An efficient SEM algorithm for Gaussian Mixtures with missing data

V. Vandewalle[1] ,C. Biernacki[2]

[1] : University Lille 2, EA 2694 & Inria Lille, Modal team
[2] : University Lille 1, UMR CNRS 8524 & Inria Lille, Modal team

ERCIM 2015
London
December 12th, 2015

Outline

Clustering with missing data

X_1	X_2	X_3	Cluster
1.23	?	3.42	?
?	?	4.10	?
4.53	1.50	5.35	?
?	5.67	?	?

Usual solutions to deal with missing data

- Suppress units and/or variables with missing data $\Rightarrow$ loss of information
- Imputation of the missing data by the mean or more evolved methods $\Rightarrow$ uncertainty of the prediction not taken into account

Illustration of the risk of imputation in clustering

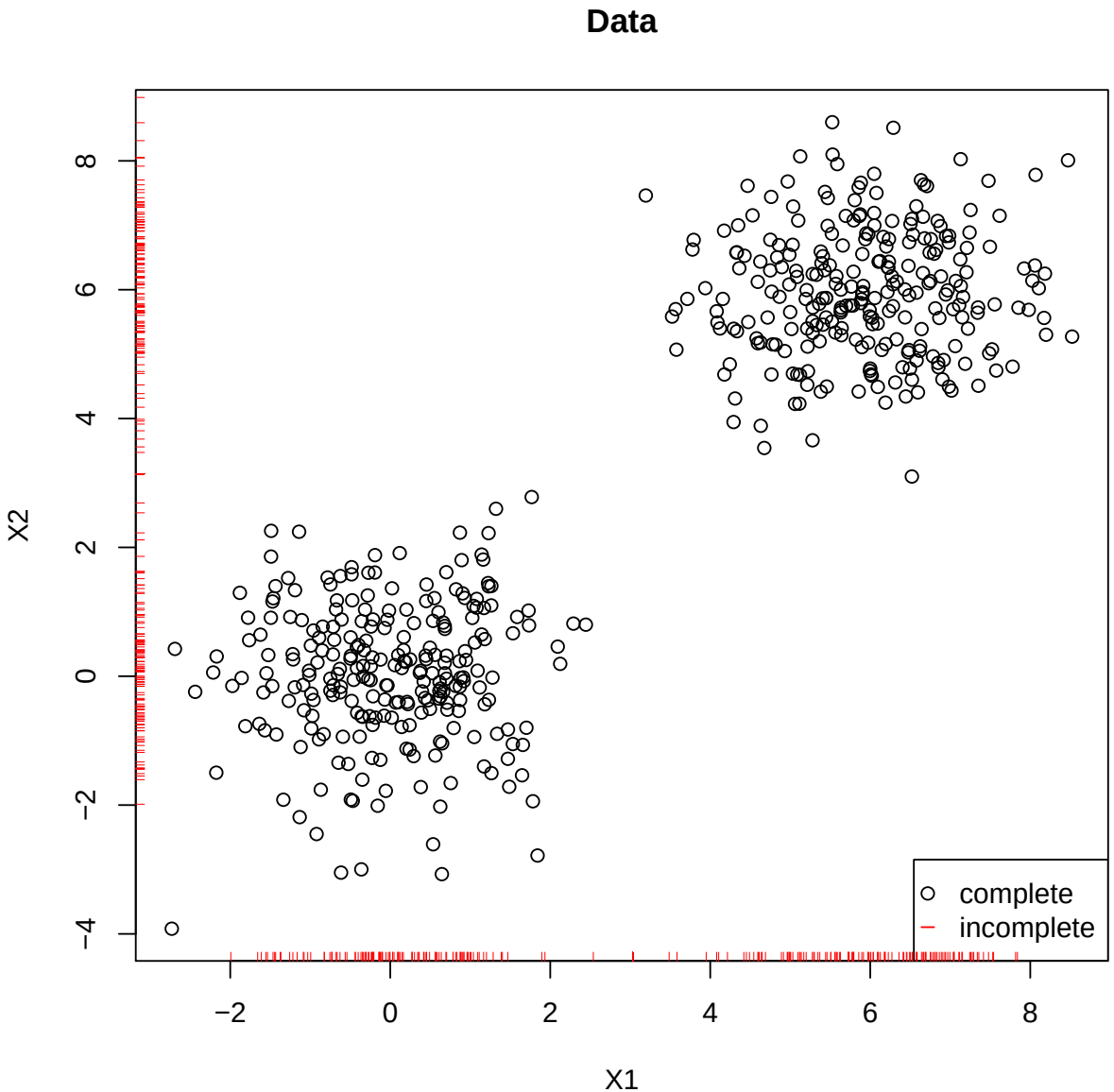


Illustration of the risk of imputation in clustering

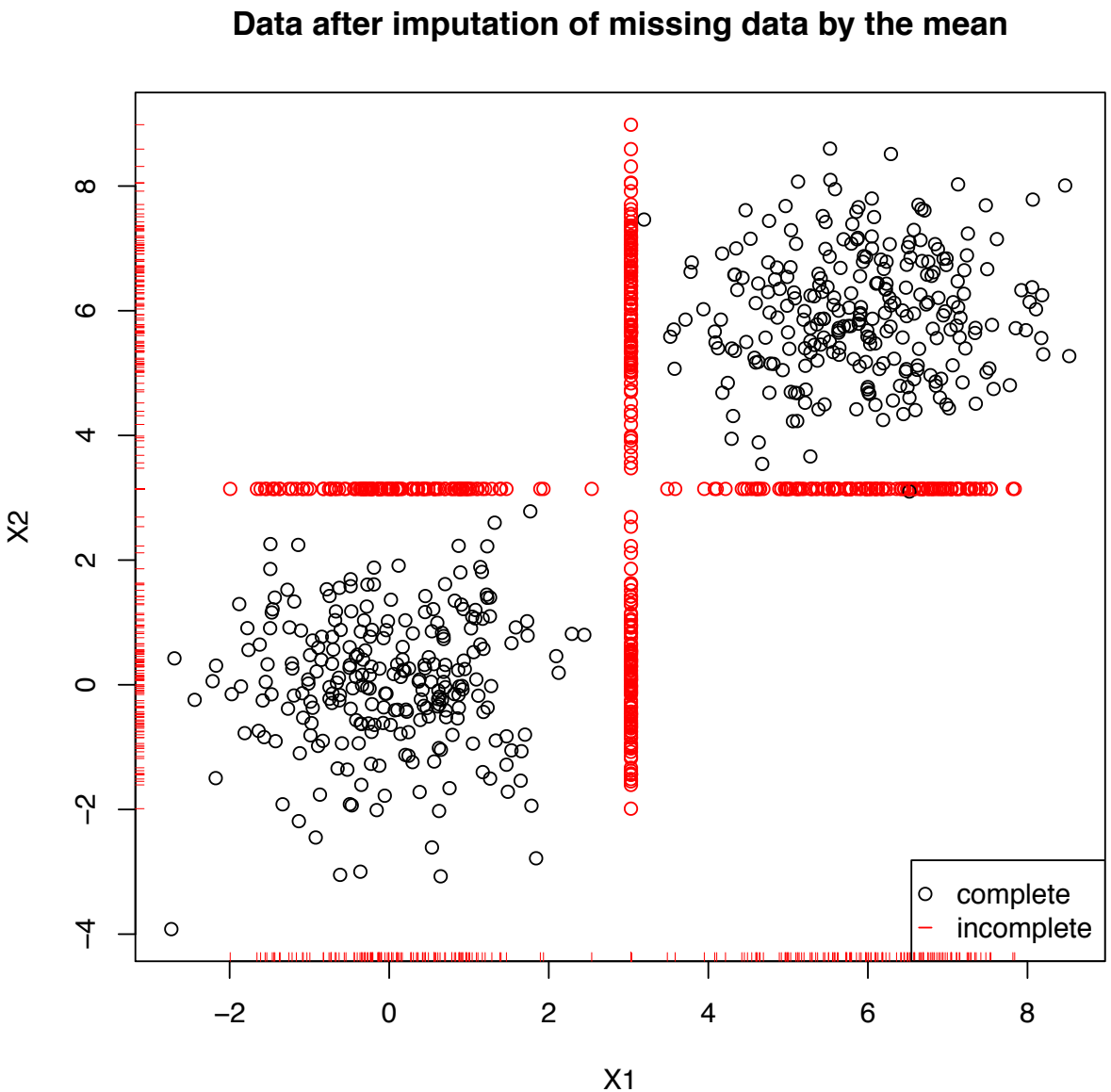


Illustration of the risk of imputation in clustering

An integrated clustering approach

Problem

Preliminary imputation step can lead to an over estimation of the number of clusters.

Solution

Use an integrated approach which allows to take into account all the available information to perform clustering $\Rightarrow$ use **mixture models**.

The two levels of missing data considered are

- the cluster
- the covariates

The **EM algorithm** allows to integrate these two levels of missing data.

Model and assumption

Model

- $\mathbf{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ the data, coming from a mixture of K Gaussian d-variates components
- $O_i \subseteq \{1, \ldots, d\}$ the set of the observed variables from sample i
- $\boldsymbol{x}_i^{\mathrm{O}}$ the observed data from sample i
- M_i the set of the missing variables for sample i
- $\boldsymbol{z}_i$ the binary coding of the cluster of sample i:
  $z_{ik} = \begin{cases} 1 \text{ if sample } i \text{ comes from class } k, \\ 0 \text{ otherwise.} \end{cases}$

Assumption on the missingness mecanism

Missing At Randon (MAR): the probability that a variable is missing does not depend on its own value given the observed variables.

Parameters

- $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$: the cluster proportions
- $\boldsymbol{\mu}_k$: vector of means for cluster k
- $\boldsymbol{\Sigma}_k$: variance-covariance matrix for cluster k
- $\boldsymbol{\lambda}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $\boldsymbol{\theta}$: the global parameter
- $\phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: the Gaussian probability density function
- $\boldsymbol{\mu}_{ik}^{\mathrm{O}}$ the sub-vector of $\boldsymbol{\mu}_k$ associated to index O_i (the same for M_i)
- $\boldsymbol{\Sigma}_{ik}^{\mathrm{OM}}$ the sub-matrix of $\boldsymbol{\Sigma}_k$ associated to row O_i and columns M_i (the same for any other combination)

Maximum likelihood estimator

Maximum likelihood estimator

Let $\ell(\boldsymbol{\theta}; \mathbf{x}^{\mathsf{O}})$ the log-likelihood:

$$\ell(\boldsymbol{\theta}; \mathbf{x}^{\mathsf{O}}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \phi(\boldsymbol{x}_i^{\mathsf{O}}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is defined by:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathsf{O}})$$

EM algorithm

- Starting from $\boldsymbol{\theta}^{(0)}$, the algorithm defines a sequence such that $\ell(\boldsymbol{\theta}^{(r+1)}; \mathbf{x}^{\mathsf{O}}) \geq \ell(\boldsymbol{\theta}^{(r)}; \mathbf{x}^{\mathsf{O}})$.
- It may converge to a root of $\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathsf{O}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$.

Maximum likelihood estimator and unbounded likelihood

Unbounded likelihood

If Σ_k is free then $\ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})$ is unbounded: $\boldsymbol{\mu}_k = \boldsymbol{x}_i$ and $|\Sigma_k| \to 0$.

$\Rightarrow \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})$

Consistent root

A root of $\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ is a consistent estimator of the parameters.

Maximum likelihood estimator and unbounded likelihood

Unbounded likelihood

If Σ_k is free then $\ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})$ is unbounded: $\boldsymbol{\mu}_k = \boldsymbol{x}_i$ and $|\Sigma_k| \to 0$.

$\Rightarrow \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})$

Consistent root

A root of $\frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ is a consistent estimator of the parameters.

New choice of the parameter estimator

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}}) \text{ s.c. } \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

Practical solution

Use the EM algorithm and discard solutions associated to unbounded likelihood.

E step: missing data

θ and θ^+ the parameters for two successive steps (*idem* for missing data)

$$z_{ik}^{+} = P(Z_{ik} = 1 | \boldsymbol{x}_i^{\mathrm{O}}; \boldsymbol{\theta}) = \frac{\pi_k \phi(\boldsymbol{x}_i^{\mathrm{O}}; \boldsymbol{\lambda}_k)}{\sum_{\ell=1}^{K} \pi_\ell \phi(\boldsymbol{x}_i^{\mathrm{O}}; \boldsymbol{\lambda}_\ell)}$$

$$\boldsymbol{x}_{ik}^{M^+} = E\left[\boldsymbol{X}_i^{M} \middle| \boldsymbol{x}_i^{\mathrm{O}}, Z_{ik} = 1; \boldsymbol{\theta}\right] = \boldsymbol{\mu}_{ik}^{\mathrm{M}} + \boldsymbol{\Sigma}_{ik}^{\mathrm{MO}} \left(\boldsymbol{\Sigma}_{ik}^{\mathrm{OO}}\right)^{-1} (\boldsymbol{x}_i^{\mathrm{O}} - \boldsymbol{\mu}_{ik}^{\mathrm{O}}).$$

Interpretation

- z_{ik}^{+}: class posterior probability membership given the available information $\boldsymbol{x}_i^{\mathrm{O}}$.
- $\boldsymbol{x}_{ik}^{\mathrm{M}^+}$: conditional imputation of the missing data given the cluster.

M step: parameter

$$\pi_k^+ = \frac{1}{n_k^+}\sum_{i=1}^{n} z_{ik}^+,\ \boldsymbol{\mu}_k^+ = \frac{1}{n_k^+}\sum_{i=1}^{n} z_{ik}^+ \boldsymbol{x}_{ik}^+$$

$$\boldsymbol{\Sigma}_k^+ = \frac{1}{n_k^+}\sum_{i=1}^{n} z_{ik}^+ \left[(\boldsymbol{x}_{ik}^+ - \boldsymbol{\mu}_k^+)(\boldsymbol{x}_{ik}^+ - \boldsymbol{\mu}_k^+)' + \boldsymbol{\Sigma}_{ik}^+\right]$$

where $n_k^+ = \sum_{i=1}^{n} z_{ik}^+$, $\boldsymbol{x}_{ik}^+ = \begin{pmatrix} \boldsymbol{x}_i^{\mathrm{O}} \\ \boldsymbol{x}_{ik}^{\mathrm{M}^+} \end{pmatrix}$, $\boldsymbol{\Sigma}_{ik}^+ = \begin{pmatrix} 0_i^{\mathrm{O}} & 0_i^{\mathrm{OM}} \\ 0_i^{\mathrm{MO}} & \boldsymbol{\Sigma}_{ik}^{\mathrm{M}^+} \end{pmatrix}$

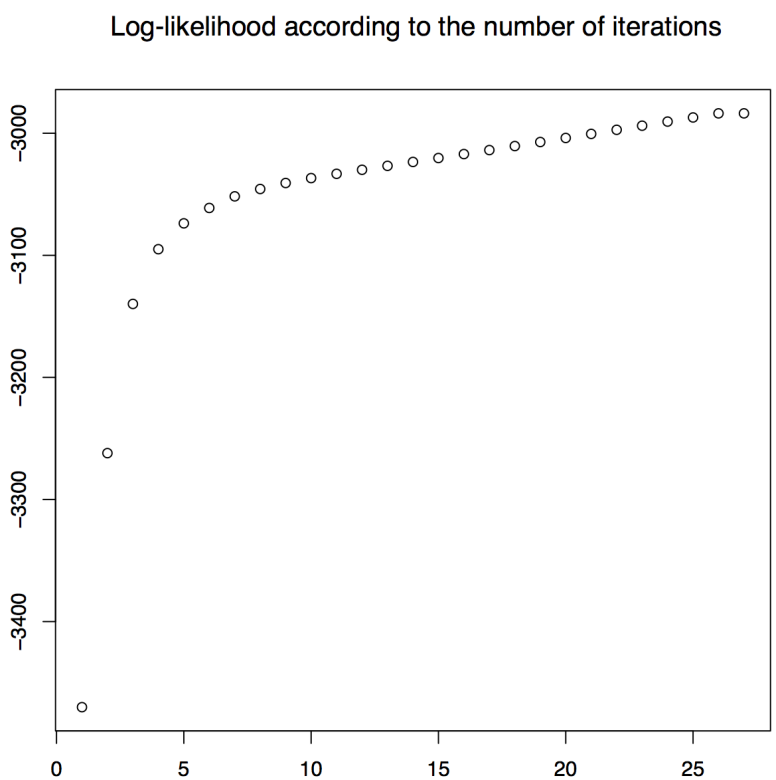
with 0 the $d \times d$ null matrix, and $\boldsymbol{\Sigma}_{ik}^{\mathrm{M}^+} = \boldsymbol{\Sigma}_{ik}^{\mathrm{MO}} \left(\boldsymbol{\Sigma}_{ik}^{\mathrm{O}}\right)^{-1} \boldsymbol{\Sigma}_{ik}^{\mathrm{OM}}$.

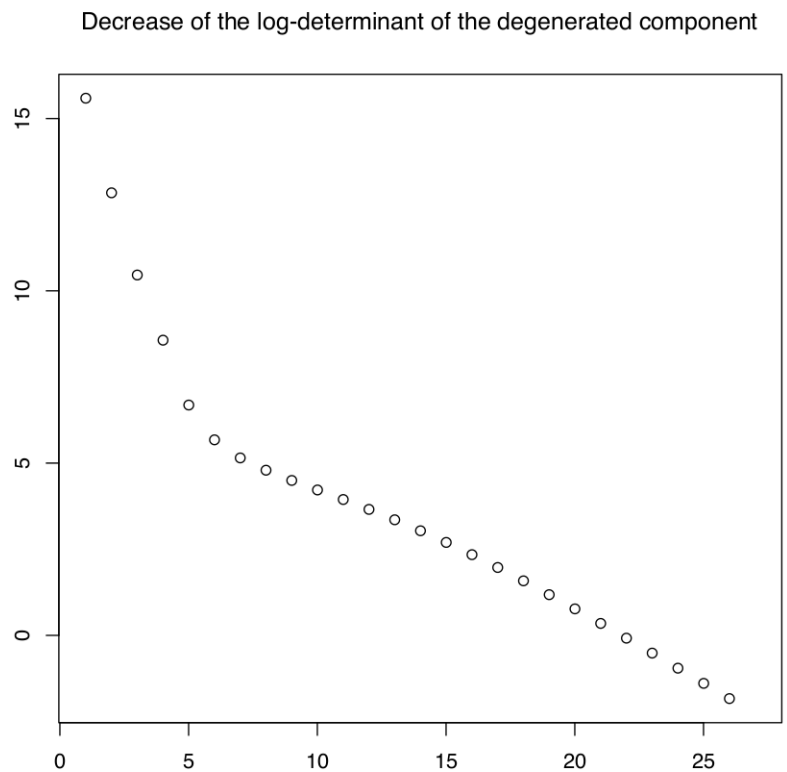
Interpretation of $\boldsymbol{\Sigma}_{ik}^{\mathrm{M}^+}$

Variance correction due to the under-estimation of variability caused by the imputation of missing data.

Example

- Breast cancer tissue of the UCI database repository : 106 units, 9 variables.
- 10% of missing data randomly generated
- $K = 4$ clusters





	1	2	3	4	5	6	7	8	9
1	211.00		0.09	30.75	151.98	4.94	14.27	27.24	217.13
2	196.86	0.02	0.09	28.59	82.06	2.87	7.97	27.66	200.75
3	144.00	0.12	0.05	19.65	70.43	3.58		7.57	160.37
4	172.52	0.13	0.04		192.22	5.12	19.32	32.19	174.93
5	121.00	0.17	0.09	24.44	144.47	5.91	22.02	10.59	141.77
6	223.00	0.12	0.08	33.10	197.01	5.95	30.45	12.96	252.48
7		0.17	0.23	34.22	94.35	2.76	31.28	13.88	180.61
8	303.00	0.06	0.04	22.57		4.54	21.83	5.72	321.65
9	250.00	0.09	0.09	29.64	180.76	6.10	26.14	13.96	280.12
10	391.00	0.06	0.01	35.78		7.41	22.13	28.11	400.99
11	176.00	0.09	0.08	20.59	79.71		18.23	9.58	191.99
12	145.00		0.11	21.22	82.46	3.89	20.30	6.17	162.51
13	124.13	0.13	0.11	20.59			18.46	9.12	134.89
14	103.00	0.16	0.29	23.75	78.26	3.29	22.32	8.12	124.98

Table: Data belonging to the degenerated component.

Remarks

- Convergence towards a degenerated component
- Convergence relatively slow : log-likelihood linear according to the number of iterations
- Number of points of the degenerated solution greater than the space dimension d (but the number of complete points lower than d)

Risks

- Consider a degenerated solution as valid
- Lose a lot of time in useless iterations

Bibliography on degeneracy

Complete data (Biernacki & Chrétien (2003), Ingrassia & Rocci (2007))

- Unbounded likelihood
- Very fast degeneracy : $\sigma^{2^+} \leq \alpha \frac{\exp(-\frac{\beta}{\sigma^2})}{\sigma^2}, \alpha, \beta \in \mathbb{R}^{+*}$
- Practical solution: restart the EM algorithm with another starting point

Bibliography on degeneracy

Complete data (Biernacki & Chrétien (2003), Ingrassia & Rocci (2007))

- Unbounded likelihood
- Very fast degeneracy : $\sigma^{2^+} \leq \alpha \frac{\exp(-\frac{\beta}{\sigma^2})}{\sigma^2}, \alpha, \beta \in \mathbb{R}^{+*}$
- Practical solution: restart the EM algorithm with another starting point

Grouped data (Biernacki (2007))

- Bounded likelihood
- Very slow degeneracy: $\sigma^{2^+} = \sigma^2 - \gamma u, \gamma \in \mathbb{R}^+, u \in v(0^+)$

Degeneracy with missing data

Missing data: an intermediary framework between complete and grouped data

- Unbounded likelihood like complete data
- Slow degeneracy like grouped data

Degeneracy speed on a toy example

Univariate framework, no mixture, only one observed data: x

- Maximum likelihood estimator:
  - $\hat{\mu} = x$
  - $\hat{\Sigma} = 0$
- Unbounded likelihood

Degeneracy speed on a toy example

Univariate framework, no mixture, only one observed data: x

- Maximum likelihood estimator:
  - $\hat{\mu} = x$
  - $\hat{\Sigma} = 0$
- Unbounded likelihood

Suppose now that $n - 1$ data have not been observed:

Useless EM algorithm

$$\mu^+ = \frac{(n-1)\mu + x}{n} \text{ et } \Sigma^+ = \frac{(n-1)\Sigma + (x - \mu^+)^2}{n}.$$

Degeneracy speed on a toy example

Univariate framework, no mixture, only one observed data: x

- Maximum likelihood estimator:
  - $\hat{\mu} = x$
  - $\hat{\Sigma} = 0$
- Unbounded likelihood

Suppose now that $n-1$ data have not been observed:

Useless EM algorithm

$$\mu^+ = \frac{(n-1)\mu + x}{n} \text{ et } \Sigma^+ = \frac{(n-1)\Sigma + (x-\mu^+)^2}{n}.$$

This lead to a linear grow of the log-likelihood :

$$L(\theta^{(r)}; x) \sim -0,5r \log \frac{n-1}{n}.$$

Influence of the missing data rate

% missing data	0	5	10	15	20	25	30
% deg.	16	4	12	11	46	51	100
% deg. given deg. for lower rate	-	25	100	91	100	100	100
Average number of iterations before deg.	2	13	13	82	304	138	215

Table: Frequency and speed of degeneracy (deg.) according to the rate of missing data on the breast cancer data set.

When the rate of missing data increases:

- The rate of degeneracy increases
- The number of iteration before degeneracy decreases

A constraint of the minimum number of points per cluster

Key idea

If each cluster contains at least $d+1$ complete data points, then no degeneracy with probability 1:

- $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ a partition of $\mathbf{x}$
- $n_k = \sum_{i=1}^{n} z_{ik}$
- $\mathcal{Z}^* = \{\mathbf{z} : n_k \geq d+1\}$

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \underbrace{\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z} \in \mathcal{Z}^*)}_{< \infty \text{ with proba. } 1} + \underbrace{\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z} \notin \mathcal{Z}^*)}_{\text{can degenerate}}$$

Discarding some $\mathbf{z}$ values to avoid degeneracy

Solution

Impose $\mathbf{z} \in \mathcal{Z}^*$ in the EM algorithm. The new parameter estimator becomes:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta; \mathbf{x}, \mathbf{z} \in \mathcal{Z}^*)$$

Remarks

- $\mathbf{z} \in \mathcal{Z}^*$ natural in the supervised setting to obtain non-singular covariance matrices
- $\hat{\theta}$ approaches the maximum likelihood estimator as the number of data increases

Specific EM algorithm

Algorithm

- E step: $\tilde{z}_{ik}^{+} \propto p(\mathbf{Z} \in \mathcal{Z}^* | \mathbf{x}, Z_{ik} = 1; \theta) \overbrace{p(Z_{ik} = 1 | \mathbf{x}; \theta)}^{z_{ik}^{+}}$
- M step: use the standard M step using $\tilde{z}_{ik}^{+}$ instead of z_{ik}^{+}

Problem

$$p(\mathbf{Z} \in \mathcal{Z}^* | \mathbf{x}, Z_{ik} = 1; \theta) = \sum_{\mathbf{z} \in \mathcal{Z}^*} p(\mathbf{Z} = \mathbf{z} | \mathbf{x}, Z_{ik} = 1; \theta)$$

involves the Stirling number of the second king $\Rightarrow$ tractable only for $K = 2$.

This algorithm will be later be called **EMgood**.

Comparison of EM and EMgood

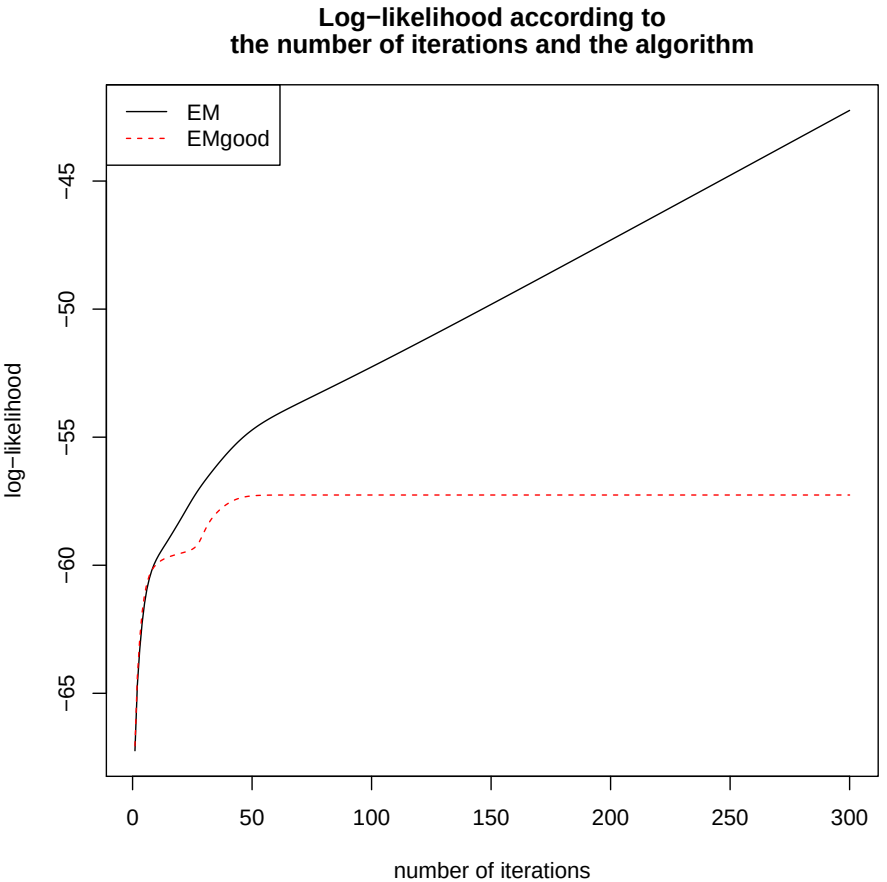
$\pi_1 = \pi_2 = 0.5$

$$\boldsymbol{X}_i | Z_{i1} = 1 \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}; \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

$$\boldsymbol{X}_i | Z_{i2} = 1 \sim \mathcal{N}\left(\begin{pmatrix}2\\2\end{pmatrix}; \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

$n = 30$ data, $p = 80\%$ of missing data.
Results on 100 simulations, 300 iterations, 10 starting values.

Algorithm	Adjusted Rand Index
EM	0.171 (0.015)
EMgood	0.200 (0.015)



A tractable SEM constrained algorithm

Stochastic EM

The SEM algorithm introduces a stochastic step between the E and the M step of the EM algorithm:

- S step : $\mathbf{z}^+ \sim \mathbf{Z}|\mathbf{x};\theta$

Partition constraints easy to include in the S step

Add the the constraint that $\mathbf{Z} \in \mathcal{Z}^*$ in the S step by various methods:

- Rejection sampling
- Gibbs sampling
- ...

This algorithm will be later be called **SEMc**.

Performances of the SEMc

Estimated parameter

- The SEMc produces a sequence $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$
- $\hat{\boldsymbol{\theta}}^{\text{SEMc}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}} \in \ell(\boldsymbol{\theta}; \mathbf{x})$

Comparison of EM and SEMc

- Start the algorithm from 10 random values, for each initialization iterate 300 times
- Keep the parameter associated to the best likelihood
- Compute the rand index between the estimated and the true partition

Example on the breast cancer tissu data set

Dataset

- Dataset: Breast cancer tissue of the UCI database repository : $n = 106$, $d = 9$.
- Draw 5% missing data completely at random
- Try to find the 6 clusters in the data

Results

- EM degenerates for each initialisation $\Rightarrow$ no performances available
- SEMc never degenerates, the solution with the higher likelihood has an adjusted rand index of 0.30

Comparison on simulated data

$\pi_1 = \pi_2 = 0.5$

$$\boldsymbol{X}_i | Z_{i1} = 1 \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}; \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

$$\boldsymbol{X}_i | Z_{i2} = 1 \sim \mathcal{N}\left(\begin{pmatrix}2\\2\end{pmatrix}; \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

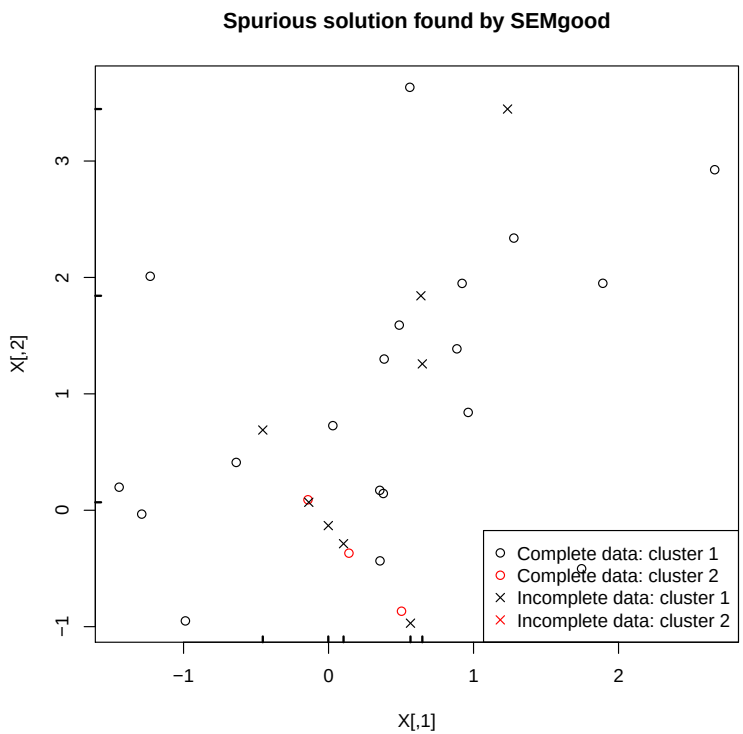
$n = 50$ data, $p = 10\%$ of missing data. Results on 100 simulations, 10 starting values, 300 iterations by starting value.

Algorithm	EM	SEMc
ARI	0.217	0.067
#best $\ell(\boldsymbol{\theta}; \mathbf{x})$	24	76

SEMgood and spurious maxima

Problem

- SEMgood efficient in finding local maxima of $\ell(\boldsymbol{\theta}; \mathbf{x})$
- But maximum likelihood can be jeopardized by spurious local maxima



Solutions

- Impose constraints on variance ratio (Hataway, 1983) $\Rightarrow$ constraints difficult to tune
- Modify $\mathcal{Z}^*$ to discard spurious solutions with high probability

Conclusion et perspectives

Degeneracy with missing data

- Degeneracy relatively slow with missing data
- The degenerate rate increases with the number of missing data

Partition constrained algorithms

- Solution to avoid degeneracy
- Can sometime be trapped by spurious maxima

Perspectives

- Limit the spurious problem by modifying $\mathcal{Z}^*$