The evolution of the k-means algorithm to the present day

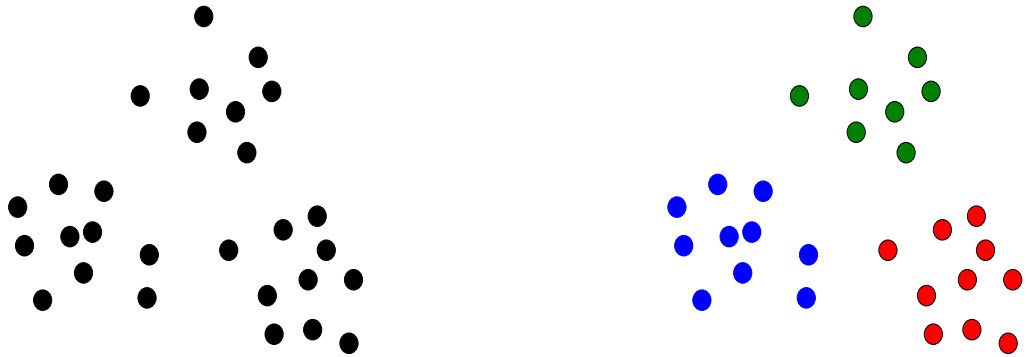
Mohamed Nadif

Centre Borelli UMR9010, Université Paris Cité, France

Outline

Clustering

- Aim: Organization of data into homogeneous subsets "clusters" or "classes"
- It seeks to obtained a reduced representation of the initial data by grouping together similar objects



- Clustering is an important unsupervised learning
- Terminoloy can depend on the field:
  - Taxonomy science of clustering of human being
  - Nosology science of clustering of diseases in medicine
- Not confuse with the classification
- It can take different forms: partitions, sequence of encased partitions or hierarchical, overlapping clusters, clusters with high density, fuzzy clusters.
- Many approches and algorithms according the nature of data

This talk is devoted to the partitioning methods or nonhierarchical clustering inspired by kmeans.

Notation

$$\mathbf{X} = (x_{ij}) \text{ of size } (n \times d) \quad \text{and} \quad z_{ik} \in \{0, 1\}$$

	a	b	c	z
i1	0	0	0	?
i2	1	1	1	?
i3	1	0	0	?
i4	0	1	0	?
i5	1	1	1	?
i6	1	1	1	?
i7	1	0	0	?
i8	0	0	0	?
i9	1	0	0	?
i10	1	1	1	?

Clustering $\Rightarrow$

	a	b	c	z	z_{i1}	z_{i2}	z_{i3}
i1	0	0	0	1	1	0	0
i4	0	1	0	1	1	0	0
i8	0	0	0	1	1	0	0
i2	1	1	1	2	0	1	0
i5	1	1	1	2	0	1	0
i6	1	1	1	2	0	1	0
i10	1	1	1	2	0	1	0
i3	1	0	0	3	0	0	1
i7	1	0	0	3	0	0	1
i9	1	0	0	3	0	0	1

- i denotes the indices of rows
- j denotes the indices of columns
- k denotes the indices of clusters
- $\sum_{i=1}^{n} z_{ik} = \#z_k$ denotes the cardinality of the kth cluster

- Problem: Many algorithms and many approaches

Brief history of k-means

- Steinhaus, H. (1956): Sur la division des corps matériels en parties. Bulletin de l'Académie Polonaise des Sciences, Classe III, vol. IV, no. 12, 801-804.
- Forgy E-W. : "Cluster analysis of multivariate data: efficiency versus inter-pretability of classifications"
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probabilities, Vol. 1, pp. 281-296.
- Diday E. (1971). Une nouvelle méthode de classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. Revue de Statistique Appliquée XIX (2), 19-33.
- Bock H-H. (2008). Origins and extensions of the k-means algorithm in cluster analysis. Journal Electronique d'Histoire des Probabilités et de la Statistique Electronic Journal for History of Probability and Statistics,Vol. 4, No. 2.

Description

- We keep the previous notation and we begin by describing the well-know k-means when the set to classify Ω is measured by d continuous variables
- To look for the optimal partition $\mathbf{Z}$ it suffices to minimize the within-cluster variance

$$\mathcal{W}(\mathbf{Z}) = \sum_{k=1}^{g} \sum_{i \in \mathbf{z}_k} ||\boldsymbol{x}_i - \overline{x}_{\mathbf{z}_k}||^2 = \sum_{k=1}^{g} \sum_{i=1}^{n} z_{ik} ||\boldsymbol{x}_i - \overline{x}_{\mathbf{z}_k}||^2 \text{ since } z_{ik} \in \{0, 1\}$$

This minimization of $\mathcal{W}(\mathbf{Z})$ is equivalent to maximize the between-cluster variance

$$\mathcal{B}(\mathbf{Z}) = \sum_{k=1}^{g} \pi_k ||\overline{x}_{\mathbf{z}_k} - \overline{x}||^2,$$

where π_k is the weight of the kth cluster and $\overline{x}$ is the vector center of all data. This equivalence is due to the decomposition of the total variance T of data

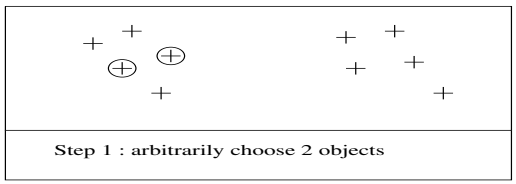
$$\mathcal{I} = \sum_{i=1}^{n} \pi_i ||\mathbf{x}_i - \overline{x}||^2 = \mathcal{W}(\mathbf{Z}) + \mathcal{B}(\mathbf{Z})$$

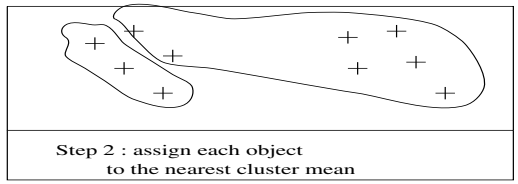
Description of k-means

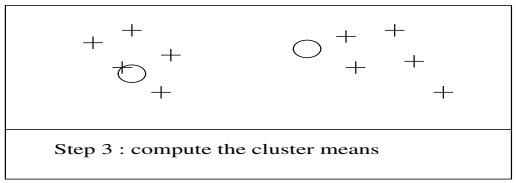
- Let $\mathbf{X}$ be an $n \times d$ continuous data matrix.
- Criterion to optimize

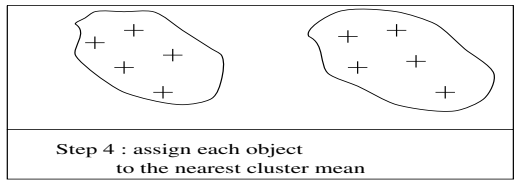
$$\mathcal{C}(\mathbf{Z}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} D(\mathbf{x}_i, \boldsymbol{\mu}_k) \text{ where } D(\mathbf{x}_i, \boldsymbol{\mu}_k) = \sum_{j=1}^{d} (x_{ij} - \mu_{kj})^2$$

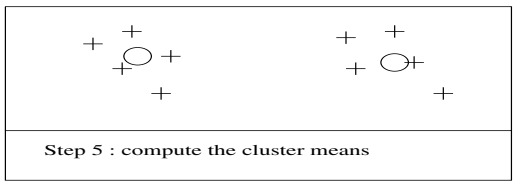
Process of k-means into g=2 clusters

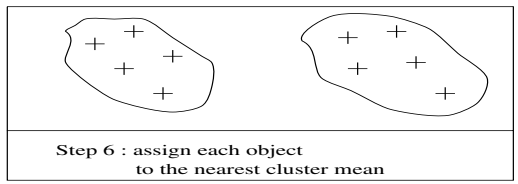












Strengths and weaknesses of k-means

+ Simple and flexible; it can be adapted to the nature of data by modifying D
+ Scalable
- Depends on the initialisation
- Tends to give spherical balanced; it depends on the shapes, proportions and volume of clusters
- It requires the number of clusters but this can be solved with different ways.
- It does not give a good job in high dimensionality and sparsity contexts

- Most of non-hierachical clustering approaches rely on the k-means idea

A partition is relevant means that we are/experts able to interpret it

Dataset

num	Z1	Z2	Z3	Z4
1	22	35	24	19
2	24	31	21	22
3	27	36	25	15
4	27	36	24	23
5	21	33	23	18
6	26	35	23	32
7	27	37	26	15
8	22	30	19	20
9	25	33	22	22
10	30	41	28	17
11	24	39	27	21
12	29	39	27	17
13	29	40	27	17
14	28	36	23	24
15	22	36	24	20
16	23	30	20	20
17	28	38	26	16
18	25	34	23	14
19	26	35	24	15
20	23	37	25	20
21	31	42	29	18
22	26	34	22	21
23	24	38	26	21

Clustering by k-means

k-means	1	2	3	4
1	6	2	0	0
2	0	7	0	0
3	0	0	5	0
4	0	0	2	1

Table: Confusion matrix. Accuracy: $1 - \frac{4}{23} = 82\%$

Normalized Dataset

num	Z1	Z2	Z3	Z4
1	0.2200000	0.3500000	0.2400000	0.1900000
2	0.2448980	0.3163265	0.2142857	0.2244898
3	0.2621359	0.3495146	0.2427184	0.1456311
4	0.2454545	0.3272727	0.2181818	0.2090909
5	0.2210526	0.3473684	0.2421053	0.1894737
6	0.2241379	0.3017241	0.1982759	0.2758621
7	0.2571429	0.3523810	0.2476190	0.1428571
8	0.2417582	0.3296703	0.2087912	0.2197802
9	0.2450980	0.3235294	0.2156863	0.2156863
10	0.2586207	0.3534483	0.2413793	0.1465517
11	0.2162162	0.3513514	0.2432432	0.1891892
12	0.2589286	0.3482143	0.2410714	0.1517857
13	0.2566372	0.3539823	0.2389381	0.1504425
14	0.2522523	0.3243243	0.2072072	0.2162162
15	0.2156863	0.3529412	0.2352941	0.1960784
16	0.2473118	0.3225806	0.2150538	0.2150538
17	0.2592593	0.3518519	0.2407407	0.1481481
18	0.2604167	0.3541667	0.2395833	0.1458333
19	0.2600000	0.3500000	0.2400000	0.1500000
20	0.2190476	0.3523810	0.2380952	0.1904762
21	0.2583333	0.3500000	0.2416667	0.1500000
22	0.2524272	0.3300971	0.2135922	0.2038835
23	0.2201835	0.3486239	0.2385321	0.1926606

Clustering by k-means

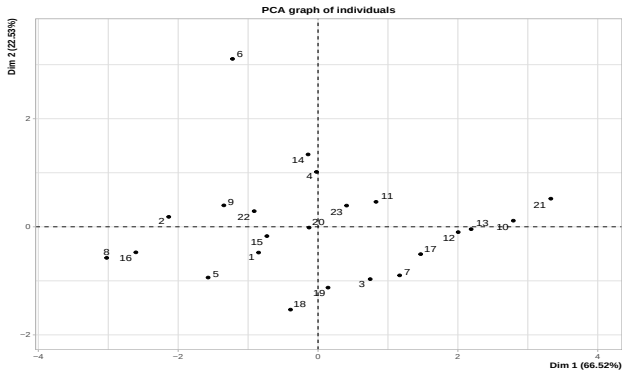
k-means	1	2	3	4
1	6	0	0	0
2	0	7	0	0
3	0	0	9	0
4	0	0	0	1

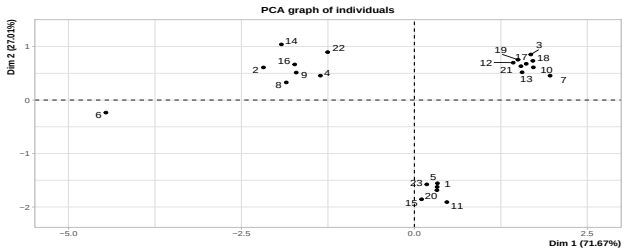
Table: Confusion matrix. Accuracy: $1 - 0/23 = 100\%$

PCA

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

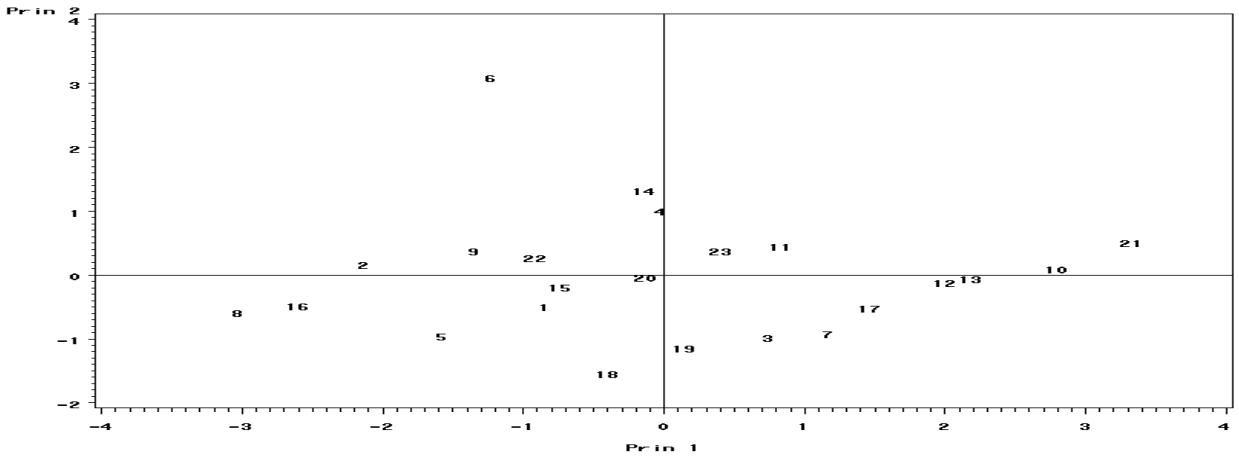
PCA



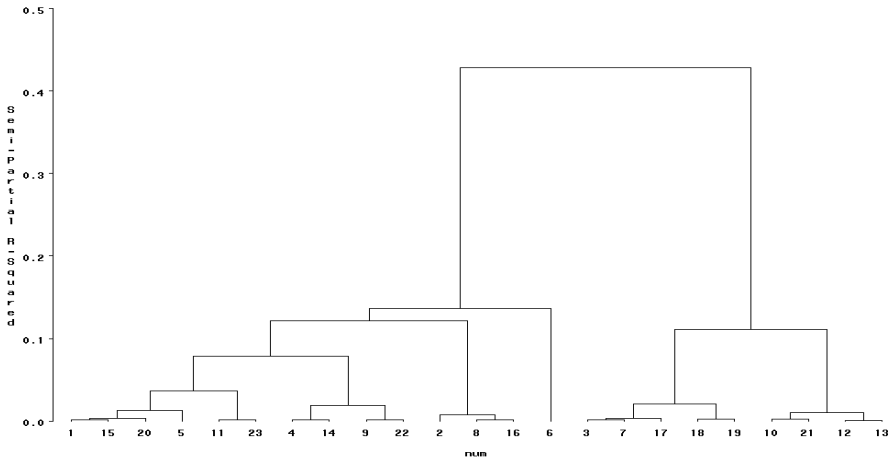


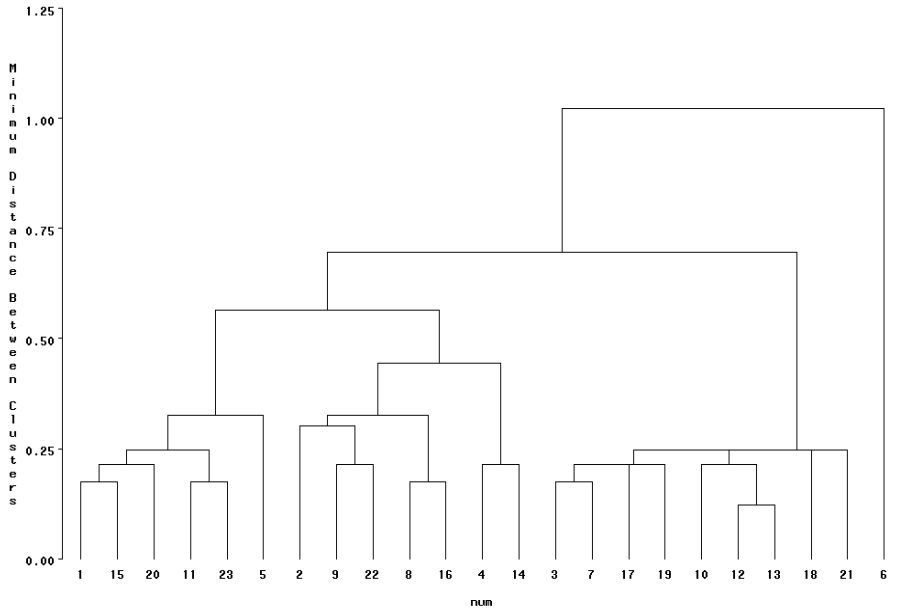
- Cluster1={8,16,2,9,22}
- Cluster2={4,14,6}
- Cluster3={5,1,15,20,23,11,18,19}
- Cluster4={3,7,17,12,13,10,21}

- Cluster1={8,16,2,9,22,4,14}
- Cluster2={6}
- Cluster3={5,1,15,20,23,11}
- Cluster4={18,19,3,7,17,12,13,10,21}



Dendrograms with ward and single criterion





The χ^2 distance

- Expression of this distance on the set of objects (rows):

$$d_{\chi^2}(i,i') = \sum_{j=1}^{p} \frac{1}{x_{.j}} \left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}}\right)^2$$

where $x_{i.} = \sum_{j=1}^{p} x_{ij}$, $x_{.j} = \sum_{i=1}^{n} x_{ij}$.

or,

$$d_{\chi^2}(i,i') = \sum_{j=1}^{p} \left(\frac{x_{ij}}{\sqrt{x_{.j}}x_{i.}} - \frac{x_{i'j}}{\sqrt{x_{.j}}x_{i'.}}\right)^2$$

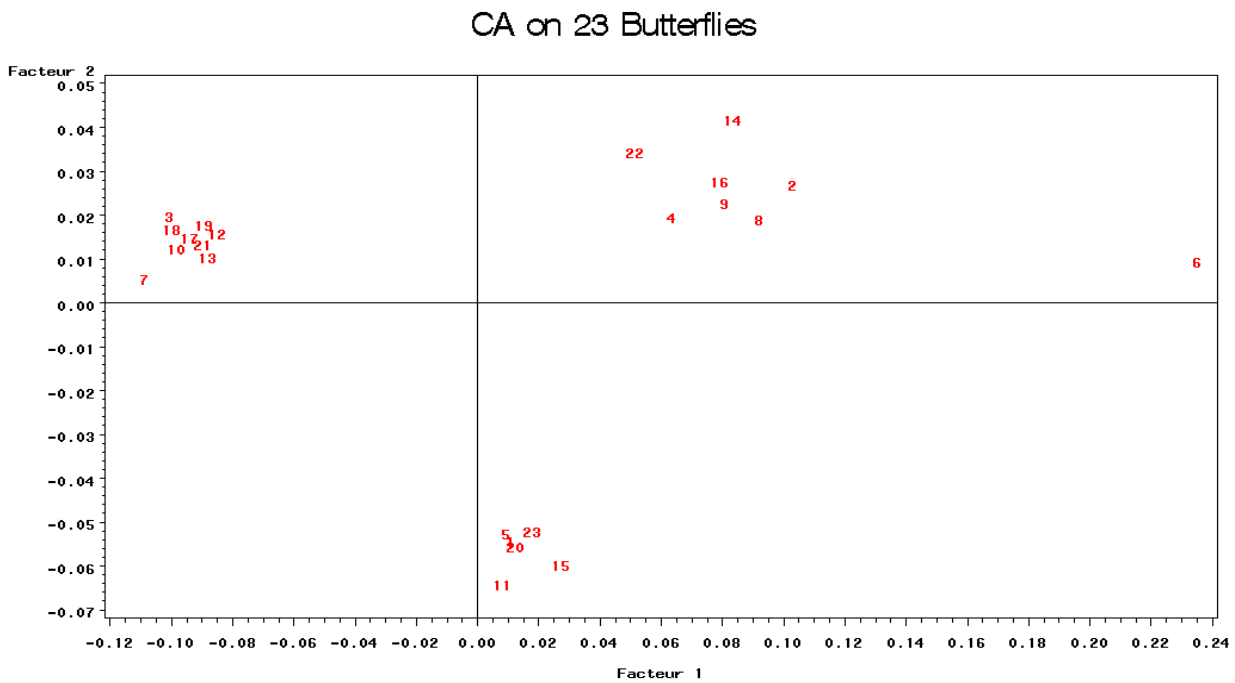
- Expression of this distance on the set of columns:

$$d_{\chi^2}(j,j') = \sum_{i=1}^{n} \left(\frac{x_{ij}}{\sqrt{x_{i.}}x_{.j}} - \frac{x_{ij'}}{\sqrt{x_{i.}}x_{.j'}}\right)^2$$

- When can we use this distance ?
  - Data is a contingency table
  - Data can be viewed as a contingency table and we aim to work on the row/columns profiles instead of the original data

Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns.

Correspondence analysis

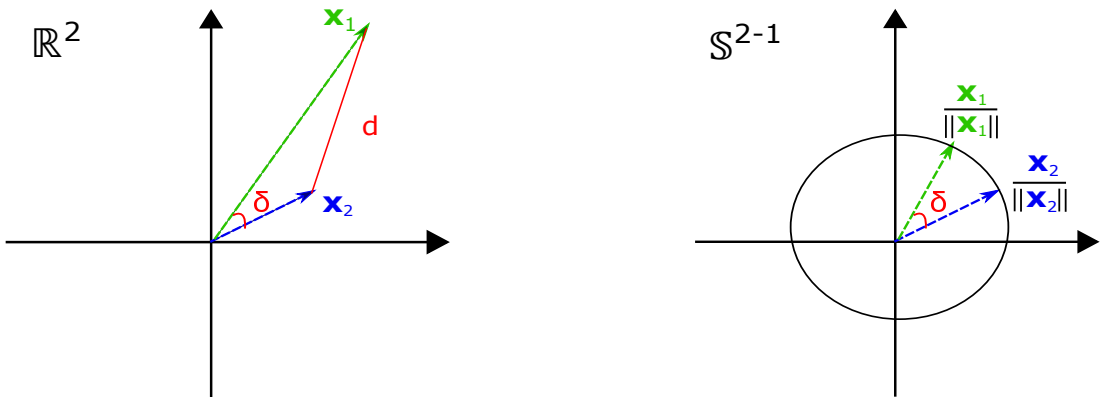


Family of the k-means algorithm

Types	Algorithms	Criteria	dissimilarity/similarity measures
Continuous	k-means	$\sum_{i,k} z_{ik} D(\mathbf{x}_i, \boldsymbol{\mu}_k)$ $\mathbf{x}_i, \boldsymbol{\mu}_k \in \mathbb{R}^d$	$D(\mathbf{x}_i, \boldsymbol{\mu}_k) = \sum_j (x_{ij} - \mu_{kj})^2$
Contingency	k-means-χ^2	$\sum_{i,k} z_{ik} D_{\chi^2}(\mathbf{x}_i, \boldsymbol{\mu}_k)$ $\mathbf{x}_i = (\frac{x_{i1}}{x_{i.}}, \ldots, \frac{x_{id}}{x_{i.}})^\top$ $\mathbf{x}_i, \boldsymbol{\mu}_k \in [0,1]^d$	$D_{\chi^2}(\mathbf{x}_i, \boldsymbol{\mu}_k) = \sum_j \frac{1}{x_{.j}} (\frac{x_{ij}}{x_{i.}} - \mu_{kj})^2$
Categorical (com. disjunc.)	k-means-χ^2	$\sum_{i,k} z_{ik} D_{\chi^2}(\mathbf{x}_i, \boldsymbol{\mu}_k)$ $\mathbf{x}_i = (\frac{x_{i1}}{x_{i.}}, \ldots, \frac{x_{im}}{x_{i.}})^\top$	$D_{\chi^2}(\mathbf{x}_i, \boldsymbol{\mu}_k) = \sum_j \sum_c \frac{1}{x_{.j}} (\frac{x_{ij}}{x_{i.}} - \mu_{kj})^2$
Binary	k-modes	$\sum_{i,k} z_{ik} D(\mathbf{x}_i, \mathbf{a}_k)$ $\mathbf{x}_i, \mathbf{a}_k \in \{0,1\}^d$	$D(\mathbf{x}_i, \mathbf{a}_k) = \sum_j \lvert x_{ij} - a_{kj} \rvert$
Categorical	k-modes	$\sum_{i,k} z_{ik} D(\mathbf{x}_i, \boldsymbol{\lambda}_k)$ $\mathbf{x}_i, \boldsymbol{\lambda}_k \in \{1, \ldots, m^j\}^d$	$D(\mathbf{x}_i, \boldsymbol{\lambda}_k) = \sum_j \delta(x_{ij}, \lambda_{kj})$ $\delta(x_{ij}, \lambda_{kj}) = 0$ if $x_{ij} = \lambda_{kj}$ $\delta(x_{ij}, \lambda_{kj}) = 1$ if $x_{ij} \neq \lambda_{kj}$
Directional	Sk-means	$\sum_{i,k} z_{ik} \cos(\mathbf{x}_i, \boldsymbol{\mu}_k)$ $\mathbf{x}_i, \boldsymbol{\mu}_k \in [0,1]^d$	$\cos(\mathbf{x}_i, \boldsymbol{\mu}_k)$
Directional	Axial k-means	$\sum_{i,k} z_{ik} D(\mathbf{x}_i, \boldsymbol{\mu}_k)$ $\mathbf{x}_i \in [0,1]^d$	Hellinger distance
Mixed data	k-means	$\sum_{i,k} z_{ik} D(\mathbf{x}_i, \boldsymbol{\mu}_k)$	Gower distance

- Several other extensions/connections such as with kernel k-means, symmetric-NMF

- The high dimensionality and sparsity are the characteristics the data sets arising in some areas, such as Recommender systems and text mining
- Such data sets consist of more than 1000 features and 95% of zero entries
- The data sets from these domains are also directional in nature. Only the direction of a data vector is important, not its magnitude



Remarks

- The spherical k-means[1] is tailored for directional data distributed on the surface of a unit-hypersphere
- Text document clustering, microarray-data, and item recommendation are the popular domains where this algorithm is effective.

[1] Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42(1), 143-175.

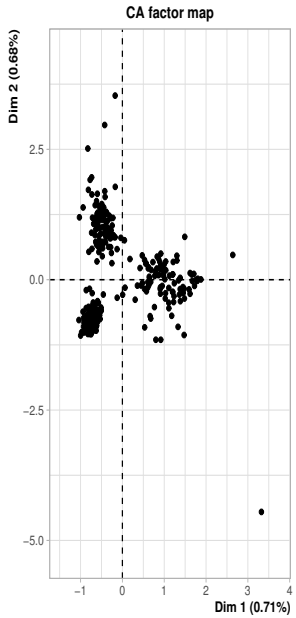
What is the performance of kmeans on this type of data ?

Clustering of classic300 (300×5577) (the true clusters are known)

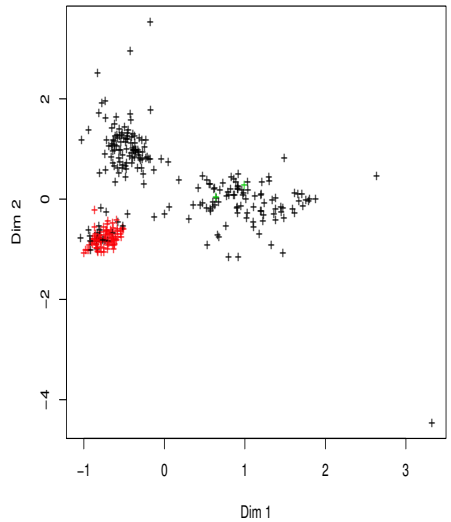
- Each cell x_{ij} denotes the occurrence of word j in document i
- The true clusters are known and we aim to evaluate the performances of k-means and Sk-means to detect the 3 clusters

```
setwd("/Users/nadif/Desktop/MLDS/docdata")
Importation des donnés
library(R.matlab)
classic300 <- readMat("classic300.mat")
classic <- classic300$dtm)
X=as.matrix(classic)
# La derniére colonne correspond á une manuelle classification. Cette information est stockée dans Y
Z=as.matrix(classic300$classid)
# Correspondence Analysis
library(FactoMineR)
res.ca=CA(X)
plot.CA(res.ca,choix="CA",col.row="black",invisible ="col",shadowtext = 0,label="none")
# Application of kmeans on the original data and visualisation of clusters
z.kmeans <- kmeans(X, 3, nstart = 100)
plot(res.ca$row$coord,col=z.kmeans$cluster,pch="+")
table(z.kmeans$cluster,Z)
#Application de skmeans on the original data and visualisation of clusters
library(skmeans)
zs.skmeans <- skmeans(X, 3)
plot(res.ca$row$coord,col=zs.skmeans$cluster,pch="+")
table(zs.skmeans$cluster,Z)
```

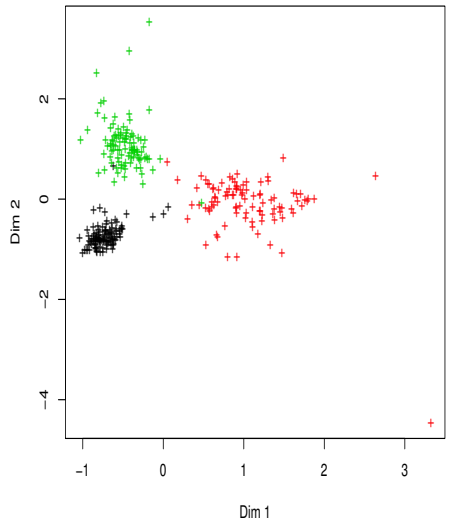
Original data



k-means



Spherical *k*-means



k-means	1	2	3
1	2	0	0
2	2	25	0
3	96	75	100

Sk-means	1	2	3
1	100	3	1
2	0	97	2
3	0	0	97

Table: Confusion matrices

- Accuracy (3 balanced clusters): k-means:$1 - \frac{2+96+75}{300} = 42\%$,
- Sk-means:$1 - \frac{3+1+2}{300} = 98\%$

Challenges

- Popular clustering assumptions based on Euclidean distance are inadequate for certain type of data
- The high dimensionality and sparsity characterising the data sets arising in many fields require appropriate similarity or dissimilarity measures

Challenges and issues

- Choice of the clustering method
- Choice of the objective function which can based on proximity measure or derived from a model

Goals of this talk

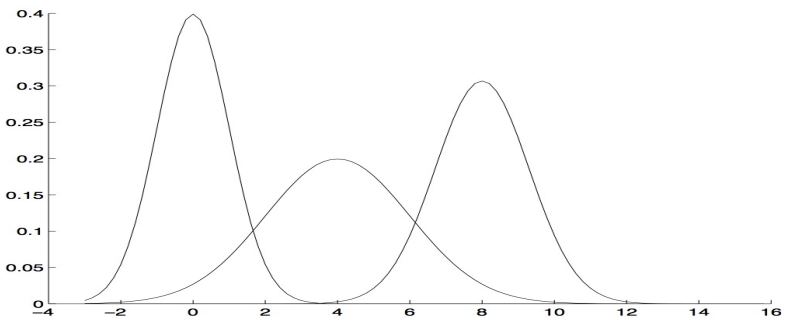
- Curse of high dimensionality and the sparsity
- $n << d$
- Shapes of clusters
- Consensus

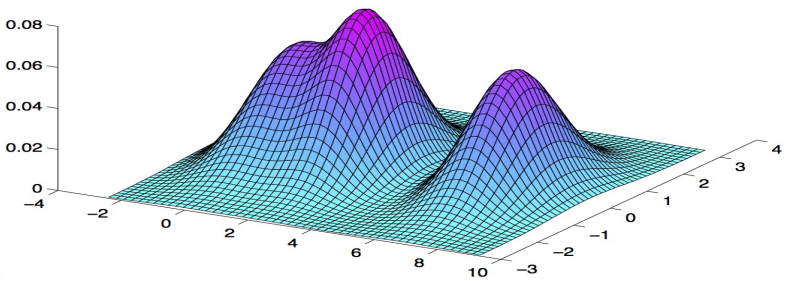
Outline

Finite Mixture Model

- In model-based clustering it is assumed that the data are generated by a mixture of underlying probability distributions, where each component k of the mixture represents a cluster. Thus, the data matrix is assumed to be an i.i.d sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^\top \in \mathbb{R}^d$ from a probability distribution with density

$$f(\boldsymbol{x}_i; \boldsymbol{\Theta}) = \sum_{k=1}^{g} \pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k),$$





- $\varphi(.\,; \boldsymbol{\alpha}_k)$ is the density of an observation $\boldsymbol{x}_i$ from the k-th component
- $\boldsymbol{\alpha}_k$'s are the corresponding class parameters.
- The parameter π_k corresponds to the probability to choose the k-th component
- g, which is assumed to be known, is the number of components in the mixture

ML and CML approaches

- The problem of clustering can be studied in the mixture model using two different approaches: the maximum likelihood approach (ML) and the classification likelihood approach (CML)

1. The ML approach (Day, 1969): It estimates the parameters of the mixture, and the partition on the objects is derived from these parameters using the maximum a posteriori principle (MAP). The maximum likelihood estimation of the parameters results in an optimization of the log-likelihood of the observed sample

$$L_M(\boldsymbol{\Theta}) = L(\boldsymbol{\Theta}; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{g} \pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right)$$

2. The CML approach (Symons, 1981): It estimates the parameters of the mixture and the partition *simultaneously* by optimizing the classification log-likelihood

$$L_C(\mathbf{z}; \boldsymbol{\Theta}) = L(\boldsymbol{\Theta}; \mathbf{X}, \mathbf{z}) = \log f(\mathbf{X}, \mathbf{z}; \boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log \left(\pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right)$$

or

$$L_C(\mathbf{z}; \boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log (\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \underline{\log \left(\varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right)}$$

Maximisation of the Likelihood by the EM algorithm[2]

$Q(\Theta|\Theta')$

In the mixture context

$$Q(\Theta|\Theta') = \mathbb{E}(L_C(\mathbf{z};\Theta|\mathbf{X},\Theta')) = \sum_{i,k} \mathbb{E}(z_{ik}|\mathbf{X},\Theta') \log(\pi_k f(\boldsymbol{x}_i;\alpha_k))$$

where $\mathbb{E}(z_{ik}|\mathbf{X},\Theta') = p(z_{ik} = 1|\mathbf{X},\Theta')$

[2] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), 1-22.

The steps of EM

- The EM algorithm involves constructing, from an initial $\boldsymbol{\Theta}^{(0)}$, the sequence $\boldsymbol{\Theta}^{(c)}$ satisfying

$$\boldsymbol{\Theta}^{(c+1)} = \text{argmax}\, Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(c)})$$

and this sequence causes the criterion $L_M(\boldsymbol{\Theta})$ to grow. The EM algorithm takes the following form

  - Initialize by selecting an initial solution $\boldsymbol{\Theta}^{(0)}$
  - Repeat the two steps until convergence

    1. E-step: compute $Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(c)})$. Note that in the mixture case this step reduces to the computation of the conditional probabilities $s_{ik}^{(c)}$
    2. M-step: compute $\boldsymbol{\Theta}^{(c+1)}$ maximizing $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(c)})$. This leads to $\pi_k^{(c+1)} = \frac{1}{n}\sum_i s_{ik}^{(c+1)}$ and the exact formula for the $\boldsymbol{\alpha}_k^{(c+1)}$ will depend on the involved parametric family of distribution probabilities

Properties of EM

- Under certain conditions, it has been established that EM always converges to a local likelihood maximum
- Simple to implement and it has good behavior in clustering and estimation contexts

Hathaway interpretation of EM: classical mixture model context

- EM = alternated maximization of the fuzzy clustering criterion

$$F_C(\tilde{\mathbf{Z}}, \boldsymbol{\Theta}) = L_C(\tilde{\mathbf{Z}}; \boldsymbol{\Theta}) + H(\tilde{\mathbf{Z}})$$

  - $\tilde{\mathbf{Z}} = (\tilde{z}_{ik})$: fuzzy partition
  - $L_C(\tilde{\mathbf{Z}}, \boldsymbol{\Theta}) = \sum_{i,k} \tilde{z}_{ik} \log(\pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k))$: fuzzy classification log-likelihood
  - $H(\tilde{\mathbf{Z}}) = -\sum_{i,k} \tilde{z}_{ik} \log \tilde{z}_{ik}$: entropy function

Algorithm

- Maximizing $F_C(\tilde{\mathbf{Z}}, \boldsymbol{\Theta})$ w.r. to $\tilde{\mathbf{Z}}$ yields the E-step
- Maximizing $F_C(\tilde{\mathbf{Z}}, \boldsymbol{\Theta})$ w.r. to $\boldsymbol{\Theta}$ yields the M-step

Fuzzy clustering to hard clustering

	a	b	c	$\tilde{z}_{i1}$	$\tilde{z}_{i2}$	$\tilde{z}_{i3}$	z_{i1}	z_{i2}	z_{i3}	z
i1	x	x	x	0.7	0.1	0.2	1	0	0	1
i2	x	x	x	0.1	0.6	0.3	0	1	0	2
i3	x	x	x	0.1	0.1	0.8	0	0	1	3
i4	x	x	x	0.6	0.2	0.2	1	0	0	1
i5	x	x	x	0.2	0.6	0.2	0	1	0	2
i6	x	x	x	0.1	0.7	0.2	0	1	0	2
i7	x	x	x	0.2	0.1	0.7	0	0	1	3
i8	x	x	x	0.8	0.1	0.1	1	0	0	1
i9	x	x	x	0.2	0.2	0.6	0	0	1	3
i10	x	x	x	0.1	0.8	0.1	0	1	0	2

The Gaussian model[ab]

[a] Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 803-821.

[b] Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern recognition, 28(5), 781-793.

- The density can be written as: $f(\boldsymbol{x}_i; \boldsymbol{\Theta}) = \sum_k \pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k)$ where

$$\varphi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_k|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x}_i - \mu_k)\}$$

- Spectral decomposition of the variance matrix

$$\Sigma_k = \lambda_k D_k A_k D_k^\top$$

  - $\lambda_k = |\Sigma_k|^{1/p}$ positive real represents the volume of the kth component
  - $A_k = Diag(a_{k1}, \ldots, a_{kd})$ whose elements are proportional to the eigenvalues of Σ_k. It defines the shape of the kth cluster
  - D_k formed by the eigenvectors. It defines the direction of the kth cluster

- Remark: number of parameters to estimate: $(g-1) + g \times d + g \times \frac{d(d+1)}{2}$

Different Gaussian models

- The Gaussian mixture depends on: proportions, centers, volumes, shapes and Directions then different models can be proposed
- In the following models proportions can be assumed equal or not

1. Spherical models: $A_k = I$ then $\Sigma_k = \lambda_k I$. Two models $[\lambda I]$ and $[\lambda_k I]$
2. Diagonal models: Four models $[\lambda A]$, $[\lambda_k A]$, $[\lambda A_k]$ and $[\lambda_k A_k]$
3. General models: the eight models assuming equal or not volumes, shapes and directions $[\lambda DAD^\top]$, $[\lambda_k DAD^\top]$, $[\lambda DA_k D^\top]$, $[\lambda_k DA_k D^\top]$, $[\lambda D_k AD_k^\top]$,$[\lambda_k D_k AD_k^\top]$, $[\lambda D_k A_k D_k^\top]$ and $[\lambda_k D_k A_k D_k^\top]$

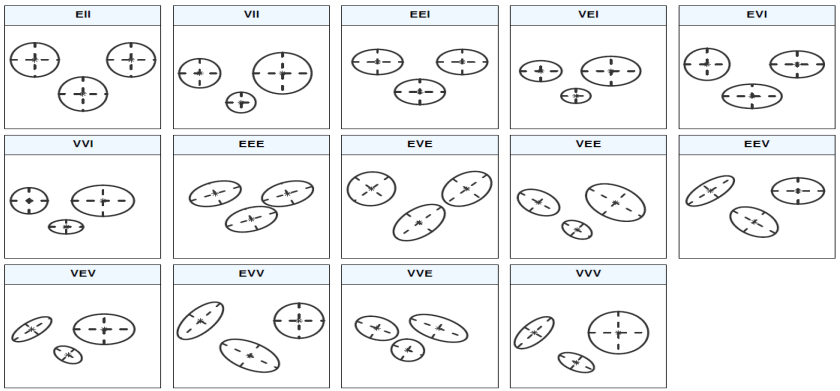
- Finally we have 28 models
- See for instance **mclust**[a] and **Rmixmod**[b]

[a] "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models" by Luca Scrucca, Michael Fop, T. Brendan Murphy and Adrian E. Raftery, 2016

[b] Rmixmod: "The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library" by Lebret, R., Iovleff, S., Langrognet, F., Biernacki, Ch, Celeux, G., Govaert, G. Journal of Statistical Software, 2015

Library `mclust`

- Spherical models: By fixing $A_k = I$, we place ourselves in the case where the classes are of spherical shapes, that is to say that the variances of all the variables are equal inside of the same class.
- Diagonal models: Considering that the matrices D_k are diagonal, we force the classes to be aligned on the axes. It is in fact the hypothesis of conditional independence in which the variables are independent of each other within the same class.
- General models: By fixing equality constraints on the A_k, the D_k or the λ_k, we can generate 8 different models.



Classification EM (CEM)

- In clustering step, each $\boldsymbol{x}_i$ is assigned to the cluster maximizing $s_{ik} \propto \pi_k \varphi(\boldsymbol{x}_i; \mu_k, \Sigma_k)$ or equivalently the cluster that minimizes

$$-\log(\pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)) = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log(\pi_k) + cste$$

- From density to Distance (or dissimilarity), $\boldsymbol{x}_i$ is assigned to the cluster according the following dissimilarity

$$D^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i; \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log(\pi_k)$$

where $D^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i; \mu_k) = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)$ is the Mahanalobis distance

- Note that when the proportions are supposed equal and the variances identical, the assignation is based only on

$$D^2_{\Sigma_k^{-1}}(\boldsymbol{x}_i; \boldsymbol{\mu}_k)$$

- When the proportions are supposed equal and for the spherical model $[\lambda I]$ ($\Sigma_k = I$), one uses the usual euclidean distance

$$D^2(\boldsymbol{x}_i; \boldsymbol{\mu}_k)$$

The von Mises-Fisher distribution (vMF)

Let $\mathbf{x}_i \in \mathbb{S}^{d-1}$ be a data point following a vMF distribution, then its pdf is

$$f(\mathbf{x}_i|\boldsymbol{\mu}, \kappa) = c_d(\kappa) \exp^{\kappa \boldsymbol{\mu}^\top \mathbf{x}_i}, \quad (1)$$

$\boldsymbol{\mu}$: centroid parameter, κ: concentration parameter, such that $\|\boldsymbol{\mu}\| = 1$ and $\kappa \geq 0$. $c_d(\kappa) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)}$ $I_r(\kappa)$: the modified Bessel function of the first kind and order r.

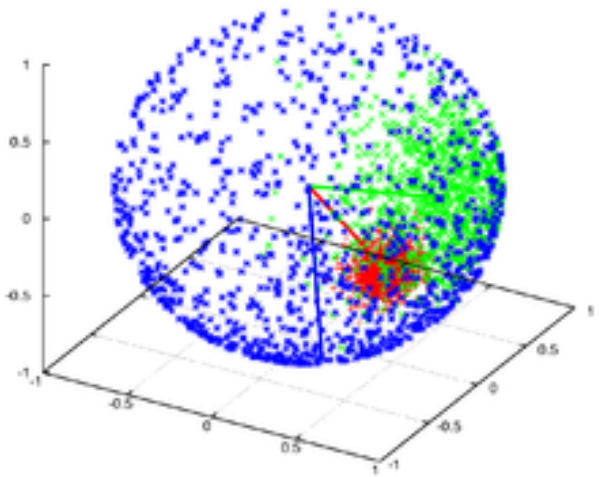


Figure: Impact of κ. blue: $\kappa = 1$, green: $\kappa = 10$, red: $\kappa = 100$

The Mixture of vMF distributions (movMFs)

The data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are supposed to be i.i.d and generated from a mixture of g vMF distributions, with pdf:

$$f(\mathbf{x}_i|\boldsymbol{\Theta}) = \sum_{k=1}^{g} \pi_k \varphi(\mathbf{x}_i|\boldsymbol{\mu}_k, \kappa_k), \quad (2)$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \pi_1, \ldots, \pi_g, \kappa_1, \ldots, \kappa_g\}$

Algorithms

Log-likelihood

$$L(\Theta; \mathbf{X}) = \sum_i \log \left(\sum_k \pi_k \varphi(\mathbf{x}_i | \boldsymbol{\mu}_k, \kappa_k) \right),$$

Complete data log-likelihood

$$\begin{aligned} L_C(\Theta; \mathbf{X}, \mathbf{Z}) &= \sum_{i,k} z_{ik} \log \pi_k + \sum_{i,k} z_{ik} \log c_d(\kappa_k) + \sum_{i,k} z_{ik} \kappa_k \boldsymbol{\mu}_k^\top \mathbf{x}_i \\ &= \sum_{i,k} z_{ik} \log \pi_k + \sum_{i,k} z_{ik} \log c_d(\pi_k) + \sum_{i,k} z_{ik} \kappa_k \cos(\boldsymbol{\mu}_k, \mathbf{x}_i) \end{aligned}$$

CEM

- E-step: finds the conditional expectation $\tilde{z}_{ik} = \mathbb{E}(z_{ik} = 1 | \mathbf{x}_i, \Theta^{(t)})$
- C-step $\tilde{z}_{ik} \Rightarrow z_{ik}$
- M-step: finds the new parameters $\Theta^{(t+1)}$ maximizing $Q(\Theta, \Theta^{(t)}) = \mathbb{E}\left(L(\Theta; \mathbf{x}, \mathbf{z}) | \mathbf{X}, \Theta^{(t)} \right)$ s.t. $\sum_k \pi_k = 1$, $\|\boldsymbol{\mu}_k\| = 1$ and $\kappa_k > 0$

Hypotheses: $\forall k, \pi_k = 1/g$ and $\kappa_k = \kappa$ the maximization of $L_C(\Theta; \mathbf{X}, \mathbf{Z})$ and $\sum_{i,k} z_{ik} \cos(\mathbf{x}_i, \boldsymbol{\mu}_k)$ are equivalent

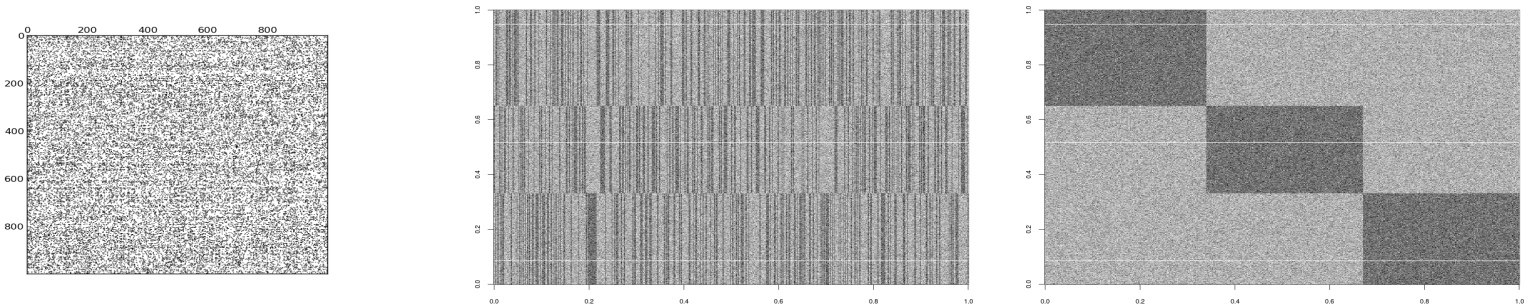
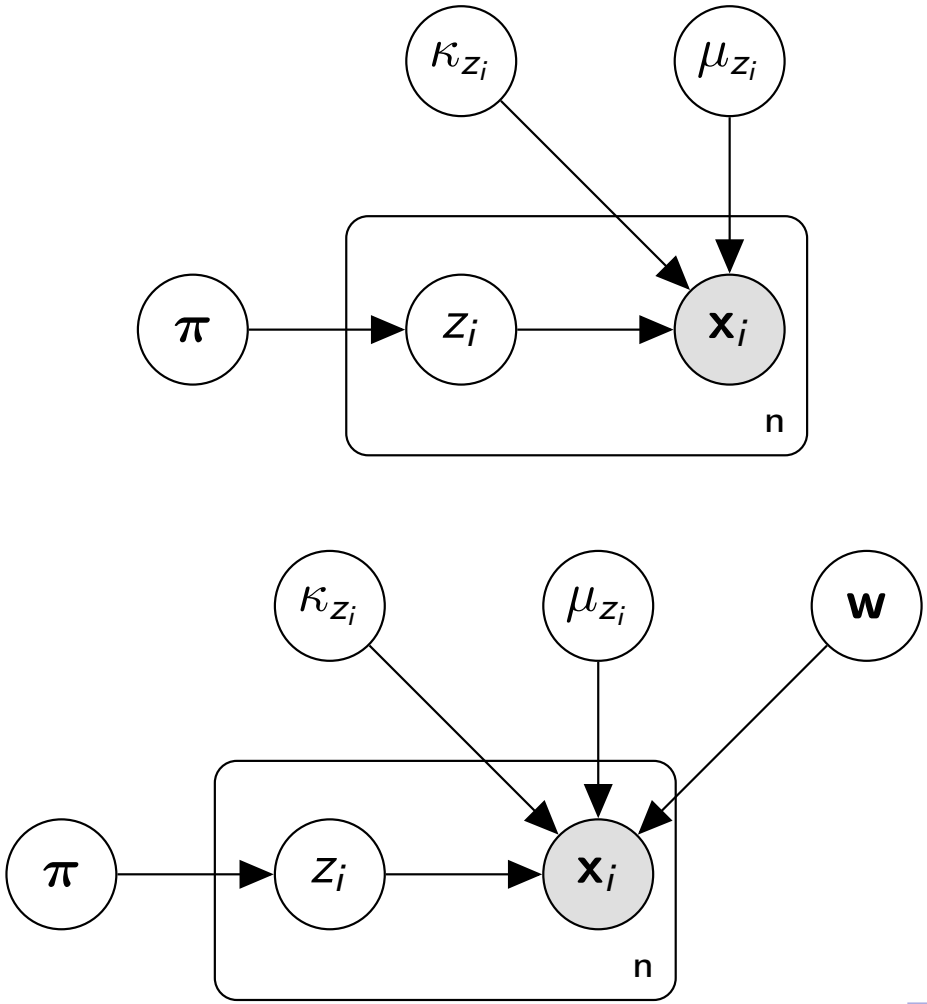


Figure: dbmovMFs-based co-clustering: (left) original data, (middle) data reorganized according to $\mathbf{z}$, (right) data reorganized data according to $(\mathbf{z}, \mathbf{w})$

- Produces directly interpretable clusters
- Very parsimonious: we need to estimate only few parameters
- It focuses only on the most useful co-clusters and ignores noisy co-clusters

Co-clustering[a]

[a]Salah, A., & Nadif, M. (2017). Model-based von mises-fisher co-clustering with a conscience. In Proceedings of the 2017 SIAM International Conference on Data Mining (pp. 246-254). Society for Industrial and Applied Mathematics.



A mixture of vMF distribution for co-clustering: dbmovMFs

dbmovMF's pdf

Let $\mathbf{x}_i \in \mathbb{S}^{d-1}$ be generated according to dbmovMFs, then

$$f(\mathbf{x}_i|\Theta) = \sum_k \pi_k \varphi(\mathbf{x}_i|\boldsymbol{\mu}_k, \kappa_k, \mathbf{w}),$$

where $\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \pi_1, \ldots, \pi_g, \kappa_1, \ldots, \kappa_g, \mathbf{w}\}$, such that
$w_j = k$ if column j is in cluster k

Let $g = 3$, $k = 1, 2, 3$, $w_{.k}$ size of the k^{th} column cluster, then $\boldsymbol{\mu}_h$ takes this form:

$$\boldsymbol{\mu}_k = (\underbrace{\mu_{h1}, \ldots, \mu_{h1}}_{(1\times w_{.1})}, \underbrace{\mu_{h2}, \ldots, \mu_{h2}}_{(1\times w_{.2})}, \underbrace{\mu_{h3}, \ldots, \mu_{h3}}_{(1\times w_{.3})})^T, \ \mu_{hk} = 0 \ \forall k \neq h$$

Complete data likelihood

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\pi}, \kappa; \mathbf{Z}) = \prod_i \prod_k \left(\pi_k c_d(\kappa_k) \times \prod_j (\exp^{\kappa_h \mu_{kk} x_{ij}})^{w_{jk}} \right)^{z_{ik}}$$

$z_{ik} = 1$ if $\mathbf{x}_i$ arises from cluster k and $z_{ik} = 0$, otherwise
$w_{jk} = 1$ if column j arises from cluster k and $w_{jk} = 0$, otherwise

A mixture of vMF distribution for co-clustering: dbmovMFs

dbmovMF's pdf

Let $\mathbf{x}_i \in \mathbb{S}^{d-1}$ be generated according to dbmovMFs, then

$$f(\mathbf{x}_i|\Theta) = \sum_k \pi_k \varphi(\mathbf{x}_i|\boldsymbol{\mu}_k, \kappa_k, \mathbf{w}),$$

where $\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \pi_1, \ldots, \pi_g, \kappa_1, \ldots, \kappa_g, \mathbf{w}\}$, such that
$w_j = k$ if column j is in cluster k

Let $g = 3$, $k = 1, 2, 3$, $w_{.k}$ size of the k^{th} column cluster, then $\boldsymbol{\mu}_h$ takes this form:

$$\boldsymbol{\mu}_k = (\underbrace{\mu_{h1}, \ldots, \mu_{h1}}_{(1\times w_{.1})}, \underbrace{\mu_{h2}, \ldots, \mu_{h2}}_{(1\times w_{.2})}, \underbrace{\mu_{h3}, \ldots, \mu_{h3}}_{(1\times w_{.3})})^T, \ \mu_{hk} = 0 \ \forall k \neq h$$

Complete data likelihood

$$\mathcal{L}(\mathbf{w}, \mu, \pi, \kappa; \mathbf{Z}) = \prod_i \prod_k \left(\pi_k c_d(\kappa_k) \times \prod_j (\exp^{\kappa_h \mu_{kk} x_{ij}})^{w_{jk}} \right)^{z_{ik}}$$

$z_{ik} = 1$ if $\mathbf{x}_i$ arises from cluster k and $z_{ik} = 0$, otherwise
$w_{jk} = 1$ if column j arises from cluster k and $w_{jk} = 0$, otherwise

A mixture of vMF distribution for co-clustering: dbmovMFs

dbmovMF's pdf

Let $\mathbf{x}_i \in \mathbb{S}^{d-1}$ be generated according to dbmovMFs, then

$$f(\mathbf{x}_i|\Theta) = \sum_k \pi_k \varphi(\mathbf{x}_i|\boldsymbol{\mu}_k, \kappa_k, \mathbf{w}),$$

where $\Theta = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \pi_1, \ldots, \pi_g, \kappa_1, \ldots, \kappa_g, \mathbf{w}\}$, such that
$w_j = k$ if column j is in cluster k

Let $g = 3$, $k = 1, 2, 3$, $w_{.k}$ size of the k^{th} column cluster, then $\boldsymbol{\mu}_h$ takes this form:

$$\boldsymbol{\mu}_k = (\underbrace{\mu_{h1}, \ldots, \mu_{h1}}_{(1\times w_{.1})}, \underbrace{\mu_{h2}, \ldots, \mu_{h2}}_{(1\times w_{.2})}, \underbrace{\mu_{h3}, \ldots, \mu_{h3}}_{(1\times w_{.3})})^T, \ \mu_{hk} = 0 \ \forall k \neq h$$

Complete data likelihood

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\pi}, \kappa; \mathbf{Z}) = \prod_i \prod_k \left(\pi_k c_d(\kappa_k) \times \prod_j (\exp^{\kappa_h \mu_{kk} x_{ij}})^{w_{jk}} \right)^{z_{ik}}$$

$z_{ik} = 1$ if $\mathbf{x}_i$ arises from cluster k and $z_{ik} = 0$, otherwise
$w_{jk} = 1$ if column j arises from cluster k and $w_{jk} = 0$, otherwise

Models and algorithms[3]

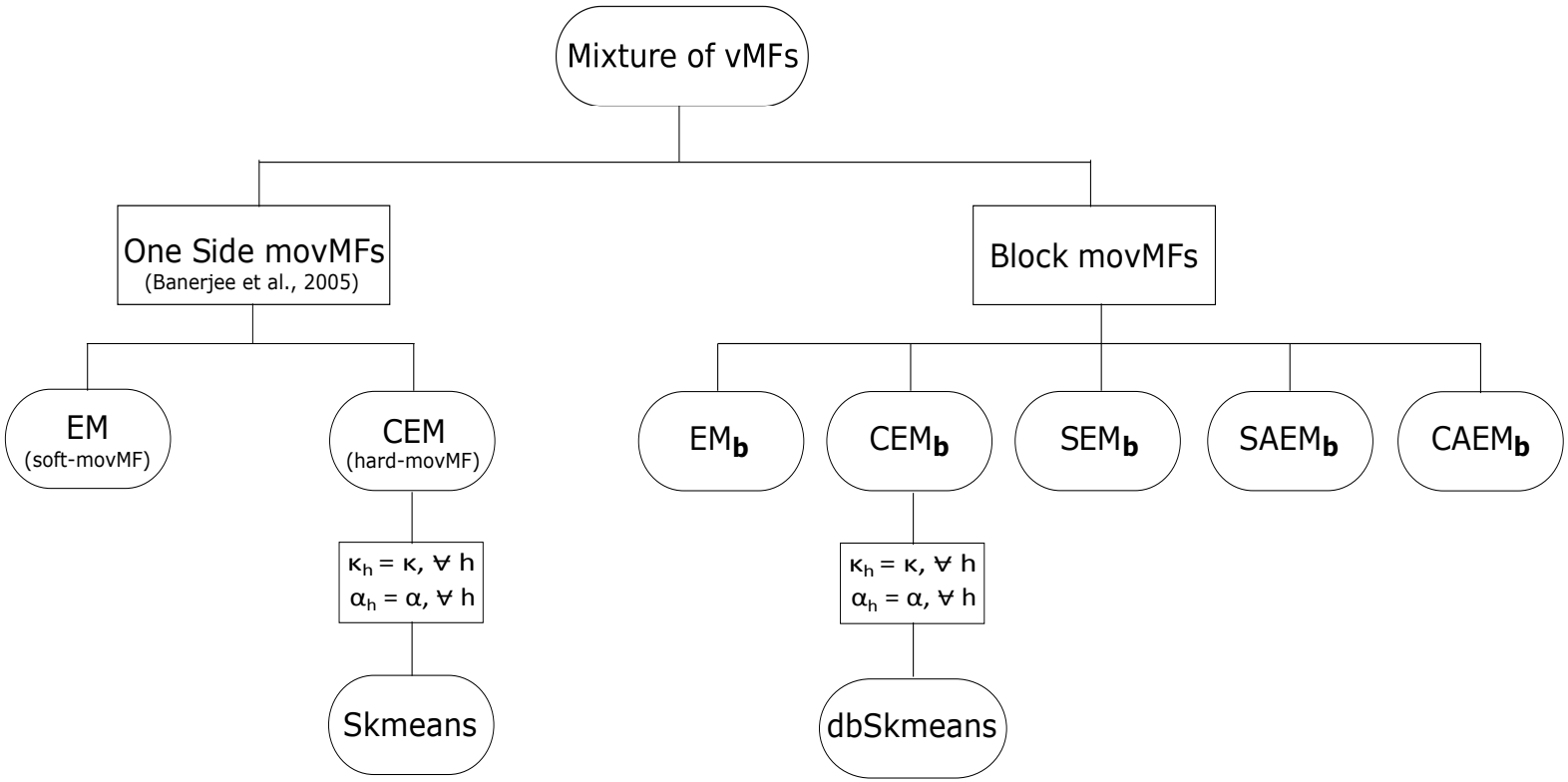


Figure: von Mises-Fisher mixture models-based clustering (left) and co-clustering (right) algorithms.

[3] Salah, A., & Nadif, M. (2019). Directional co-clustering. Advances in Data Analysis and Classification, 13(3), 591-620.

Comparison of some variants

Table: Mean±sd Normalized Mutual Information (NMI and Adjusted Rand Index (ARI) on several real-world *documents* × *terms* matrices (50 runs).

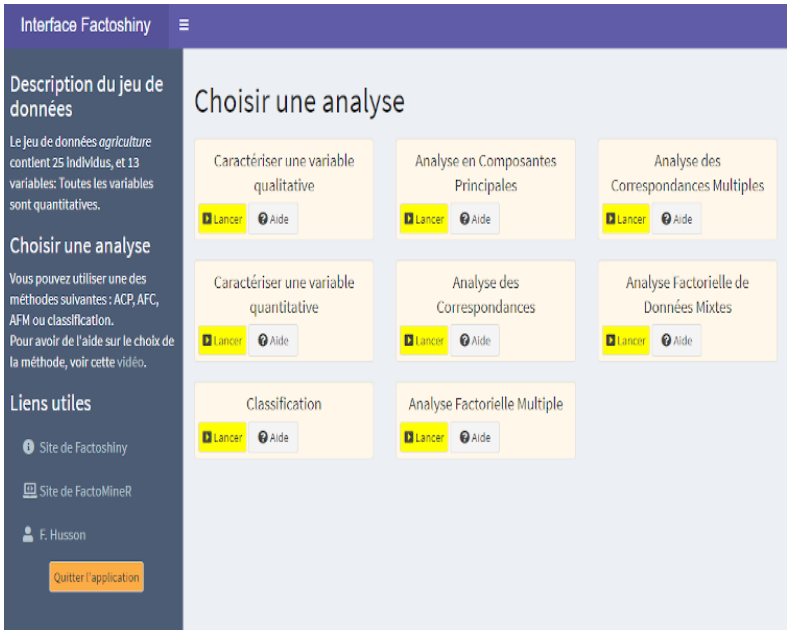
Methods	CSTR		CLASSIC4		WEBACE	
	NMI	ARI	NMI	ARI	NMI	ARI
Skmeans	0.732±0.026	0.772±0.025	0.591±0.020	0.468±0.011	0.613±0.008	0.423±0.026
	(0.759)	(0.807)	(0.595)	(0.476)	(0.620)	(0.394)
CEM	0.734±0.025	0.774±0.024	0.413±0.011	0.199±0.018	0.619±0.011	0.398±0.021
	(0.759)	(0.807)	(0.410)	(0.194)	(0.623)	(0.412)
EM	0.741±0.026	0.777±0.026	0.406±0.013	0.190±0.015	0.614±0.014	0.385±0.034
	(0.768)	(0.808)	(0.403)	(0.184)	(0.623)	(0.397)
DAEM	0.779±0.013	0.813±0.014	0.591±0.002	0.471±0.002	0.620±0.008	0.427±0.022
	(0.783)	(0.817)	(0.592)	(0.472)	(0.628)	(0.468)
CoclusMod	0.701±0.02	0.693±0.04	0.709±0.020	0.675±0.050	0.602±0.020	0.566±0.03
	(0.722)	(0.730)	(0.727)	(0.680)	(0.615)	(0.570)
CEM_b	0.754±0.024	0.804±0.022	0.660±0.003	0.467±0.003	0.623±0.011	0.479±0.038
	(0.789)	(0.830)	(0.665)	(0.473)	(0.637)	(0.519)
EM_b	0.754±0.022	0.803±0.02	0.660±0.002	0.466±0.002	0.624±0.008	0.481±0.02
	(0.792)	(0.837)	(0.668)	(0.473)	(0.639)	(0.523)
SEM_b	0.776±0.022	0.82±0.02	0.691±0.031	0.705±0.05	0.567±0.04	0.582±0.03
	(0.807)	(0.846)	(0.721)	(0.735)	(0.597)	(0.588)
$CAEM_b$	0.794±0.014	0.833±0.013	0.735±0.033	0.751±0.048	0.640±0.007	0.666±0.019
	(0.817)	(0.851)	(0.746)	(0.772)	(0.658)	(0.688)
$SAEM_b$	0.795±0.011	0.830±0.010	0.746±0.023	0.756±0.039	0.644±0.015	0.656±0.021
	(0.821)	(0.851)	(0.773)	(0.798)	(0.661)	(0.689)

Strengths

- New algorithms that improve the performances of SK-means
- With such approach we can use popular criteria to estimate the number of clusters

Outline

- Many strategies: The most popular (PCA followed by k-means)
- Interesting for Exploratory Data Analysis and useful for other machine learning techniques [4][5]



[4] Husson, F., Josse, J., Le, S., Mazet, J., & Husson, M. F. (2016). Package 'factominer'. An R package, 96, 698.

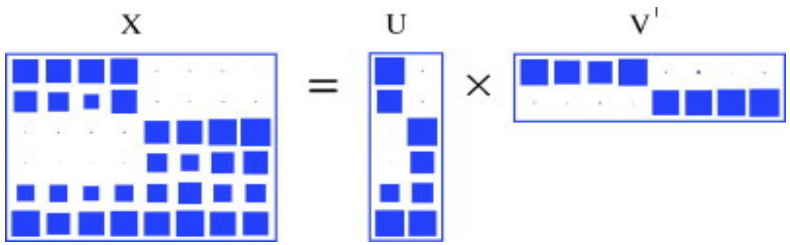
[5] Kassambara, A., & Mundt, F. (2017). Package 'factoextra'. Extract and visualize the results of multivariate data analyses, 76(2).

Nonnegative data

NMF: Nonnegative Matrix Factorization[a]

[a] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

- Problem: $\underset{\mathbf{U},\mathbf{V}\geq 0}{\operatorname{argmin}} ||\mathbf{X} - \mathbf{U}\mathbf{V}^\top||^2$ where factor matrices, $\mathbf{U} \in \mathbb{R}_+^{n\times g}$ and $\mathbf{V} \in \mathbb{R}_+^{d\times g}$
- I-divergence: $\underset{\mathbf{U},\mathbf{V}\geq 0}{\operatorname{argmin}} = \sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}\log\frac{X_{ij}}{(UV^\top)_{ij}} - X_{ij} + (UV^\top)_{ij}$
- The clustering problem is not the main objective of NMF
- Each row of $\mathbf{X}$ is treated as a data point in d-dimensional space
- Each U_{ik} of $\mathbf{U}$ corresponds to the degree to which row i belongs to kth cluster
- Each row of $\mathbf{V}$ is associated with a prototype vector for the kth cluster



NMF and K-means

Expressions of U and V

A typical constrainted optimization problem can be solved using the Lagrange multiplier method: $U_{ik} \leftarrow U_{ik} \frac{(\mathbf{XV})_{ik}}{(\mathbf{UV}^T\mathbf{V})_{ik}}$ and $V_{kj} \leftarrow V_{kj} \frac{(\mathbf{X}^T\mathbf{U})_{kj}}{(\mathbf{VU}^T\mathbf{U})_{kj}}$

NMF towards clustering

1. Perform the NMF on $\mathbf{X}$ to obtain $\mathbf{U}$ and $\mathbf{V}$
2. Normalize $\mathbf{U}$ and $\mathbf{V}$ ($\widetilde{\mathbf{U}} = \mathbf{U}\mathbf{D}_U^{-1}$ and $\widetilde{\mathbf{V}} = \mathbf{V}\mathbf{D}_U$, where $\mathbf{D}_U = Diag(e^T\mathbf{U})$
3. Use matrix $\tilde{\mathbf{U}}$ to determine the cluster label of each document. Examine each row of matrix $\widetilde{\mathbf{U}}$ and assign a document $\mathbf{d}_i$ to cluster k^* if $k^* = \arg\max_k \widetilde{\mathbf{U}}_{ik}$

Orthogonal NMF[a]

[a] Yoo, J., & Choi, S. (2008, November). Orthogonal nonnegative matrix factorization: Multiplicative updates on Stiefel manifolds. In International conference on intelligent data engineering and automated learning (pp. 140-147). Springer, Berlin, Heidelberg.

$\text{argmin}_{\mathbf{U},\mathbf{V}\geq 0} ||\mathbf{X} - \mathbf{UV}^T||^2$ where $\mathbf{U} \in \mathbb{R}_+^{n\times g}, \mathbf{V} \in \mathbb{R}_+^{d\times g}$, $\mathbf{UU}^T = \mathbf{I}_n$

Table: Description of Datasets, # denotes the cardinality

Data	Characteristics					
	#Documents	#Terms	#Clusters	Density$_X$ (%)	balance	Density$_M$ (%)
CSTR	475	1000	4	3.40	0.399	13.80
CLASSIC4	7095	5896	4	0.59	0.323	6.21
RCV1	6387	16921	4	0.25	0.080	3.07
NG5	4905	10167	5	0.92	0.943	22.04
NG20	18846	14390	20	0.59	0.628	25.06

Table: Mean±sd Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) on several real-world *documents* × *terms* matrices (50 runs).

Data	metric	NMF	ONMF	PNMF	GNMF	SNMF
CSTR	NMI	0.65±0.01	0.65±0.05	0.66±0.01	0.57±0.08	**0.75**±0.01
	ARI	0.54±0.01	0.56±0.04	0.56±0.01	0.53±0.11	**0.80**±0.01
CLASSIC4	NMI	0.51±0.09	0.55±0.09	0.59±0.05	0.65±0.04	**0.72**±0.06
	ARI	0.36±0.10	0.39±0.09	0.44±0.01	0.49±0.05	**0.70**±0.09
RCV1	NMI	0.39±0.03	0.49±0.002	0.46±0.00	0.48±0.04	**0.56**±0.01
	ARI	0.29±0.02	0.39±0.00	0.37±0.00	0.39±0.03	**0.57**±0.01
NG5	NMI	0.65±0.05	0.65±0.04	0.65±0.05	0.63±0.07	**0.72**±0.04
	ARI	0.48±0.09	0.48±0.08	0.47±0.09	0.62±0.09	**0.70**±0.06
NG20	NMI	0.43±0.01	0.44±0.02	0.45±0.02	0.52±0.01	**0.53**±0.01
	ARI	0.24±0.01	0.22±0.02	0.24±0.02	0.35±0.05	**0.37**±0.01

- ONMF (Othogonal NMF). PNMF (Projective NMF)[6] GNMF (Graph NMF)[7]. SNMF (Semantic NMF) [8]

Strengths

- Popular methods in computer vision and text-mining. Simple to implement. Many packages are available

[6] Yuan, Z., & Oja, E. (2005). Projective nonnegative matrix factorization for image compression and feature extraction. In Scandinavian Conference on Image Analysis (pp. 333-342).

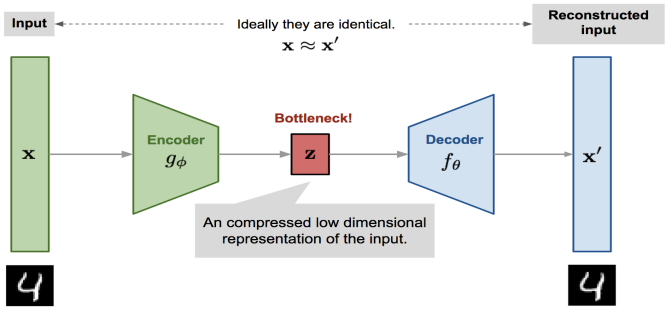
[7] Cai, D., He, X., Han, J., & Huang, T. S. (2010). Graph regularized nonnegative matrix factorization for data representation. IEEE TPMAI, 33(8), 1548-1560.

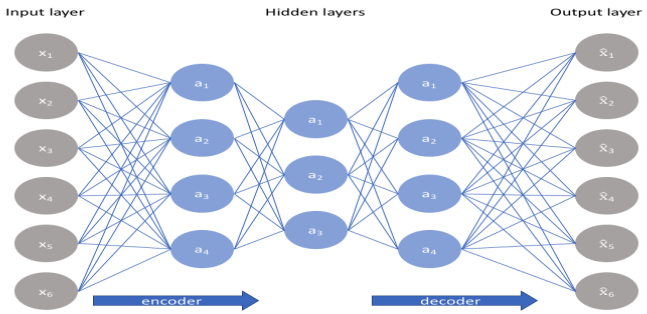
[8] Febrissy, M., Salah, A., Ailem, M., & Nadif, M. (2022). Improving NMF clustering by leveraging contextual relationships among words. Neurocomputing, 495, 105-117.

Deep learning

Autoencoder

- An autoencoder is a type of artificial neural network
- It is an unsupervised learning algorithm that applies backpropagation to adjust its weights, attempting to learn to make its target values (outputs) to be equal to its inputs. In other words, it is trying to learn an approximation to the identity function, so as its output is similar to its input, for all training examples.
- If linear activations are used, or only a single sigmoid hidden layer, then the optimal solution to an autoencoder is strongly related to PCA





$$L_{rec} = \sum_{\mathbf{x}} \delta(\mathbf{x}, g_\Phi of_\theta(\mathbf{x})) \quad \delta \text{ is a dissmilarity function}$$

This loss function is trained by backpropagation.

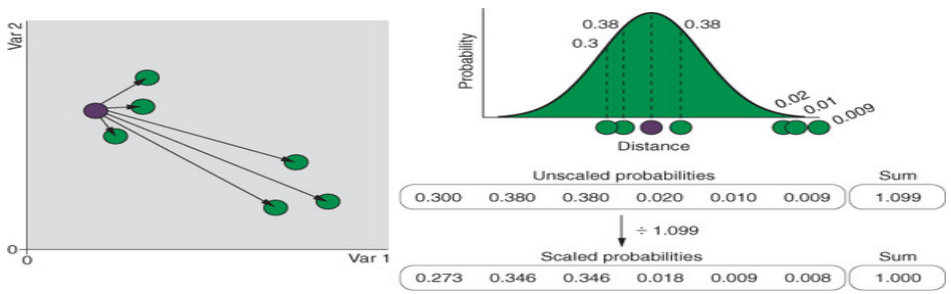
SNE: Stochastic Neighbor Embedding[a]

[a] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. JMLR, 9(11).

Aim : Modeling pairwise similarities

$$p_{j|i} = \frac{exp(-||\mathbf{x}_i - \mathbf{x}_j||^2)/2\sigma_i^2}{\sum_{k \neq i} exp(-||\mathbf{x}_i - \mathbf{x}_k||^2)/2\sigma_i^2} \quad with \quad p_{i|i} = 0$$

where σ_i^2 is the variance of the Gaussian that is centered on datapoint $\mathbf{x}_i$.



For the low-dimensional conterparts $\mathbf{y}_i$ and $\mathbf{y}_j$ of $\mathbf{x}_i$ and $\mathbf{x}_j$ it is possible to compute a similar conditional probability

$$q_{j|i} = \frac{exp(-||\mathbf{y}_i - \mathbf{y}_j||^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2)} \quad with \quad q_{i|i} = 0$$

The cost function to optimize: $\sum_{i,j} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$

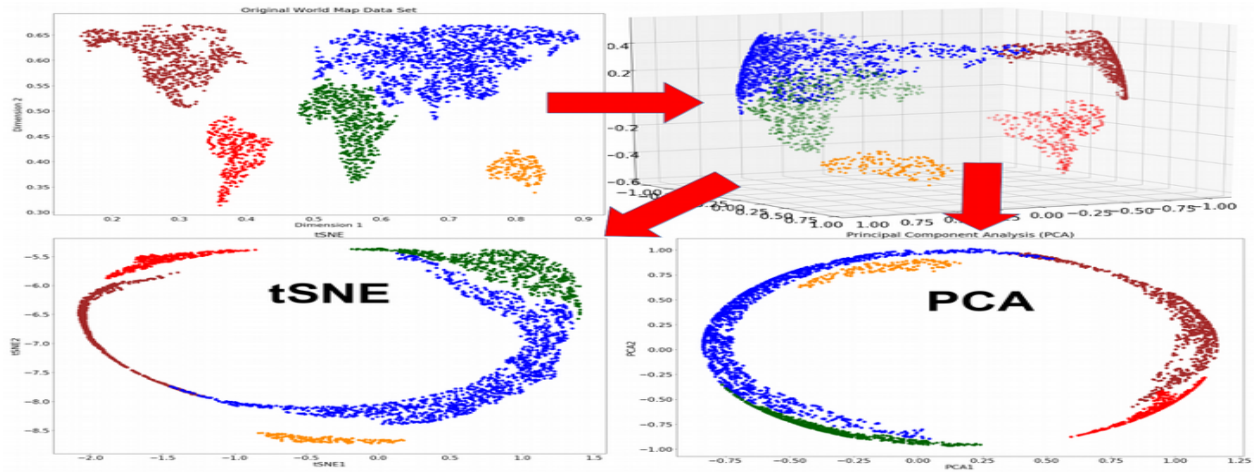
t-SNE: Stochastic Neighbor Embedding

- $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$
- Perplexity $2^{-\sum_j p_{j|i} \log_2 p_{j|i}}$ and σ_i optimal
- $q_{ij} = \frac{(1+||\mathbf{y}_i-\mathbf{y}_j||^2)^{-1}}{\sum_{k\neq\ell}(1+||\mathbf{y}_k-\mathbf{y}_\ell||^2)^{-1}}$
- The cost function to optimize $KL(P,Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$
- Gradient descent method $\frac{\partial KL(P,Q)}{\partial \mathbf{y}_i} = 4\sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}$

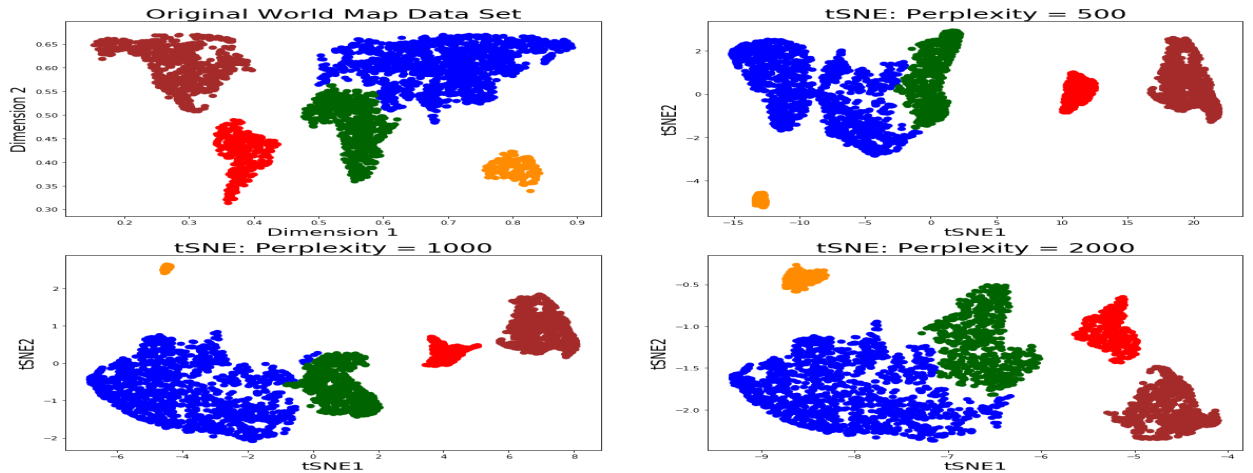
Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $X = \{x_1, x_2, ..., x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
  set $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  for t=1 **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  end
end

t-SNE



Perplexity can be interpreted as a smooth measure of the effective number of neighbors.The typical values are between 5 and 50 but...



UMAP: Uniform Manifold Approximation and Projection[a]

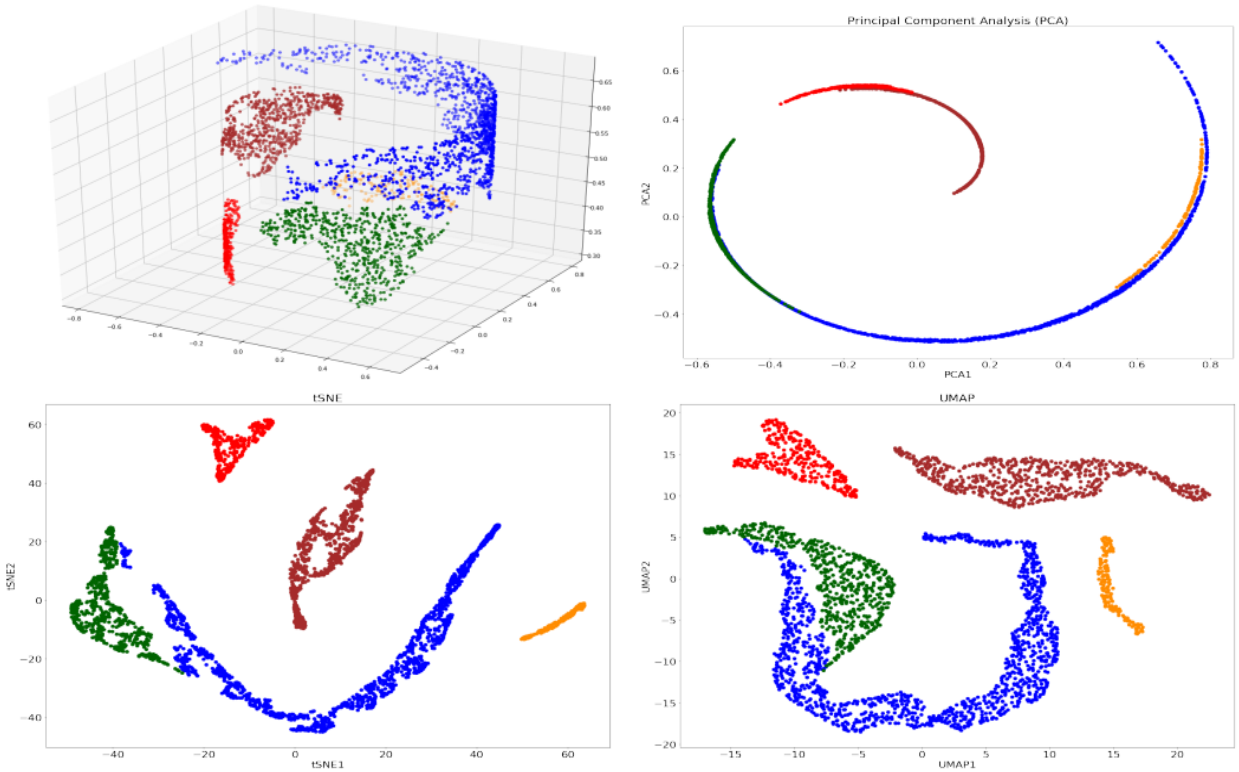
[a] McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

The t-SNE performance suffers with large datasets and using it correctly can be challenging.

- $p_{i|j} = e^{-\frac{d(\mathbf{x}_i, \mathbf{x}_j) - \rho_i}{\sigma_i}}$
- Instead of perplexity, UMAP is based on the number of *Nearest Neighbours* $2^{\sum_i p_{ij}}$
- $p_{ij} = p_{i|j} + p_{j|i} - p_{i|j} p_{j|i}$
- $q_{ij} = \frac{1}{1 + a(\mathbf{y}_i - \mathbf{y}_j)^{2b}}$
- Cross entropy as objective function

$$CE(P, Q) = \sum_{i,j} \left(p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}} \right)$$

- Unlike t-SNE(random initialisation), the initialisation dans UMAP is initialized with Graph Laplacian
- Stochastic Gradient Descent (SGD) was applied instead of the regular (GD)


Principal Component Analysis (PCA)
PCA1
PCA2
tSNE
tSNE1
tSNE2
UMAP
UMAP1
UMAP2

- Evaluation on document embedding

Table: Datasets' description.

	classic3	classic4	BBC	DBPedia	AG-news
Clusters	3	4	5	14	4
Balance	0.71	0.32	0.76	0.92	0.97
Samples	3,891	7,095	2,225	12,000	8,000

Table: Clustering scores (NMI in %) obtained using the tandem approach on different text representations. Three DR techniques are used: PCA, PCA_w and UMAP (all with $d' = 10$) and are compared to the "raw" representations (without any post-processing).

Representation		classic3				classic4				BBC				DBPedia				AG-news			
		raw	pca	pca_w	umap	raw	pca	pca_w	umap	raw	pca	pca_w	umap	raw	pca	pca_w	umap	raw	pca	pca_w	umap
Word2vec		86.7	86.5	91.1	96.2	22.8	22.8	45.9	75.0	79.6	79.1	81.8	88.4	66.8	61.4	60.6	71.7	55.7	54.6	46.9	59.5
GloVe		88.7	88.3	89.6	96.2	54.7	54.2	65.1	73.2	73.8	72.7	79.4	87.8	72.5	63.7	63.0	74.9	52.9	52.4	50.5	55.9
fastText		79.4	78.6	91.4	93.1	21.8	21.8	45.1	71.8	43.9	42.7	69.7	75.8	45.3	37.8	34.7	68.0	3.0	3.2	37.8	30.1
BERT	last	93.3	93.1	94.8	97.1	20.3	20.1	55.1	72.8	76.7	75.6	66.9	77.9	53.3	45.1	43.7	55.8	0.2	0.2	0.2	19.6
	last2	93.3	92.8	95.0	95.7	55.2	55.0	57.3	73.2	77.0	76.2	64.1	77.3	47.4	44.3	44.6	55.6	18.1	17.7	21.0	19.1
	all	90.0	89.9	94.8	96.4	68.0	67.7	71.2	74.4	78.5	76.3	79.7	86.7	67.7	62.0	61.3	71.8	48.5	43.6	47.4	55.6
SBERT	last	87.3	86.9	88.8	93.5	60.7	59.9	62.6	67.7	79.7	72.4	79.7	81.0	37.5	27.2	26.9	39.7	19.8	18.6	24.9	38.8
	last2	88.7	87.7	89.7	94.1	60.4	59.7	62.8	67.6	78.8	75.6	81.4	81.6	43.5	30.1	30.2	40.2	26.9	24.9	27.0	38.7
	all	89.5	89.0	90.8	95.1	46.2	45.8	66.3	61.8	69.4	68.2	67.3	82.7	50.5	49.2	54.9	54.4	35.5	34.1	33.6	41.8
SBERT-CT	last	91.2	90.4	93.2	96.1	66.1	65.0	67.0	68.3	80.7	76.9	80.7	81.4	51.7	37.8	40.0	56.6	41.5	39.8	42.3	53.3
	last2	90.7	90.4	93.0	95.9	66.4	65.9	67.9	69.9	82.1	79.3	84.5	82.3	62.9	41.6	43.3	60.2	43.9	43.4	44.9	53.3
	all	88.7	88.3	90.8	94.9	64.3	63.9	67.5	71.7	74.8	74.2	74.9	83.6	62.9	54.5	57.3	71.4	51.2	49.8	50.6	57.3

- UMAP+kmeans » PCA+kmeans
- UMAP+kmeans » kmeans on original data

Spectral clustering (NJW[a])

[a] Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 14.

$$\underset{\mathbf{B}\in\mathbb{R}^{n\times g}}{\text{argmax}}\ Tr(\mathbf{B}^\top\mathbf{W}\mathbf{B}) \text{ s.t } \mathbf{B}^\top\mathbf{B}=\mathbf{I}$$

1. Construct an $n \times n$ positive semi-definite similarity matrix (or kernel) $\mathcal{K}$, where κ_{ij} quantifies the similarity between data samples i and j.
2. Compute the normalized graph Laplacian defined by $\mathbf{W} = \mathbf{D}^{-\frac{1}{2}}\mathcal{K}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is a diagonal matrix with $d_{ii} = \sum_j \kappa_{ij}$.
3. Let $\mathbf{B}$ denote a $n \times g$ matrix with columns as the top g eigenvectors of $\mathbf{W}$.
4. Normalize each row of $\mathbf{B}$ to obtain $\mathbf{V}$.
5. Run the `k-means` algorithm to cluster the row vectors of $\mathbf{V}$ into g clusters $\{\mathcal{C}_1, \ldots, \mathcal{C}_g\}$.
6. Assign example $\mathbf{x}_i$ to cluster $\mathcal{C}_k$ if the i-th row of $\mathbf{V}$ belongs to $\mathcal{C}_k$.

SC-EDAE[a].

[a] Affeldt, S., Labiod, L., & Nadif, M. (2020). Spectral clustering via ensemble deep autoencoder learning (SC-EDAE). Pattern Recognition, 108, 107522

$$\underset{\mathbf{B}\in\mathbb{R}^{n\times g}}{\text{argmax}}\ Tr(\mathbf{B}^\top\tilde{\mathbf{W}}\mathbf{B})\ \text{s.t}\ \mathbf{B}^\top\mathbf{B}=\mathbf{I}$$

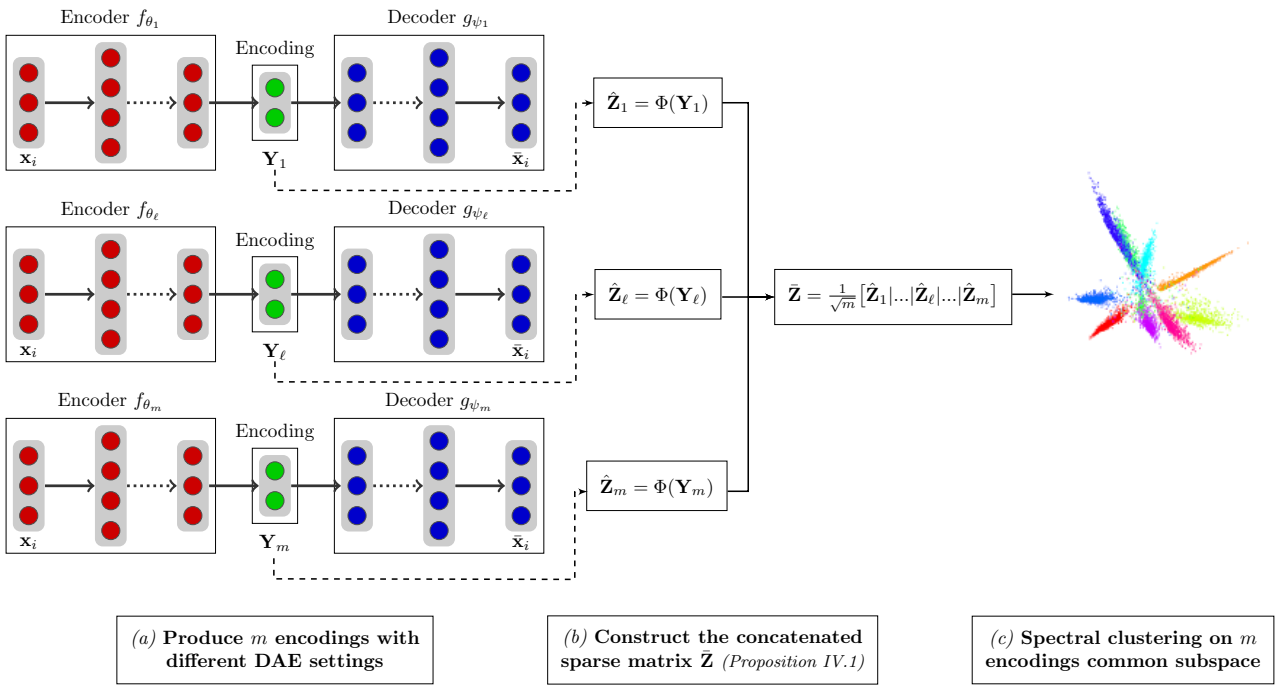


Figure: Scheme of SC-EDAE The SC-EDAE algorithm computes first m encodings from DAE with different hyperparameters settings (a), then generates m sparse affinity matrix, $\{\hat{\mathbf{Z}}_\ell\}_{\ell\in[\mathbf{1},m]}$, that are concatenated in $\bar{\mathbf{Z}}$ (b), and finally performs a SVD on the ensemble graph affinity matrix $\bar{\mathbf{Z}}$ (c).

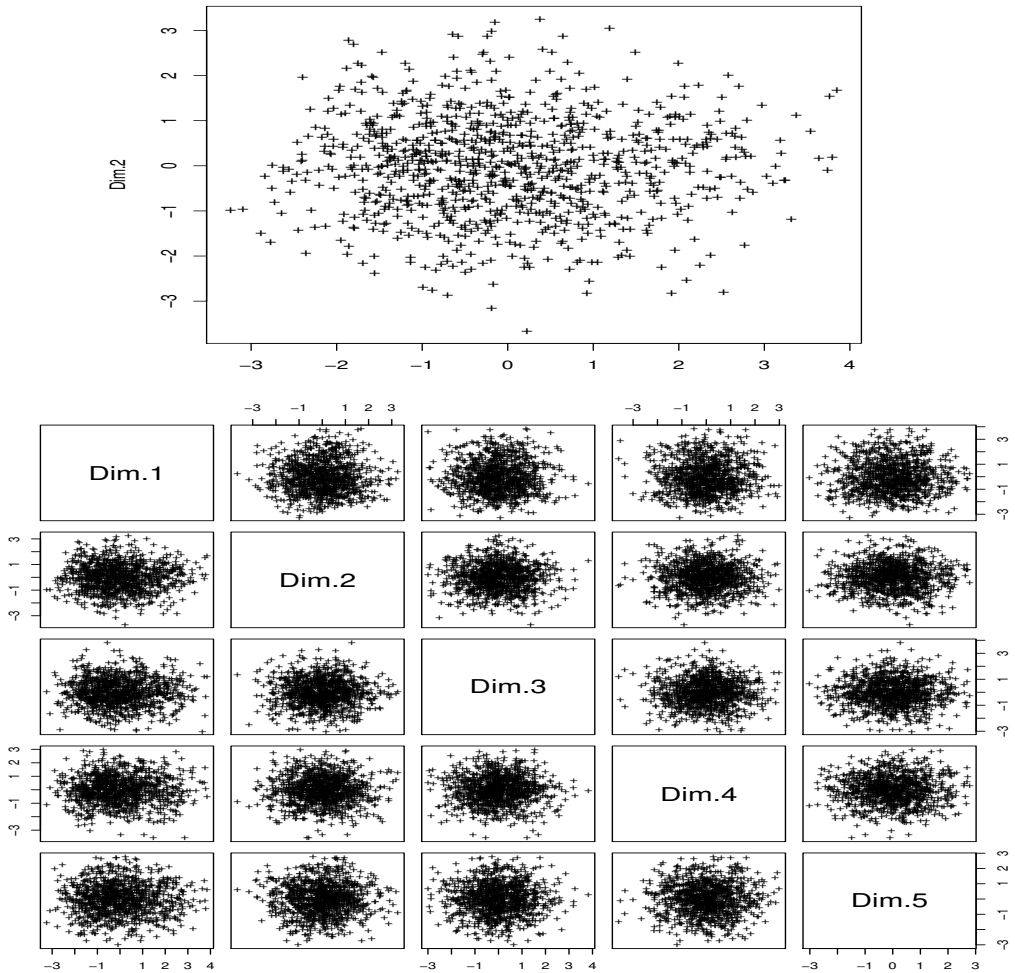
Limitations of tandem approach

- The disadvantage of this approach is that consists in optimizing two different objectives.
- Spectral low-dimensional embedding and clustering are successively and not simultaneously used.
- Certain obtained continuous low-dimensional embedding can deviate far from the clustering solution, thereby affecting the partition quality.
- Finally, due to the computational complexity of $O(n^3)$ in general, with n the number of data points, the applicability of spectral clustering for large-scale problems remains limited.

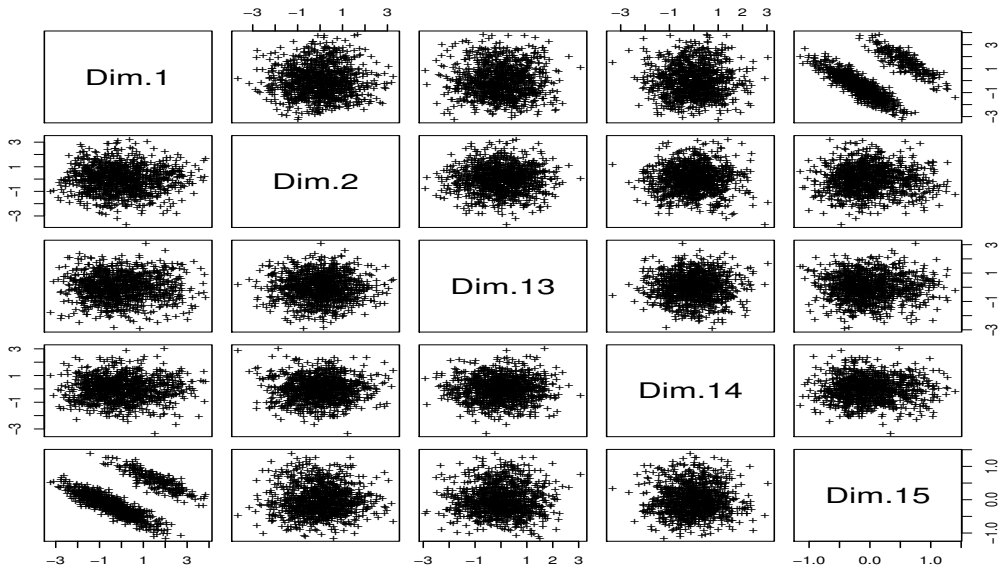
Outline

Example

Dataset (1000 $\times$ 15): Often we tend to make a PCA and from the first axes (components) we apply a clustering method.



Chang dataset



- Only EM far outperforms other methods (accuracy: 100%)

K-means

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote a centered and standardized $n \times d$ data matrix,

- $\mathbf{Z} \in \{0,1\}^{n \times g}$ represents cluster memberships of
  - n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ into the g clusters.
- $\mathbf{G} \in \mathbb{R}^{g \times d}$ represents cluster centroids $\mathbf{g}_1, \ldots, \mathbf{g}_g$

The criterion

$$W(\mathbf{Z}, \mathbf{G}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} ||\mathbf{x}_i - \mathbf{g}_k||^2$$

where $z_{ik} \in \{0,1\}$, $\sum_k z_{ik} = 1$

Matrical expression

$$W(\mathbf{Z}, \mathbf{G}) = ||\mathbf{X} - \mathbf{Z}\mathbf{G}||^2$$

Principal Component Analysis

Let $\mathbf{X}$ denote a centered and standardized $n \times d$ data matrix,

- $\mathbf{B}$ is a $d \times p$ columnwise orthonormal loadings matrix, i.e., $\mathbf{B}^\top \mathbf{B} = \mathbf{I}$, where p is the user supplied dimensionality of the reduced space
- PCA finds $\mathbf{C}$ (principal components) and $\mathbf{B}$ (loadings matrix) by minimizing

$$\underset{\mathbf{C},\mathbf{B}}{\operatorname{argmin}} ||\mathbf{X} - \mathbf{C}\mathbf{B}^\top||^2 \quad s.t \quad \mathbf{B}^\top \mathbf{B} = \mathbf{I}$$

Solving fo the optimal $\mathbf{C}$ while fixing $\mathbf{V}$ is given by $\mathbf{C} = \mathbf{X}\mathbf{B}$

PCA objective function

$$\underset{\mathbf{B}}{\operatorname{argmin}} ||\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^\top||^2 \quad s.t \quad \mathbf{B}^\top \mathbf{B} = \mathbf{I}$$

Let $\mathbf{X}$ denote a centered and standardized $n \times d$ data matrix,

- $\mathbf{B}$ is a $d \times p$ columnwise orthonormal loadings matrix, i.e., $\mathbf{B}^\top \mathbf{B} = \mathbf{I}$, where p is the user supplied dimensionality of the reduced space.
- $\mathbf{Z}$ is the $n \times g$ binary matrix indicating cluster memberships.
- $\mathbf{G}$ denotes the $g \times p$ cluster centroid matrix.

- Objective Function of Reduced k-means[a]:

$$\underset{\mathbf{Z},\mathbf{G},\mathbf{B}}{\operatorname{argmin}} ||\mathbf{X} - \mathbf{Z}\mathbf{G}\mathbf{B}^\top||^2 \quad s.t \quad \mathbf{Z} \in \{0,1\}^{n \times g}, \mathbf{B}^\top \mathbf{B} = \mathbf{I}$$

- We can show that this leads to minimize

$$||\mathbf{X} - \mathbf{Z}\mathbf{G}\mathbf{B}^\top||^2 = ||\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^\top||^2 + \underline{||\mathbf{X}\mathbf{B} - \mathbf{Z}\mathbf{G}||^2}$$

- Yamamoto and Wang (2014) insert the solution for cluster means $\mathbf{G} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X}\mathbf{B}$

$$||\mathbf{X} - \mathbf{Z}\mathbf{G}\mathbf{B}^\top||^2 = ||\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}^\top||^2 + ||\mathbf{X}\mathbf{B} - \mathbf{P}\mathbf{X}\mathbf{B}||^2$$

where $\mathbf{P} = \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$

[a] DE SOETE, Geert et CARROLL, J. Douglas. K-means clustering in a low-dimensional Euclidean space. In : New approaches in classification and data analysis. Springer, Berlin, Heidelberg, 1994. p. 212-219.

Joint approach

methods	criterion to minimize	constraints
FKM[9]	$\|\|\mathbf{XB} - \mathbf{ZG}\|\|^2$	$\mathbf{Z} \in \{0,1\}^{n\times g}, \mathbf{B}^\top\mathbf{B} = \mathbf{I}$
RKM	$\|\|\mathbf{X} - \mathbf{ZGB}^\top\|\|^2$	$\mathbf{Z} \in \{0,1\}^{n\times g}, \mathbf{B}^\top\mathbf{B} = \mathbf{I}$
ClusPCA[10]	$\alpha\|\|\mathbf{X} - \mathbf{XBB}^\top\|\|^2 + (1-\alpha)\|\|\mathbf{XB} - \mathbf{PXB}\|\|^2$	$\mathbf{Z} \in \{0,1\}^{n\times g}, \mathbf{B}^\top\mathbf{B} = \mathbf{I}$
Semi-NMF-PCA[11]	$\|\|\mathbf{X} - \mathbf{ZGB}^\top\|\|^2$	$\mathbf{Z} \geq 0, \mathbf{B}^\top\mathbf{B} = \mathbf{I}$
F-Semi-NMF-PCA	$\|\|\mathbf{X} - \mathbf{ZGB}^\top\|\|^2$	$\mathbf{Z} \in \{0,1\}^{n\times g}, \mathbf{B}^\top\mathbf{B} = \mathbf{I}$
RF-Semi-NMF-PCA	$\|\|\mathbf{X} - \mathbf{ZGB}^\top\|\|^2 + \alpha\mathbf{Trace}(\mathbf{Z}^\top(\mathbf{D}-\mathbf{W})\mathbf{Z})$ $\mathbf{W}$ is a k-nearest neighbor data graph $\mathbf{D}$ is a diagonal matrix where $D_{ii} = \sum_j W_{ij}$ α is trade-off the contribution of graph regularizing	$\mathbf{Z} \in \{0,1\}^{n\times g}, \mathbf{B}^\top\mathbf{B} = \mathbf{I}$

- When $\alpha = 1$ ClusPCA is a tandem approach
- RKM and F-Semi-NMF-PCA are comparable
- With RKM $\mathbf{B}$ is obtained by eignedecomposition of $\frac{1}{2}\mathbf{X}^\top\mathbf{PX}$ where
- With F-Semi-NMF-PCA $\mathbf{B}$ is obtained by svd of $\mathbf{X}^\top\mathbf{PXB}$

[9] Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. Computational Statistics and Data Analysis, 37(1), 49-64.

[10] Yamamoto, M., & Hwang, H. (2014). A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika, 41(1), 115-129.

[11] Allab, K., Labiod, L., & Nadif, M. (2016). A semi-NMF-PCA unified framework for data clustering. IEEE Transactions on Knowledge and Data Engineering, 29(1), 2-16.

Package R: clustrd

FKM, RKM and CLusPCA and other variants for categorical data are available [12]

Arguments	Description
`data`	Data set with metric variables.
`nclus`	Number of clusters.
`ndim`	Dimensionality of the solution.
`method`	Specifies the method. Options are `"RKM"` for reduced K-means and `"FKM"` for factorial K-means (default = `"RKM"`).
`alpha`	Adjusts for the relative importance of RKM and FKM in the objective function; `alpha` = 0.5 leads to reduced K-means, `alpha` = 0 to factorial K-means, and `alpha` = 1 reduces to the tandem approach.
`center`	A logical value indicating whether the variables should be shifted to be zero centered before the analysis takes place (default = `TRUE`).
`scale`	A logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place (default = `TRUE`).
`rotation`	Specifies the method used to rotate the factors. Options are `none` for no rotation, `varimax` for varimax rotation with Kaiser normalization and `promax` for promax rotation (default = `"none"`).
`nstart`	Number of random starts (default = 100).
`smartStart`	If `NULL` then a random cluster membership vector is generated. Alternatively, a cluster membership vector can be provided as a starting solution.
`seed`	An integer that is used as argument by `set.seed()` for offsetting the random number generator when `smartStart = NULL`. The default value is `NULL`.

[12] Markos, A., Iodice D'Enza, A., & van de Velden, M. (2019). Beyond tandem analysis: Joint dimension reduction and clustering in R. Journal of Statistical Software (Online), 91(10).

Spectral clustering and data embedding

Many ways to achieve this goal [13], RSDE[14] considers clustering and data embedding simultaneously.

RSDE

- RSDE alternates spectral clustering and the dimensionality reduction iteratively.
- RSDE relies on a matrix decomposition technique to simultaneously learn a spectral data embedding B, a clustering matrix G and a rotation matrix Q which closely maps out the continuous spectral embedding.

$$\min_{B,Q,Z,M} ||\mathbf{W} - \mathbf{B}\mathbf{M}^\top||^2 + \lambda||\mathbf{B} - \mathbf{Z}\mathbf{Q}||^2 \quad \text{s.t.}, \mathbf{B}^\top\mathbf{B} = \mathbf{I}, \quad \mathbf{Q}^\top\mathbf{Q} = \mathbf{I} \quad \mathbf{Z} \in \{0,1\}^{n\times g}.$$

- $\mathbf{W} = \mathbf{D}^{-\frac{1}{2}}\mathcal{K}\mathbf{D}^{-\frac{1}{2}}$, where $\mathbf{D}$ is a diagonal matrix with $d_{ii} = \sum_j \kappa_{ij}$.
- $\mathbf{M}$ of size $(n \times g)$ matrix introduced to improve the efficiency of the optimization
- $\mathbf{Z}$ of size $(n \times g)$ is a cluster membership matrix,
- $\mathbf{B}$ of size $(n \times g)$ is the embedding matrix;
- $\mathbf{Q}$ of size $(g \times g)$ is an orthonormal rotation matrix which most closely maps $\mathbf{B}$ to $\mathbf{Z}$.

[13] Allab, K., Labiod, L., & Nadif, M. (2018). Simultaneous spectral data embedding and clustering. IEEE transactions on neural networks and learning systems, 29(12), 6396-6401.

[14] Labiod, L., & Nadif, M. (2021). Efficient regularized spectral data embedding. Advances in Data Analysis and Classification, 15(1), 99-119.

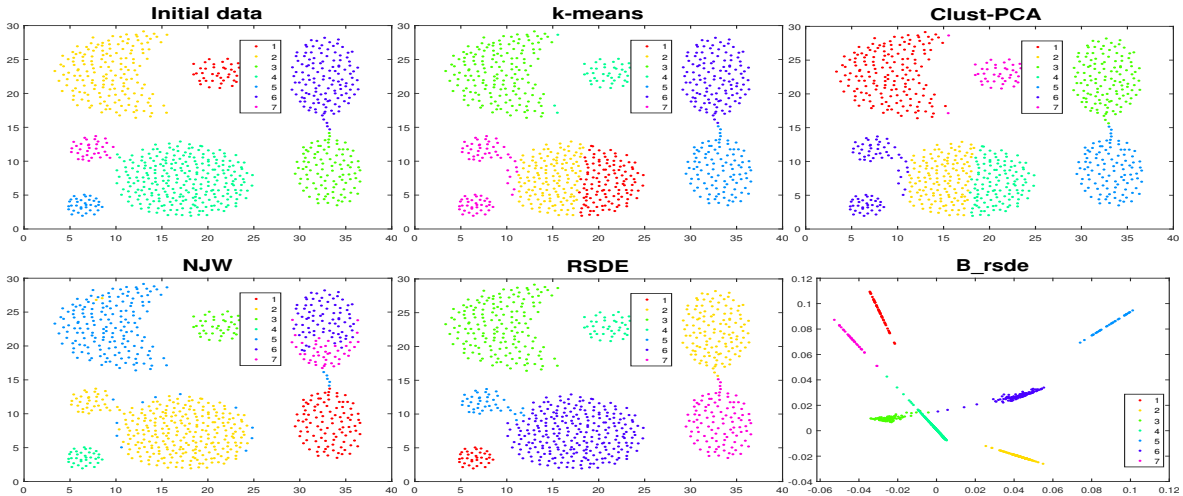
Table: Comparison of methods

Data	Mea.	PCA-KM	LDA-KM	NSC	NJW	RSDE-KM	RSDE
JAFFE	**ACC**	0.95±0.04	0.81±0.06	0.84±0.07	0.70±0.08	0.85±0.08	**0.98±0.00**
	NMI	0.95±0.02	0.85±0.04	0.90±0.04	0.82±0.05	0.91±0.04	**0.97±0.00**
UMIST	**ACC**	0.49±0.03	0.39±0.03	0.55±0.03	0.56±0.03	0.55±0.04	**0.59±0.01**
	NMI	0.64±0.02	0.56±0.02	0.70±0.02	0.68±0.02	0.68±0.02	**0.70±0.01**
MNIST5	**ACC**	0.52±0.00	0.55±0.01	0.55±0.06	**0.65±0.03**	0.63±0.05	0.64±0.02
	NMI	0.48±0.00	0.52±0.00	0.57±0.04	**0.68±0.03**	0.66±0.03	0.67±0.00
MFEA	**ACC**	0.61±0.01	0.77±0.08	0.70±0.07	0.84±0.06	0.84±0.07	**0.96±0.00**
	NMI	0.58±0.01	0.75±0.04	0.72±0.04	0.86±0.03	0.86±0.03	**0.92±0.00**
ORL	**ACC**	0.59±0.03	0.49±0.03	0.64±0.01	0.58±0.03	0.62±0.02	**0.66±0.01**
	NMI	0.76±0.02	0.68±0.02	0.79±0.01	0.75±0.02	0.78±0.01	**0.80±0.00**
	Purity	0.67±0.03	0.59±0.03	0.67±0.01	0.66±0.02	0.66±0.01	**0.68±0.01**
COIL20	**ACC**	0.58±0.02	0.59±0.04	0.50±0.03	0.63±0.04	0.77±0.05	**0.82±0.01**
	NMI	0.71±0.02	0.72±0.02	0.67±0.02	0.79±0.03	0.88±0.02	**0.90±0.01**
USPS	**ACC**	0.13±0.00	0.13±0.00	0.68±0.05	0.87±0.07	0.77±0.06	**0.81±0.00**
	NMI	0.01±0.00	0.01±0.00	0.67±0.03	0.87±0.04	0.83±0.03	**0.85±0.00**
PIE	**ACC**	0.77±0.02	0.36±0.02	0.73±0.03	0.65±0.03	0.84±0.03	**0.92±0.01**
	NMI	0.93±0.01	0.60±0.02	0.88±0.01	0.82±0.02	0.94±0.01	**0.98±0.00**

Remarks

The matrix $\mathbf{M}$ is an auxiliary variable which has a dual objective.

- $\mathbf{M}$ makes possible to reduce the computation time of $\mathbf{B}$, at each iteration, $\mathbf{B}$ is obtained by an SVD of $(\mathbf{W}^\top \mathbf{M} + \lambda \mathbf{G}\mathbf{Q})_{n \times k}$. In other words, with RSDE we implicitly use the idea of the Power method to speed up the computation of eigenvectors. Note that the update of $\mathbf{B}$ is performed on a matrix regularized by $\mathbf{G}\mathbf{Q}$.
- Unlike spectral clustering which requires $\mathbf{W}$ to be symmetric, $\mathbf{M}$ allows an implicit way of making graph $\mathbf{W}$ symmetric. Let $\mathbf{M}^*$ denote $\mathbf{W}^\top \mathbf{B}$, plugging $\mathbf{M}^*$ into the first term $\mathbf{B}$ can be derived from $\max_{\mathbf{B}} Tr\{\mathbf{B}^\top \mathbf{W}\mathbf{W}^\top \mathbf{B} + \lambda \mathbf{B}^\top \mathbf{G}\mathbf{Q})\}$ where $\mathbf{W}\mathbf{W}^\top$ is symmetric.



AGG dataset with 7 clusters

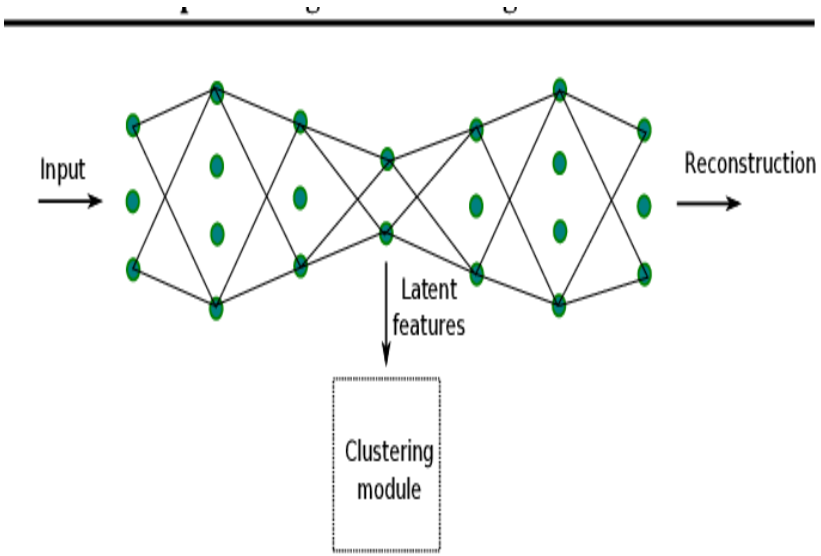
DCN: Deep Clustering Network[15]

Let $\mathbf{x}_i \in \mathbb{R}^{p\times 1}$ the k-means criterion takes the following form

$$\underset{M\in\mathbb{R}^{p\times g},\mathbf{z}_i\in\{0,1\}^{g\times 1}}{\operatorname{argmin}} \sum_i ||\mathbf{x}_i - \mathbf{M}\mathbf{z}_i||^2$$

The clustering loss is defined as:

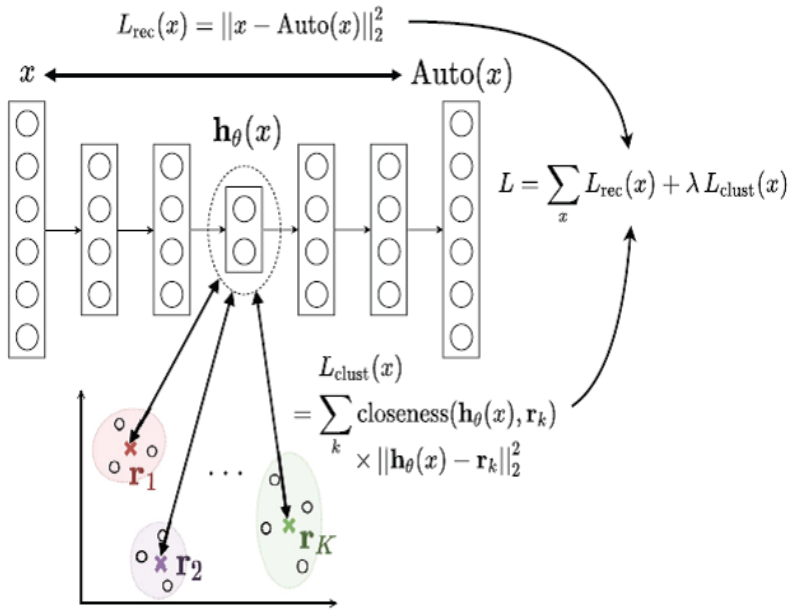
$$\underset{M\in\mathbb{R}^{p\times g},\mathbf{z}_i\in\{0,1\}^{g\times 1}\theta,\Phi}{\operatorname{argmin}} \sum_i \left(\ell(g_\Phi of_\theta(\mathbf{x}_i),\mathbf{x}_i) + \frac{\lambda}{2}||f_\theta(\mathbf{x}_i) - \mathbf{M}\mathbf{z}_i||^2\right)$$



[15] Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017, July). Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In international conference on machine learning (pp. 3861-3870).

DKM: Deep k-means [16]

In contrast to DCN, The authors propose the DKM framework such that all network and clustering parameters are updated simultaneously.



[16] Fard, M. M., Thonet, T., & Gaussier, E. (2020). Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognition Letters, 138, 185-192.

Outline

Notation

Data

- matrix $\mathbf{X} = (x_{ij})$
- $i \in I$ set of n rows, $j \in J$ set of d columns

Partition of I in g clusters

- $\mathbf{z} = (z_1, \ldots, z_i, \ldots, z_n)$ where $z_i \in \{1, \ldots, g\}$
- $\mathbf{Z} = (z_{ik})$ where $z_{ik} = 1$ if $i \in k$th cluster and $z_{ik} = 0$ otherwise

z	Z		
3	0	0	1
2	0	1	0
3	0	0	1
2	0	1	0
1	1	0	0

Partition of J in s clusters

- $\mathbf{w} = (w_1, \ldots, w_j, \ldots, w_d)$ where $w_j \in \{1, \ldots, s\}$
- $\mathbf{W} = (w_{ij})$ where $w_{j\ell} = 1$ if $j \in \ell$th cluster and $w_{j\ell} = 0$ otherwise

From Z and W

- Block or co-cluster (k, ℓ) is defined by the (i, j)'s with $z_{ik} w_{j\ell} = 1$ then

Co-clustering algorithms (1)

Hard co-clustering

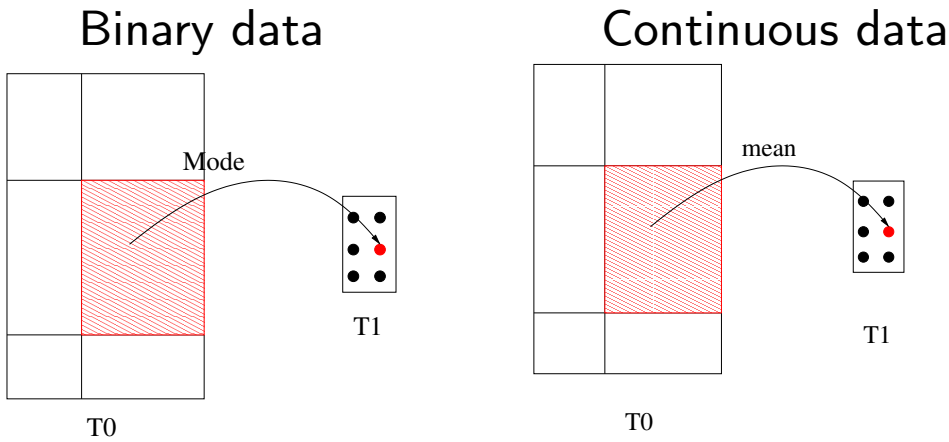
- Model: $\mathbf{X} = \mathbf{ZAW}^T + \mathbf{R}$
- $\mathbf{Z}$, $\mathbf{W}$ are binary matrices

Optimization of criterion $\mathcal{C}(\mathbf{Z}, \mathbf{W}, \mathbf{A})$

- $\mathbf{Z}$ and $\mathbf{W}$ partitions of I and J
- $\mathbf{A} = (a_{k\ell})$ summary matrix of dimensions $g \times s$ having the same structure that the initial data matrix
- $\mathcal{C}$ depends on the type of data.

Co-clustering algorithms (2)

General principle



Criteria[a]

[a] Govaert, G. (1995). Simultaneous clustering of rows and columns. Control and Cybernetics, 24, 437-458.

Data	$a_{k\ell}$	Criterion $\mathcal{C}$
Binary	Mode	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} \lvert x_{ij} - a_{k\ell} \rvert$
Continuous	Mean	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} - a_{k\ell})^2 = \lvert\lvert \mathbf{X} - \mathbf{ZAW}^T \rvert\rvert^2$

Binary data: CROBIN

Algorithm

Alternated minimization of the criterion $\mathcal{C}(\mathbf{Z}, \mathbf{W}, \mathbf{A})$

- minimization of $\mathcal{C}(\mathbf{Z}, \mathbf{A}|\mathbf{W}) = \sum_{i,k,\ell} z_{ik}|u_{i\ell} - \#w_\ell a_{k\ell}|$ where $u_{i\ell} = \sum_j w_{j\ell} x_{ij}$
  - *nuées dynamiques* on **U**
- minimization of $\mathcal{C}(\mathbf{W}, \mathbf{A}|\mathbf{Z}) = \sum_{j,k,\ell} w_{j\ell}|v_{j\ell} - \#z_k a_{k\ell}|$ where $v_{kj} = \sum_i z_{ik} x_{ij}$
  - *nuées dynamiques* on **V**

Data

	abcdefghij
y_1	1010001101
y_2	0101110011
y_3	1000001100
y_4	1010001100
y_5	0111001100
y_6	0101110101
y_7	0111110111
y_8	1100111011
y_9	0100110000
y_{10}	1010101101
y_{11}	1010001100
y_{12}	1010000100
y_{13}	1010001101
y_{14}	0010011100
y_{15}	0010010100
y_{16}	1111001100
y_{17}	0101110011
y_{18}	1010011101
y_{19}	1010001000
y_{20}	1100101100

Reorganized matrix

	a c g h	*b d e f i j*
y_2	0 0 0 0	1 1 1 1 1 1
y_6	0 0 0 1	1 1 1 1 0 1
y_7	0 1 0 1	1 1 1 1 1 1
y_8	1 0 1 0	1 0 1 1 1 1
y_9	0 0 0 0	1 0 1 1 0 0
y_{17}	0 0 0 0	1 1 1 1 1 1
y_1	1 1 1 1	0 0 0 0 0 1
y_3	1 0 1 1	0 0 0 0 0 0
y_4	1 1 1 1	0 0 0 0 0 0
y_5	0 1 1 1	1 1 0 0 0 0
y_{10}	1 1 1 1	0 0 1 0 0 1
y_{11}	1 1 1 1	0 0 0 0 0 0
y_{12}	1 1 0 1	0 0 0 0 0 0
y_{13}	1 1 1 1	0 0 0 0 0 1
y_{14}	0 1 1 1	0 0 0 1 0 0
y_{15}	0 1 0 1	0 0 0 1 0 0
y_{16}	1 1 1 1	1 1 0 0 0 0
y_{18}	1 1 1 1	0 0 0 1 0 1
y_{19}	1 1 1 0	0 0 0 0 0 0
y_{20}	1 0 1 1	1 0 1 0 0 0

Summary

0	1
1	0

Homogeneity

0.80	0.87
0.86	0.84

Continuous Data

Minimization of the criterion $\mathcal{C}(\mathbf{Z}, \mathbf{W}, \mathbf{A}) = ||\mathbf{X} - \mathbf{ZAW}^T||^2$

Alternating Exchanges [a]

[a] Gaul, W., & Schader, M. (1996). A new algorithm for two-mode clustering. In Data analysis and information systems (pp. 15-23). Springer, Berlin, Heidelberg.

- Choose initial $\mathbf{Z}$ and $\mathbf{W}$
- repeat the following steps
  - update $\mathbf{A}$, $a_{k\ell} = \sum_{i,j} z_{ik} w_{j\ell} x_{ij} / \sum_{i,j} z_{ik} w_{j\ell}$
  - update $\mathbf{Z}$, $z_{ik} = 1$ if $c_{ik} = min_{1 \leq k \leq g} c_{ik}$ where $c_{ik} = \sum_{j,\ell} w_{j\ell}(x_{ij} - a_{k\ell})^2$
  - update $\mathbf{A}$
  - update $\mathbf{W}$, $w_{j\ell} = 1$ if $d_{j\ell} = min_{1 \leq \ell \leq m} d_{j\ell}$ where $d_{j\ell} = \sum_{i,k} z_{ik}(x_{ij} - a_{k\ell})^2$

The Croeuc Algorithm

(a) minimization of $\mathcal{C}(\mathbf{Z}, \mathbf{A}|\mathbf{W}) = \sum_{i,k,\ell} z_{ik}(u_{i\ell} - a_{k\ell})^2$ where $u_{i\ell} = \sum_j w_{j\ell} x_{ij} / \#w_\ell$

(a.1) *k-means* on $\mathbf{U}$ and we obtain $\mathbf{Z}$

(b) minimization of $\mathcal{C}(\mathbf{W}, \mathbf{A}|\mathbf{Z}) = \sum_{j,k,\ell} w_{j\ell}(v_{j\ell} - a_{k\ell})^2$ where $v_{kj} = \sum_i z_{ik} x_{ij} / \#z_k$

(b.1) *k-means* on $\mathbf{V}$ and we obtain $\mathbf{W}$

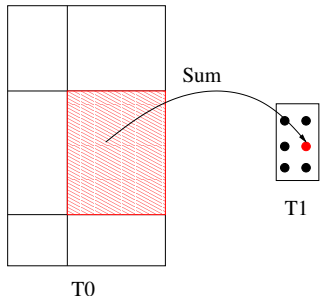
Let **X** be a two-way contingency table associated to two categorical random variables that take values in sets $I = \{1, \ldots, i, \ldots, n\}$ and $J = \{1, \ldots, j, \ldots, d\}$. The entries x_{ij} are co-occurrences of row and column categories, each of them counts the number of entities that fall simultaneously in the corresponding row and column categories.

Document-term matrices can be viewed as contingency tables

	Directional	Model	Clustering	Classification	Regression	Linear
Doc_1	1	2	5	0	0	0
Doc_2	2	1	0	5	1	0
Doc_3	10	20	50	0	0	0
Doc_4	2	0	0	0	5	3
Doc_5	0	3	0	0	1	3

Contingency table[17]

- Summary of **X** can be obtained by



- Problem: Find partitions **z** and **w** maximizing $\mathcal{A}(\mathbf{z}, \mathbf{w})$. The $(\mathbf{z}, \mathbf{w})$ is obtained in making the sums of values per block

	$v1$	$v2$	$v3$	$v4$	$v5$	**Z**
a	5	0	0	0	0	1
b	0	2	0	1	1	1
c	0	0	1	4	0	2
d	1	0	0	0	1	1
e	2	0	1	3	1	2
f	0	4	1	1	1	2
W	1	1	2	2	1	

	$v1$	$v2$	$v5$	$v3$	$v4$
a	5	0	0	0	0
b	0	2	2	0	1
d	1	0	1	0	0
c	0	0	0	1	4
e	1	0	1	1	3
f	0	0	1	1	1

11	1
3	11

- Solution: Alternated maximization of $\mathcal{A}(\mathbf{z}, J)$ and $\mathcal{A}(I, \mathbf{w})$
- Idea: Alternated application of k-means (nuées dynamiques, Diday 1971) with an appropriate metric on intermediate reduced matrices of size $(g \times d)$ and $(n \times s)$

[17] Govaert, G. (1995). Simultaneous clustering of rows and columns. Control and Cybernetics, 24, 437-458.

Measures of association

The contingency table characterizes the dependency links between the two sets, and measuring the strength of this association is a long tradition in statistics, going back to at least Pearson (1900).

Phi-squared: $\Phi^2(P_{IJ}) = \frac{\chi^2(\mathbf{X})}{N} = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_{i,j} \frac{p_{ij}^2}{p_{i.}p_{.j}} - 1$

This coefficient can be seen as an estimation of the deviation between the probabilities $\xi_{i.}\xi_{.j}$, that we would have if the two categorical random variables were independent, and the probabilities ξ_{ij}

Mutual Information: $\mathrm{MI}(P_{IJ}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.}p_{.j}}$

This measure of association is defined by $\mathrm{MI}(P_{IJ}) = H(P_I) + H(P_J) - H(P_{IJ})$ where $H(P_I)$, $H(P_J)$ are the marginal entropies, $H(P_{IJ})$ is the joint entropy of I and J.

Connections between these approaches

There exists a large variety of co-clustering methods for contingency tables which can be applied in document clustering context.

Let $P_{IJ} = (p_{ij})$ denote the sample joint probability distribution. It is a matrix of size $n \times d$ defined by $p_{ij} = \frac{x_{ij}}{N}$ where $N = \sum_{ij} x_{ij}$. The sample marginal probability distributions are defined by $p_{i.} = \sum_j p_{ij}$ and $p_{.j} = \sum_i p_{ij}$.

	$\mathbf{1}$	$\ldots$	j	$\ldots$	d	
$\mathbf{1}$	$X_{\mathbf{1}j}$	$\ldots$	$X_{\mathbf{1}j}$	$\ldots$	$X_{\mathbf{1}d}$	$X_{\mathbf{1}.}$
			$\vdots$	$\ldots$		
i	$X_{i\mathbf{1}}$	$\ldots$	X_{ij}	$\ldots$	X_{id}	$X_{i.}$
			$\vdots$	$\ldots$		
n	$X_{n\mathbf{1}}$	$\ldots$	X_{nj}	$\ldots$	X_{nd}	$X_{n.}$
	$X_{.\mathbf{1}}$	$\ldots$	$X_{.j}$	$\ldots$	$X_{.d}$	N

	$\mathbf{1}$	$\ldots$	j	$\ldots$	d	
$\mathbf{1}$	$p_{\mathbf{1}j}$	$\ldots$	$p_{\mathbf{1}j}$	$\ldots$	$p_{\mathbf{1}d}$	$p_{\mathbf{1}.}$
			$\vdots$	$\ldots$		
i	$p_{i\mathbf{1}}$	$\ldots$	p_{ij}	$\ldots$	p_{id}	$p_{i.}$
			$\vdots$	$\ldots$		
n	$p_{n\mathbf{1}}$	$\ldots$	p_{nj}	$\ldots$	p_{nd}	$p_{n.}$
	$p_{.\mathbf{1}}$	$\ldots$	$p_{.j}$	$\ldots$	$p_{.d}$	$\mathbf{1}$

Objective: Approximation of P_{IJ} while taking into account $(\mathbf{z}, \mathbf{w})$

Distribution associated to z and w defined on $I \times J$

The second distribution $Q_{IJ}^{zw} = (q_{ij}^{zw})$ defined on $I \times J$

$$q_{ij}^{zw} = q_{i.}^{z} q_{.j}^{w} \sum_{k,\ell} z_{ik} w_{j\ell} \frac{p_{k\ell}^{zw}}{p_{k.}^{z} p_{.\ell}^{w}} \qquad \forall (i,j) \in I \times J$$

Note that

$$\forall (i,j) \in bloc(g,s) \text{ we have } \sum_{k,\ell} z_{ik} w_{j\ell} \frac{p_{k\ell}^{zw}}{p_{k.}^{z} p_{.\ell}^{w}} = \frac{p_{gs}^{zw}}{p_{g.}^{z} p_{.s}^{w}} \text{ since } z_{ik} w_{j\ell} = 1$$

With this formulation we have $\sum_{i,j} q_{ij}^{zw} = 1, \quad q_{i.}^{z} = p_{i.} \quad \text{and} \quad q_{.j}^{w} = p_{.j} \qquad \forall i,j.$ We have the same margins as the initial distribution P_{IJ}.

$$q_{ij}^{zw} = p_{i.} p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \frac{p_{k\ell}^{zw}}{p_{k.}^{z} p_{.\ell}^{w}} \qquad \forall (i,j) \in I \times J$$

	1	2	3	4	5	
1	0.050	0.040	0.060	0.010	0.000	0.160
2	0.060	0.050	0.040	0.000	0.010	0.160
3	0.010	0.000	0.010	0.070	0.050	0.140
4	0.010	0.010	0.000	0.060	0.050	0.130
5	0.040	0.050	0.030	0.040	0.050	0.210
6	0.050	0.040	0.040	0.030	0.040	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

	1	2	3	4	5	
1	0.056	0.048	0.046	0.005	0.005	0.160
2	0.056	0.048	0.046	0.005	0.005	0.160
3	0.008	0.007	0.006	0.061	0.058	0.140
4	0.007	0.006	0.006	0.057	0.054	0.130
5	0.048	0.041	0.039	0.042	0.040	0.210
6	0.045	0.039	0.037	0.040	0.038	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

Table: Distributions P_{IJ} (left) and Q_{IJ}^{zw} (right)

Measures of associations associated to z and w

Using the two measures phi-squared and mutual, we obtain the following measures:

$$\Phi^2(Q_{IJ}^{zw}) = \sum_{i,j} \frac{(q_{ij}^{zw} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} \qquad \mathrm{MI}(Q_{IJ}^{zw}) = \sum_{i,j} q_{ij}^{zw} \log \frac{q_{ij}^{zw}}{p_{i.}p_{.j}}.$$

$$\Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{zw}) = D_{\Phi^2}(P_{IJ}||Q_{IJ}^{zw})$$

where $D_{\Phi^2}(P_{IJ}||Q_{IJ}^{zw}) = \sum_{i,j} \frac{(p_{ij}-q_{ij}^{zw})^2}{p_{i.}p_{.j}} = \sum_{i,j} p_{ij}\left(\frac{p_{ij}}{p_{i.}p_{.j}} - \frac{q_{ij}^{zw}}{p_{i.}p_{.j}}\right)$ can be viewed as a Φ^2 distance between the two distributions P_{IJ} and Q_{IJ}^{zw}

$$\mathrm{MI}(P_{IJ}) - \mathrm{MI}(Q_{IJ}^{zw}) = \mathrm{KL}(P_{IJ}||Q_{IJ}^{zw})$$

where $\mathrm{KL}(P_{IJ}||Q_{IJ}^{zw}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}^{zw}}$ is the Kullback-Leibler between the two distributions P_{IJ} and Q_{IJ}^{zw},

$$\mathrm{MI}(Q_{IJ}^{zw}) \leq \mathrm{MI}(P_{IJ}).$$

Co-clustering obtained by approximating the original distribution: Φ^2

- The co-clustering problem can be viewed as an approximation of the distribution P_{IJ} by a distribution according to co-clustering termed $Q_{IJ}^{\mathbf{zw}}$ by minimizing

$$\Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}})$$

To reach this goal we introduce a distribution $R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}}$ where

$$\boldsymbol{\delta} := \{\delta_{k\ell}; k = 1, \ldots, g; \ell = 1, \ldots, m\},$$

a matrix of size (g, m) where

$$\delta_{k\ell} \geq 0 \quad \forall k, \ell \quad \text{and} \quad \sum_{k,\ell} p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}} \delta_{k\ell} = 1.$$

$R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}}$ is a probability distribution. Each $\delta_{k\ell}$ plays the role of centroid of the $k\ell$th block. Using this parameter, a new distribution $R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}} := \{r_{ij}^{\mathbf{zw}\boldsymbol{\delta}}; i = 1, \ldots, n; j = 1, \ldots, d\}$ with partitions $\mathbf{z}$ and $\mathbf{w}$, and parameter $\boldsymbol{\delta}$ can be defined by

$$r_{ij}^{\mathbf{zw}\boldsymbol{\delta}} = p_{i.} p_{.j} \sum_{k,\ell} z_{ik} w_{j\ell} \delta_{k\ell} \Rightarrow r_{ij}^{\mathbf{zw}\boldsymbol{\delta}} = p_{i.} p_{.j} \delta_{k\ell} \text{ if } (i,j) \in (k\ell) \text{ and } r_{ij}^{\mathbf{zw}\boldsymbol{\delta}} = 0 \quad \textit{otherwise}$$

Phi-squared criterion and Algorithm[18]

$$\widetilde{W}_{\Phi 2}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \Phi^2(P_{IJ}) - \Phi^2(R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}}) = \sum_{i,j,k;\ell} z_{ik}\, w_{j\ell}\; p_{i.}\, p_{.j} \left(\frac{p_{ij}}{p_{i.}\, p_{.j}} - \delta_{k\ell} \right)^2$$

Algorithm 1 Croki2

input: contingency table $\mathbf{X}$, g and s the desired numbers of row column clusters;
output: partitions $\mathbf{z}$ and $\mathbf{w}$;
initialization: start with some initial partitions $\mathbf{z}$, $\mathbf{w}$; $\delta_{k\ell} \leftarrow \frac{p_{k\ell}^{\mathbf{zw}}}{(p_{k.}^{\mathbf{z}})(p_{.\ell}^{\mathbf{w}})}$;
repeat

step 1. $\mathbf{z}$: each i is assigned to the cluster k minimizing $\sum_{j,\ell} w_{j\ell} p_{.j} \left(\frac{p_{ij}}{p_{i.}\, p_{.j}} - \delta_{k\ell} \right)^2$;

step 2. $\delta_{k\ell} \leftarrow \frac{p_{k\ell}^{\mathbf{zw}}}{(p_{k.}^{\mathbf{z}})(p_{.\ell}^{\mathbf{w}})}$;

step 3. $\mathbf{w}$: each j is assigned to the cluster ℓ minimizing $\sum_{i,k} z_{ik} p_{i.} (\frac{p_{ij}}{p_{i.}\, p_{.j}} - \delta_{k\ell})^2$;

step 4. $\delta_{k\ell} \leftarrow \frac{p_{k\ell}^{\mathbf{zw}}}{(p_{k.}^{\mathbf{z}})(p_{.\ell}^{\mathbf{w}})}$;

until the change in objective function value $W_{\Phi 2}(\mathbf{z},\mathbf{w},\boldsymbol{\delta})$ is "small" (say 10^{-6})
return $\mathbf{z}$ and $\mathbf{w}$

[18] Govaert, G., & Nadif, M. (2018). Mutual information, phi-squared and model-based co-clustering for contingency tables. Advances in data analysis and classification, 12(3), 455-488.

Example

	1	2	3	4	5	
1	5	4	6	1	0	16
2	6	5	4	0	1	16
3	1	0	1	7	5	14
4	1	1	0	6	5	13
5	4	5	3	4	5	21
6	5	4	4	3	4	20
	22	19	18	21	20	100

	1	2	3	4	5	
1	0.05	0.04	0.06	0.01	0.00	0.16
2	0.06	0.05	0.04	0.00	0.01	0.16
3	0.01	0.00	0.01	0.07	0.05	0.14
4	0.01	0.01	0.00	0.06	0.05	0.13
5	0.04	0.05	0.03	0.04	0.05	0.21
6	0.05	0.04	0.04	0.03	0.04	0.20
	0.22	0.19	0.18	0.21	0.20	1.00

Table: Example of contingency table and associated joint distribution

Approximation of P_{IJ} by Q_{IJ}

	1	2	3	4	5	
1	0.050	0.040	0.060	0.010	0.000	0.160
2	0.060	0.050	0.040	0.000	0.010	0.160
3	0.010	0.000	0.010	0.070	0.050	0.140
4	0.010	0.010	0.000	0.060	0.050	0.130
5	0.040	0.050	0.030	0.040	0.050	0.210
6	0.050	0.040	0.040	0.030	0.040	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

	1	2	3	4	5	
1	0.056	0.048	0.046	0.005	0.005	0.160
2	0.056	0.048	0.046	0.005	0.005	0.160
3	0.008	0.007	0.006	0.061	0.058	0.140
4	0.007	0.006	0.006	0.057	0.054	0.130
5	0.048	0.041	0.039	0.042	0.040	0.210
6	0.045	0.039	0.037	0.040	0.038	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

Table: Distributions P_{IJ} (left) and $Q_{IJ}^{\mathbf{zw}}$ (right)

Defects of algorithms cited

- Choice of the criterion not often easily
- Implicit hypotheses unknown
- Crobin and Croeuc are not effective when the clusters are not well-separated and unbalanced
- Croeuc has the same problem where the clusters are not balanced

Aim

Propose a general framework able to formalize the hypotheses of block clustering algorithms: latent block model

- to overcome the defects of criteria and therefore to propose other criteria
- to develop other efficient algorithms

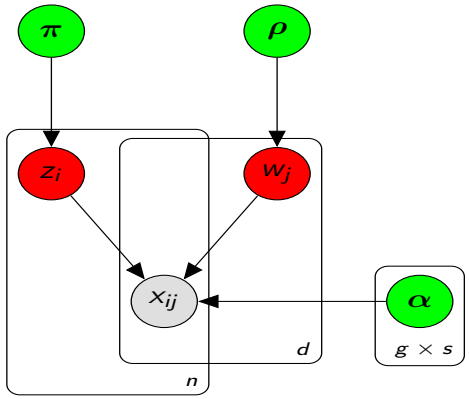
Definition[a]

[a] Govaert, G., & Nadif, M. (2003). Clustering with block mixture models. Pattern Recognition, 36(2), 463-473.

The pdf of $\mathbf{X}$:

$$f(\mathbf{X};\Theta) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_i \pi_{z_i} \prod_j \rho_{w_j} \prod_{i,j} \varphi(X_{ij};\boldsymbol{\alpha}_{z_i w_j})$$

where $\Theta = (\pi_1,\ldots,\pi_g;\rho_1,\ldots,\rho_s;\boldsymbol{\alpha}_{11},\ldots,\boldsymbol{\alpha}_{gs})$



Advantages

- Parsimonious models giving probabilistic interpretations of classical [a] criteria
- Variational EM algorithm[b]

[a] Govaert, G., & Nadif, M. (2013). Co-clustering: models, algorithms and applications. John Wiley & Sons.

[b] Govaert, G., & Nadif, M. (2005). An EM algorithm for the block mixture model. IEEE TPAMI, 27(4), 643-647.

Poisson Latent Block Models[abc]

[a] Ailem, M., Role, F., & Nadif, M. (2017). Sparse poisson latent block model for document clustering. IEEE Transactions on Knowledge and Data Engineering, 29(7), 1563-1576.

[b] Ailem, M., Role, F., & Nadif, M. (2017). Model-based co-clustering for the effective handling of sparse data. Pattern Recognition, 72, 108-122.

[c] Riverain, P., Fossier, S., & Nadif, M. (2022). Semi-supervised Latent Block Model with pairwise constraints. Machine Learning, 111(5), 1739-1764.

Flexibility of the LBMs: variant models and algorithms

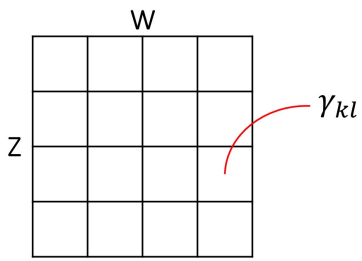


Figure: PLBM

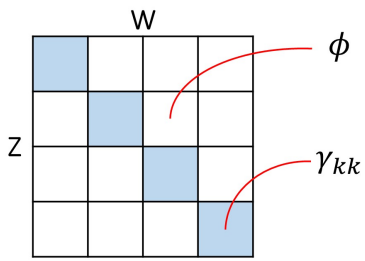


Figure: SPLBM

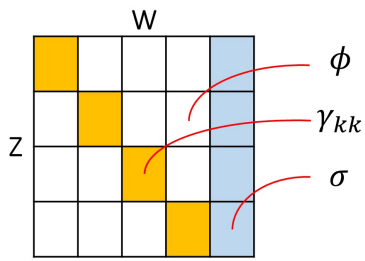


Figure: GPLBM

Datasets	Characteristics				
	#Documents	#Words	#Clusters	Sparsity (%)	Balance
CSTR	475	1000	4	96.60	0.399
CLASSIC4	7095	5896	4	99.41	0.323
SPORTS	8580	14870	7	99.14	0.036
TDT2	9394	36771	30	99.64	0.028
Yahoo_K1B	2340	21839	6	99.41	0.043
Reuters30	8067	18832	30	99.75	0.005
Reuters40	8203	18914	40	99.75	0.003

Table: Mean±sd clustering Accuracy (Acc), Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) on several real-world *documents* × *terms* matrices. For each algorithm, the column best corresponds to the trial (among the 50 runs) with the highest criterion.

datasets	per.	Skmeans	best	ITCC	best	SpecCo	best	Plbcem	best	Splbcem	best
CSTR	Acc	0.62±0.01	0.63	0.62±0.08	0.66	0.82±0.00	0.82	0.67±0.04	0.66	0.84±0.05	0.89
	NMI	0.59±0.02	0.60	0.68±0.02	0.67	0.71±0.00	0.71	0.67±0.03	0.68	**0.71±0.02**	0.74
	ARI	0.51±0.01	0.52	0.59±0.04	0.57	0.72±0.00	0.72	0.59±0.03	0.58	0.73±0.04	0.79
CLASSIC4	Acc	0.59±0.00	0.60	0.65±0.02	0.66	0.58±0.00	0.58	0.79±0.09	0.89	0.81±0.10	0.90
	NMI	0.54±0.00	0.55	0.59±0.01	0.60	0.48±0.00	0.48	0.63±0.05	0.72	0.66±0.05	0.73
	ARI	0.43±0.00	0.44	0.44±0.01	0.45	0.22±0.00	0.22	0.53±0.05	0.71	0.59±0.10	0.74
SPORTS	Acc	0.46±0.04	0.49	0.54±0.05	0.53	0.67±0.00	0.67	0.52±0.06	0.57	0.67±0.08	0.85
	NMI	0.46±0.04	0.50	0.59±0.04	0.60	0.59±0.00	0.59	0.56±0.03	0.62	0.56±0.07	0.69
	ARI	0.28±0.03	0.30	0.45±0.03	0.44	0.48±0.00	0.48	0.42±0.04	0.48	0.52±0.11	0.76
TDT2	Acc	0.55±0.02	0.57	0.57±0.03	0.59	0.81±0.02	0.83	0.56±0.02	0.59	0.81±0.04	0.83
	NMI	0.75±0.01	0.76	0.76±0.01	0.78	0.82±0.01	0.83	0.75±0.01	0.76	0.80±0.02	0.81
	ARI	0.43±0.03	0.46	0.49±0.03	0.52	0.72±0.03	0.75	0.48±0.03	0.51	0.79±0.04	0.81
Yahoo_K1B	Acc	0.57±0.03	0.57	0.57±0.04	0.61	0.79±0.03	0.79	0.60±0.03	0.63	0.79±0.05	0.79
	NMI	0.62±0.02	0.64	0.57±0.03	0.58	0.62±0.02	0.64	0.58±0.03	0.60	0.62±0.05	0.66
	ARI	0.37±0.04	0.39	0.35±0.05	0.40	0.58±0.03	0.58	0.37±0.04	0.40	0.61±0.09	0.60
REUTERS30	Acc	0.27±0.02	0.29	0.30±0.03	0.33	0.52±0.04	0.61	0.49±0.04	0.48	0.66±0.02	0.68
	NMI	0.49±0.02	0.51	0.51±0.01	0.53	0.52±0.02	0.53	0.54±0.01	0.55	0.54±0.02	0.57
	ARI	0.13±0.01	0.14	0.20±0.03	0.23	0.42±0.03	0.48	0.41±0.05	0.44	0.59±0.02	0.60
REUTERS40	Acc	0.25±0.01	0.26	0.28±0.03	0.27	0.50±0.03	0.57	0.32±0.04	0.41	0.64±0.04	0.73
	NMI	0.50±0.00	0.50	0.51±0.01	0.52	0.51±0.01	0.51	0.51±0.01	0.55	0.53±0.02	0.57
	ARI	0.11±0.00	0.11	0.18±0.03	0.18	0.41±0.02	0.46	0.22±0.04	0.31	0.59±0.06	0.71

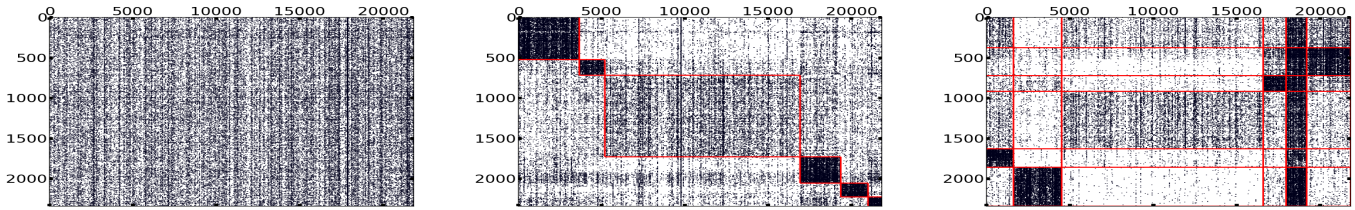


Figure: (a): Original Yahoo-K1B data - (b): Yahoo-K1B data after `splbcem` co-clustering (row NMI: 0.66 , row ARI: 0.60) - (c): Yahoo-K1B data after `plbcem` co-clustering (row NMI: 0.60 , row ARI: 0.40).

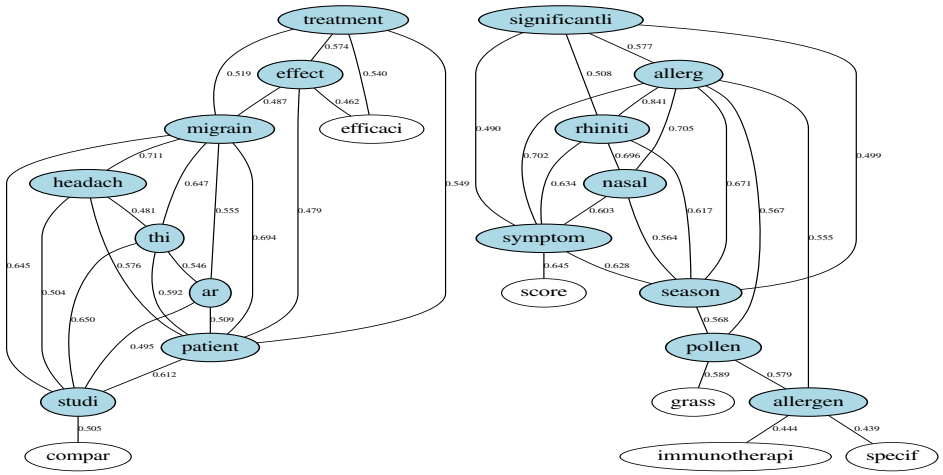


Figure: For each co-cluster, the n most frequent terms in this co-cluster are placed in a graph and connected to their k most similar neighbors according to cosine similarity. In this example, $n = 8$ and $k = 5$ and the displayed graphs correspond to 5 co-clusters obtained by analyzing the PUBMED5 dataset. It can be noted that the top terms (light-blue nodes) in each cluster are well interconnected, reflecting a real semantic cohesion

Outline

Summary

- k-means is still relevant. He has inspired and continues to inspire several approaches
- k-means should be helped to detect relevant classes

Many hurdles

- High dimensionality, n«d, Sparsity
- Shapes of classes, their volume, direction and proportion
- Diversity of solutions

Some solutions

- Finite mixture models
- Combining kmeans with embedding is an interesting option for the user (Joint Approach)
- Ensemble method [a]
- Spectral clustering[b], Co-clustering, Subspace clustering etc.

[a] Boutalbi, R., Labiod, L., & Nadif, M. (2021). Implicit consensus clustering from multiple graphs. Data Mining and Knowledge Discovery, 35(6), 2313-2340.

[b] Allab, K., Labiod, L., & Nadif, M. (2018). Simultaneous spectral data embedding and clustering. IEEE TNNLS, 29(12), 6396-6401.