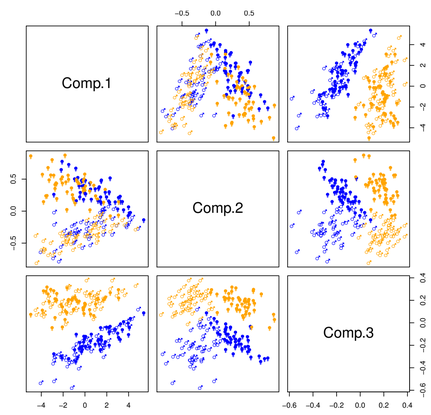

Figure 2: Crab dataset on the three first components of the PCA

Fitted model

- Model fitted for $H = 2$ class variables, each responsible of one classifying variable in dimension 1 ($p_1 = p_2 = 1$).
- Model selection with BIC for $K_1 \in \{1, \ldots, 5\}$ and $K_2 \in \{1, \ldots, 5\}$.

$K_1 \setminus K_2$	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

- Visualization of selected model presented on Figure 3, the first component well separates the males from the females, and the second component well separates the blue crabs from the oranges ones.

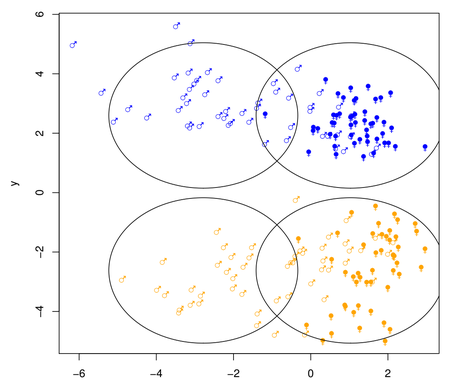

Figure 3: Crab dataset on the two clustering components

Bibliography

Charles Bouveyron and Camille Brunet. Simultaneous model-based clustering and visualization in the fisher discriminative subspace. *Statistics and Computing*, 22(1):301–324, 2012.

N. A. Campbell. Canonical variate analysis - a general model formulation. *Australian Journal of Statistics*, 26(1): 86–96, 1984.

NA Campbell and RJ Mahon. A multivariate study of variation in two species of rock crab of the genus leptograpsus. *Australian Journal of Zoology*, 22(3):417–425, 1974.

Nagendra Kumar and Andreas G Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech communication*, 26(4):283–297, 1998.

Robert Tibshirani Trevor Hastie. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):155–176, 1996.

Conclusion and perspectives

Conclusion

- Model combining visualization and clustering with several clustering view points.
- Possibility to perform model choice.

Perspectives

- Consider the high dimensional setting.
- Consider variable clustering with respect to the clustering behavior point of view.