Pitfalls in Mixtures from the Clustering Angle

C. Biernacki
(with G. Castellan, S. Chrétien, B. Guedj, V. Vandewalle)

Working Group on Model-Based Clustering Summer Session, Paris, July 17-23, 2016

Take home message

Computational estimates $\tilde{\theta}$ are the imbricated result of five factors

1. An initial practitioner target t
2. A data set $\mathbf{x}$
3. A theoretical model m
4. A theoretical estimate $\hat{\theta}$
5. An estimation algorithm A

$$\tilde{\theta} = f(t, \mathbf{x}, m, \hat{\theta}, A)$$

This talk

- Considered pitfalls in mixtures are degeneracy and label switching
- Consequences can be disastrous on $\tilde{\theta}$
- Often, solutions are sought in m or $\hat{\theta}$
- We explore here also solutions through t and A

- Focus target t : clustering
- Focus algorithms A : EM, SEM, Gibbs

Outline

Unbounded likelihood

- d-variate g-Gaussian mixture with $\boldsymbol{\theta} = (\{\pi_k\}, \{\boldsymbol{\mu}_k\}, \{\Sigma_k\})$

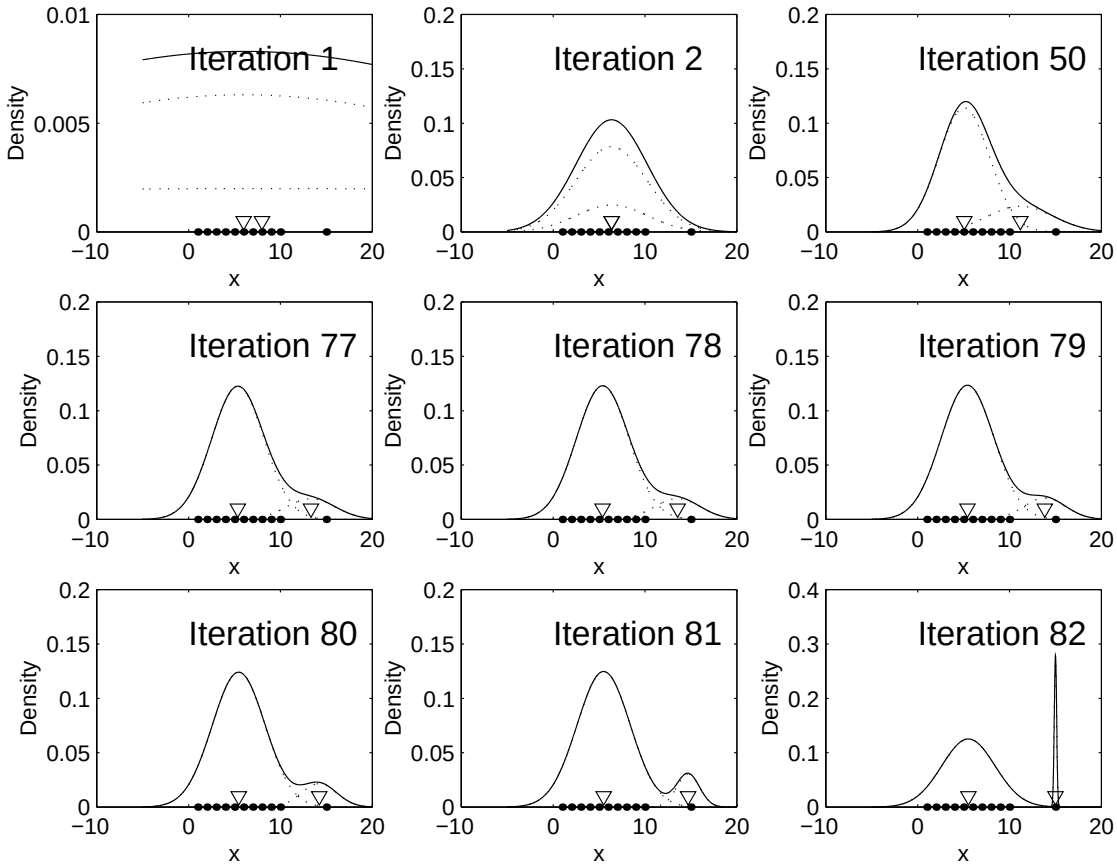
$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \underbrace{\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\Big)}_{p(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)}$$

- Sampling: $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \overset{i.i.d.}{\sim} p(.; \boldsymbol{\theta})$
- Likelihood: $\ell(\boldsymbol{\theta}; \mathbf{x}) = p(\mathbf{x}; \boldsymbol{\theta})$

$$\text{particular center } \boldsymbol{\mu}_2 = \boldsymbol{x}_i \quad \Rightarrow \quad \lim_{|\Sigma_2| \to 0} \ell(\boldsymbol{\theta}; \mathbf{x}) = +\infty$$

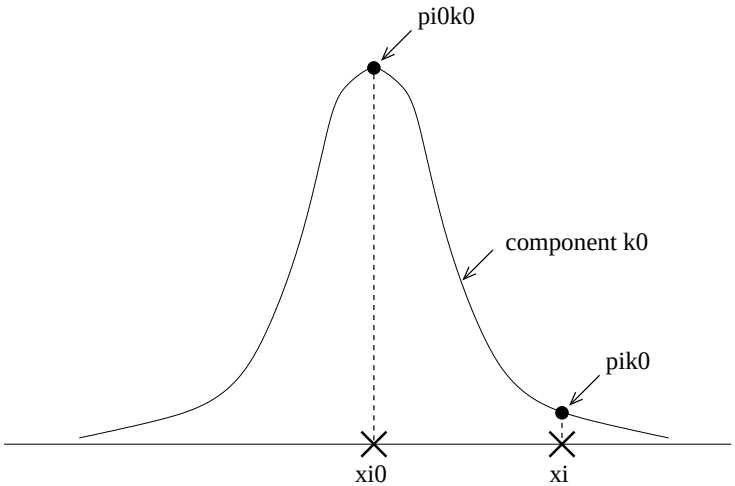
[Kiefer and Wolfowitz, 1956] [Day, 1969]

EM behaviour: illustration



- degeneracy may occur even when starting from large variances
- convergence can be slow when far from the degenerate limit
- convergence extremely fast near degeneracy

EM behaviour: results



$$\mathbf{u}_0 = \left[\frac{1}{p_{i_0 k_0}}, \{p_{ik_0}\}_{i \neq i_0} \right]$$

degeneracy of component k_0 at $\boldsymbol{x}_{i_0}$

$\Leftrightarrow$

$$\|\mathbf{u}_0\| \to 0$$

[Biernacki and Chrétien, 2003]

[Ingrassia and Rocci, 2009]

Proposition 1: Existence of a bassin of attraction

$\exists \epsilon > 0$ *s.t. if* $\|\mathbf{u}_0\| \leq \epsilon$ *then* $\|\mathbf{u}_0^+\| = o\|\mathbf{u}_0\|$ *with probability 1.*

Proposition 2: Speed towards degeneracy is exponential

$\exists \epsilon > 0$, $\alpha > 0$ *and* $\beta > 0$ *s.t. if* $\|\mathbf{u}_0\| \leq \epsilon$ *then, with probability 1,*

$$|\Sigma_{k_0}^+| \leq \alpha / |\Sigma_{k_0}| \cdot \exp\Big(- \beta / |\Sigma_{k_0}| \Big).$$

Consequences of the EM study

When EM is close to degeneracy, EM mapping is contracting and reaches numerical tolerance extremely quickly

$\Downarrow$

Simply starting again EM when numerical tolerance is reached
(pragmatic bahaviour of EM practitioners)
is now somewhat justified

$\Downarrow$

However, the numerical tolerance is finally
an arbitrary lower bound for $|\Sigma_k|\ldots$

Outline

Binned data

- A binned partition of $\mathbb{R}$ in H intervals $\Omega_1, \ldots, \Omega_H$: $\Omega_h =]\alpha_h, \beta_h[$
- Individuals x_i unknown, only the interval where x_i lies is known
- Hypothesis of Gaussian mixture on x_i's unchanged
- The log-likelihood is written

$$\ell(\theta) = \sum_{h=1}^{H} \underbrace{m_h}_{\#\,\Omega_h} \ln \underbrace{\left(\sum_{k=1}^{K} \pi_k \overbrace{\int_{\Omega_h} f_k(x)dx}^{a_{kh}} \right)}_{p(X \in \Omega_h)}$$

Question

Does degeneracy still exists since $\ell(\theta) \leq 0$?

Degeneracy may still happen!

Proposition 3

Let for all $b \in \mathbb{N}$

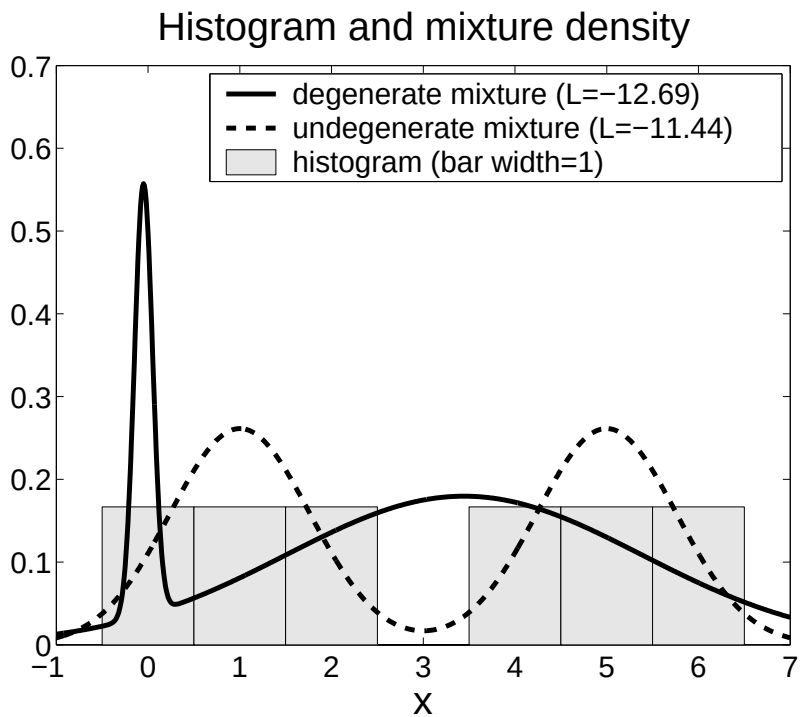
- *sequence* $\{\epsilon^b\}$*:* $\epsilon^b > 0$ *and* $\epsilon^b \to 0$ *when* $b \to \infty$
- *bins* $\left\{\Omega_h^b, h = 1, \ldots, H^b\right\}$*: if* $\beta_h^b - \alpha_h^b \geq \epsilon^b$ *then* $m_h^b = 0$
- $\Omega_{h_0^b}$ *is a non-empty interval and* $k_0 \in \{1, \ldots, K\}$ *a component*
- $\hat{\theta}^b$ *is the unique consistent root of the ML associated to* $\left\{(\Omega_h^b, m_h^b)\right\}$
- $\ell^b(\theta) \longrightarrow \ell^b_{deg}(\theta)$ *when* $\mu_{k_0} \in \Omega_{h_0}$ *et* $\Sigma_{k_0} \to 0$.

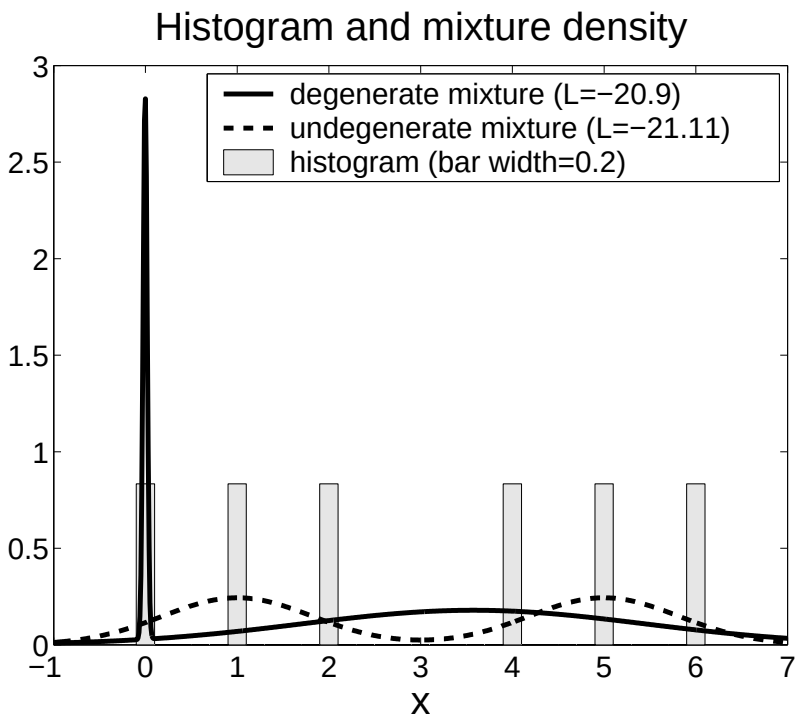
Thus, it exists $B \in \mathbb{N}$ *such that for all* $b > B$ *we have* $\ell^b_{deg}(\hat{\theta}^b) \geq \ell^b(\hat{\theta}^b)$.

Sketch of proof At a first time, we have to show that, for all θ, it exists $B_\theta \in \mathbb{N}$ such that for all $b > B_\theta$ we have $\ell^b_{deg}(\theta) \geq \ell^b(\theta)$.
Then, we conclude by noting that $B = \sup_\theta B_\theta$.

Meaning

If dimension of non-empty bins is "small enough", then the global maximum of the likelihood is obtained in a degenerate situation





EM behaviour in a degeneracy neighborough?

Remind

component k_0 degenerates inside $\Omega_{h_0} \Leftrightarrow \left(\mu_{k_0} \in \Omega_{h_0} \text{ and } \Sigma_{k_0} \to 0\right)$

Notations

- $\Omega_{h_0'}$: bin the closest to the center μ_{k_0} (left or right of Ω_{h_0})
- γ: borderline of Ω_{h_0} the closest to μ_{k_0} (either α_{h_0}, or β_{h_0})
- $\eta = |\gamma - \mu_{k_0}|$: distance between the center and the closest center
- $\sigma = \text{sign}(\gamma - \mu_{k_0})$ and $u = \Sigma_{k_0} f_{k_0}(\gamma)$
- $R_h = (\pi_{k_0} + A_{k_0 h_0})/A_{k_0 h}$ with $A_{k_0 h} = \sum_{k \neq k_0} \pi_k a_{kh}$

Possibility to be attracted around degeneracy

Proposition 4

It exists $\epsilon > 0$ *such that, if*

- $0 < \Sigma_{k_0} < \epsilon$
- $\eta \in (\delta, \Delta - \sqrt{\Sigma_{k_0}})$ *with* $0 < \delta < \Delta < (\beta_{h_0} - \alpha_{h_0})/2$
- $1 - \frac{m_{h_0'}}{m_{h_0}} R_{h_0'} > 0$

then,

$$0 < \Sigma_{k_0}^+ < \Sigma_{k_0} \left[1 - \underbrace{\left(1 - \frac{m_{h_0'}}{m_{h_0}} R_{h_0'} \right)}_{\rho} \frac{\delta}{2\sqrt{2\pi\Sigma_{k_0}}} e^{-\Delta^2/(2\Sigma_{k_0})} \right]$$

and

$$\eta^+ \in \left(\delta, \Delta - \sqrt{\Sigma_{k_0}^+} \right).$$

sketch of proof It relies on Taylor expansions around $\Sigma_{k_0} = 0$ with $\mu_{k_0} \in \Omega_{h_0}$

$$\mu_{k_0}^+ = \mu_{k_0} - \sigma\rho u + o(u) \quad \text{and} \quad \Sigma_{k_0}^+ = \Sigma_{k_0} - \eta\rho u + o(u).$$

Then the inequality on Σ_{k_0} arises easily.
For the second expression, we obtain in the same manner (for Σ_{k_0} "small enough")

$$\delta < |\gamma - \mu_{k_0}^+| < \Delta - \sqrt{\Sigma_{k_0}^+}.$$

Thus $|\gamma - \mu_{k_0}^+| < \Delta < (\beta_{h_0} - \alpha_{h_0})/2$ and so $\gamma^+ = \gamma$ (the closest borderline is kept unchanged). Since $\eta^+ = |\gamma - \mu_{k_0}^+|$, conclusion follows.

Attraction or repulsion?

Around a degenerate solution, EM runs closer or further depending on the sign of ρ which itself depends on the sample size of the "closest" bin.

Attraction: $\rho > 0$

from the theorem, if Σ_{k_0} is "close enough" to 0 and $\mu_{k_0} \in \Omega_{h_0}$ then

$$\underbrace{0 < \Sigma_{k_0}^+ < \Sigma_{k_0}\left[1 - \rho \times |\text{fcte}(\theta)|\right]}_{\Sigma_{k_0} \text{ decreases}} \quad \text{and} \quad \mu_{k_0}^+ \in \Omega_{h_0}$$

Repulsion: $\rho < 0$

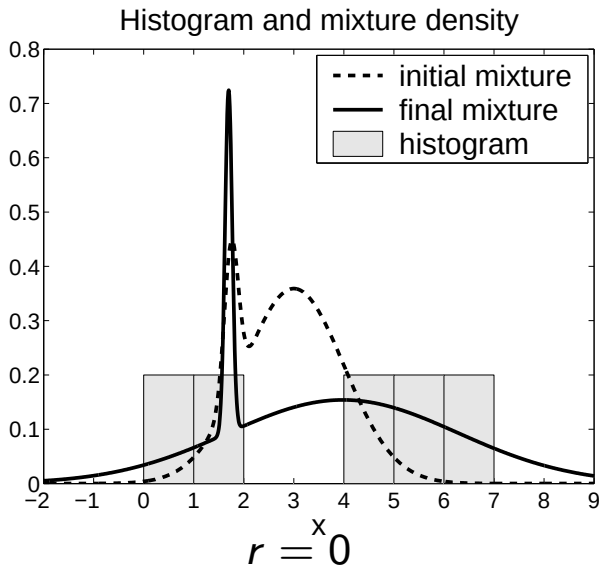
$$\text{Taylor: } \Sigma_{k_0}^+ = \Sigma_{k_0} - \eta\rho u + o(u) \quad \Rightarrow \quad \Sigma_{k_0} \text{ increases}$$

The sign of ρ if mainly controlled by the ratio of sample sizes

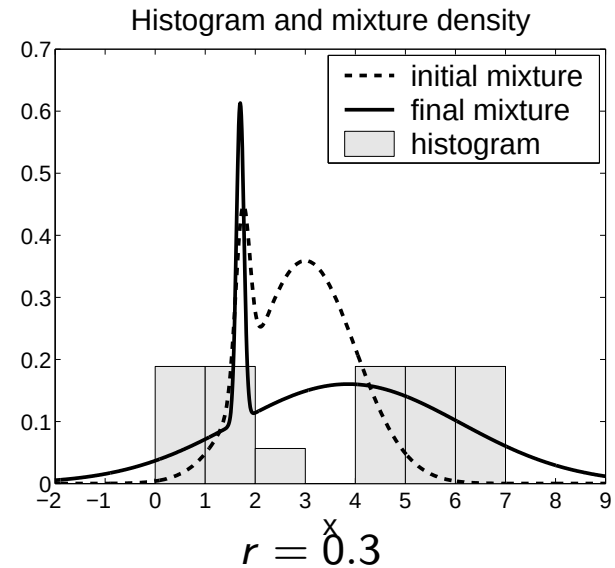
$$r = \frac{m_{h_0'}}{m_{h_0}} = \frac{\text{sample size of the closest bin}}{\text{sample size of the bin where degeneracy occurs}}$$

- r "small" favors $\rho > 0$
- r "large" favors $\rho < 0$
- $r = 0$: convergence of EM towards degeneracy established

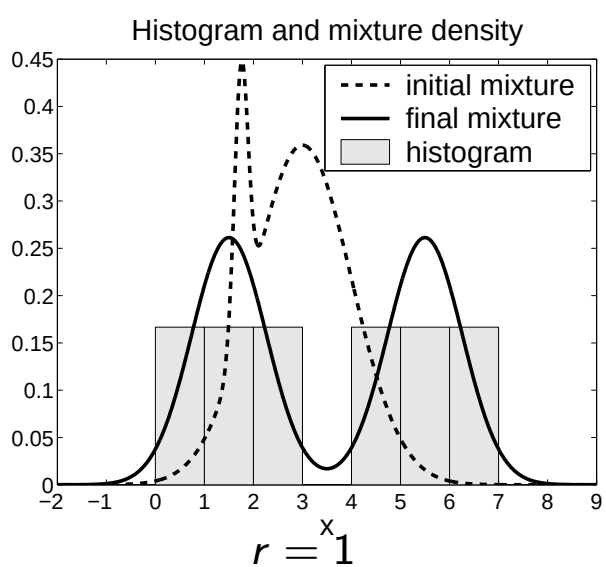
$$\Omega_{h_0} = (1\ 2) \quad \text{and} \quad \Omega_{h_0'} = (2\ 3)$$



$r = 0$



$r = 0.3$

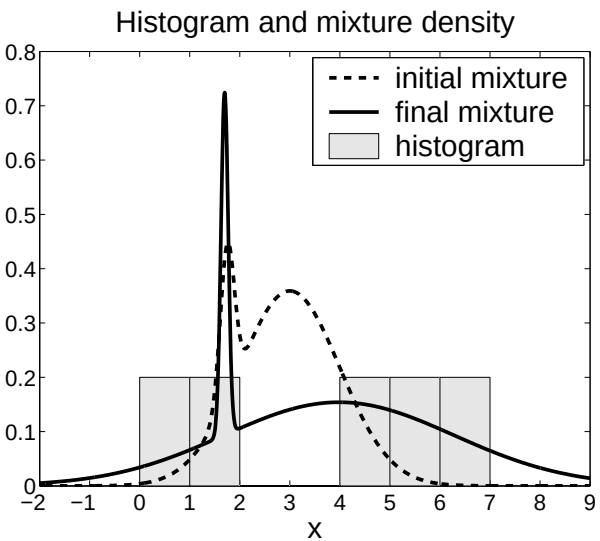


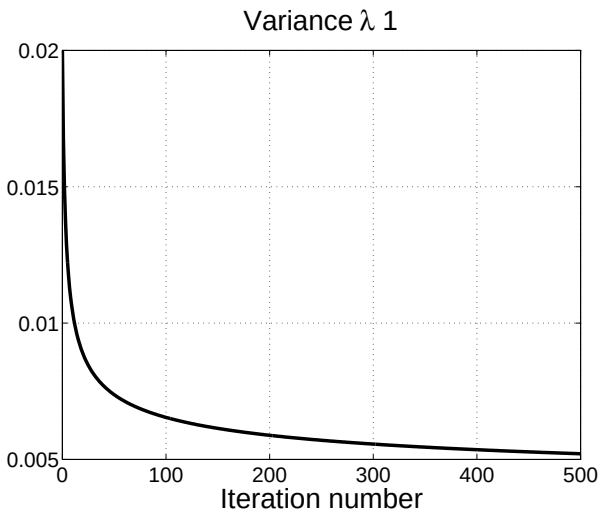
$r = 1$

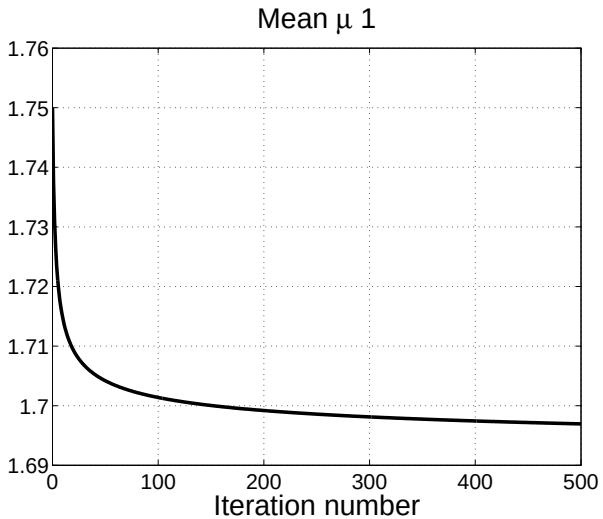
EM speed

EM is very slow around degeneracy because its global convergence rate is equal to 1

$$\Sigma_{k_0}^{+}/\Sigma_{k_0} \longrightarrow 1 \quad \text{when } \mu_{k_0} \in \Omega_{h_0} \text{ et } \Sigma_{k_0} \to 0$$







A stopping rule is required for EM!

- Danger: the ML could correspond to a degenerate solution
- Save computation time: numerous wasted iterations when $\rho > 0$
- Still running: run other iterations when $\rho < 0$

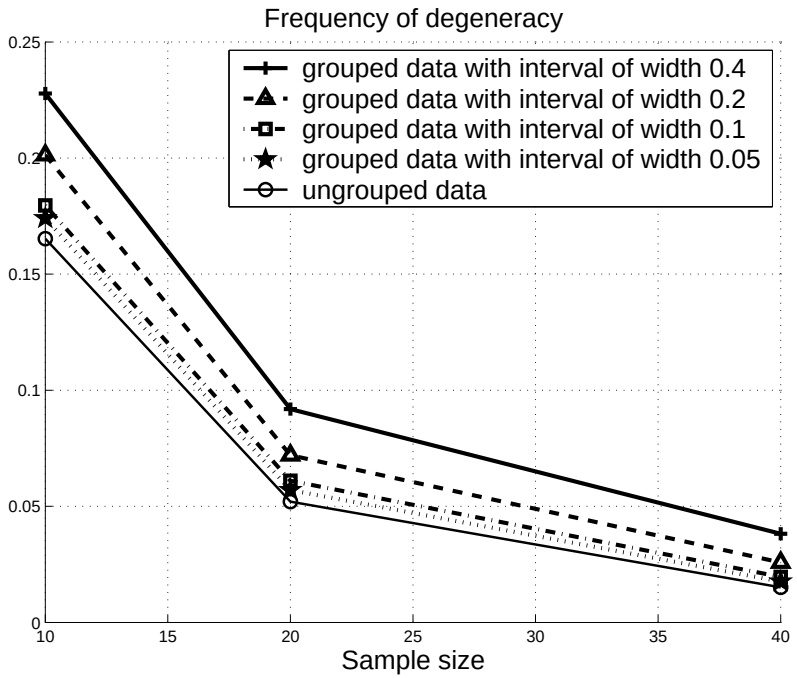
Stopping rules to be avoided

- $|\Sigma_{k_0}^{+} - \Sigma_{k_0}| < \epsilon$: confusion with convergence
- $\Sigma_{k_0} < \epsilon$: huge iteration number

Stopping rule relying on Taylor

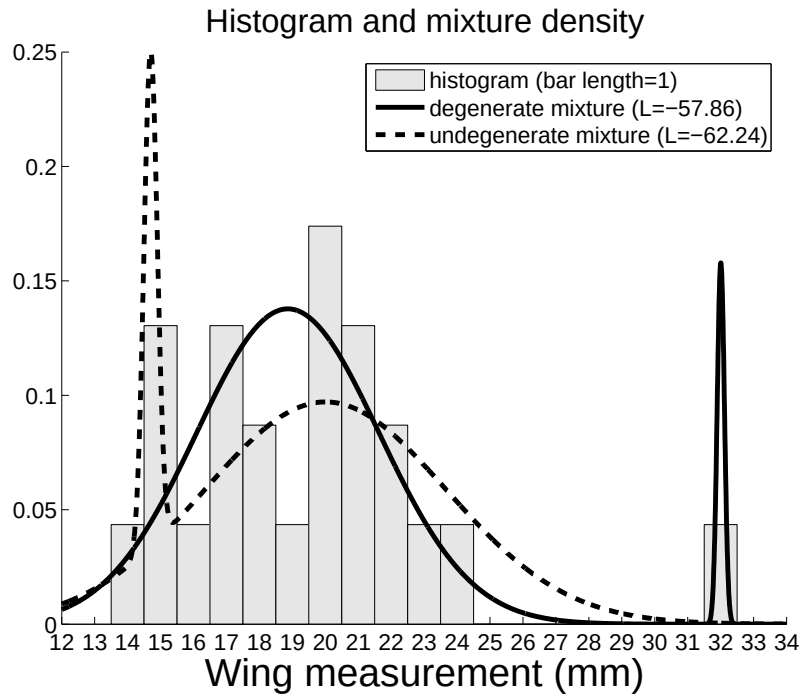
$$|\Sigma_{k_0}^{+} - \Sigma_{k_0} + \eta\rho u| < \epsilon$$

Numerical experiment 1: simulations



- $\rho < 0$ rare
- degeneracy $\rho > 0$ $\nearrow$ with bin width and $\searrow$ with n
- degeneracy binned case more frequent that the individual data case!

Numerical experiment 2: wing measurements of butterflies



- data known with 1mm precision: natural bins
- better likelihood at degeneracy
- the user could make a confusion between degeneracy and convergence
- the second variance has no meaning: DANGER

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	?	?

Discarded solutions

- 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

Retained solution

Use an integrated approach which allows to take into account all the available information to perform clustering

Notations

- $O_i \subseteq \{1, \ldots, d\}$ 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
- $\boldsymbol{\mu}_{ik}^{\mathrm{O}}$ 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)

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.

Maximum likelihood estimator

Unbounded likelihood...

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

$\boldsymbol{\mu}_k = \boldsymbol{x}_i$ and $|\boldsymbol{\Sigma}_k| \to 0 \Rightarrow \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})$ unbounded $\Rightarrow \hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}^{\mathrm{O}})$

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. So choose

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

Practical solution

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

E step

$\boldsymbol{\theta}$ and $\boldsymbol{\theta}^+$ the parameters for two successive steps (*idem* for missing data)

$$
\begin{aligned}
z_{ik}^{+} &= P(Z_{ik} = 1 | \boldsymbol{x}_i^{\mathrm{O}}; \boldsymbol{\theta}) = \frac{\pi_k \phi(\boldsymbol{x}_i^{\mathrm{O}}; \boldsymbol{\Sigma}_k)}{\sum_{\ell=1}^{K} \pi_\ell \phi(\boldsymbol{x}_i^{\mathrm{O}}; \boldsymbol{\Sigma}_\ell)} \\
\boldsymbol{x}_{ik}^{M^+} &= E\left[\boldsymbol{X}_i^{M} \middle| \boldsymbol{x}_i^{\mathrm{O}}, Z_{ik} = 1; \boldsymbol{\theta}\right] = \boldsymbol{\mu}_{ik}^{\mathrm{M}} + \boldsymbol{\Sigma}_{ik}^{\mathrm{MO}} \left(\boldsymbol{\Sigma}_{ik}^{\mathrm{OO}}\right)^{-1} (\boldsymbol{x}_i^{\mathrm{O}} - \boldsymbol{\mu}_{ik}^{\mathrm{O}}).
\end{aligned}
$$

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

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

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

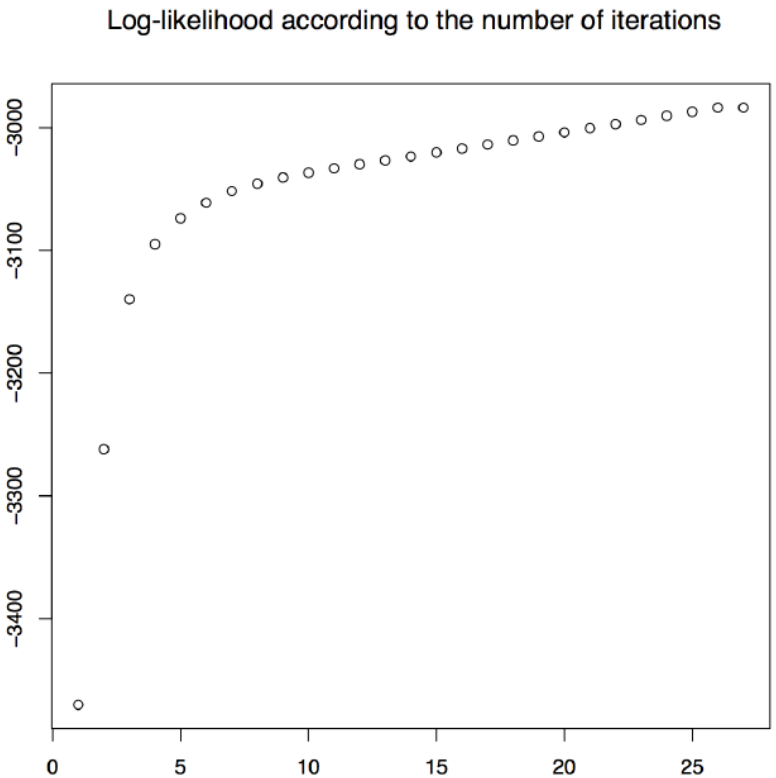
where $n_k^+ = \sum_{i=1}^{n} z_{ik}^+$, $\boldsymbol{x}_{ik}^+ = \begin{pmatrix} \boldsymbol{x}_i^{\mathrm{O}} \\ \boldsymbol{x}_{ik}^{\mathrm{M}^+} \end{pmatrix}$, $\boldsymbol{\Sigma}_{ik}^+ = \begin{pmatrix} 0_i^{\mathrm{O}} & 0_i^{\mathrm{OM}} \\ 0_i^{\mathrm{MO}} & \boldsymbol{\Sigma}_{ik}^{\mathrm{M}^+} \end{pmatrix}$ with 0 the $d \times d$ null matrix, and $\boldsymbol{\Sigma}_{ik}^{\mathrm{M}^+} = \boldsymbol{\Sigma}_{ik}^{\mathrm{MO}} \left(\boldsymbol{\Sigma}_{ik}^{\mathrm{O}}\right)^{-1} \boldsymbol{\Sigma}_{ik}^{\mathrm{OM}}$.

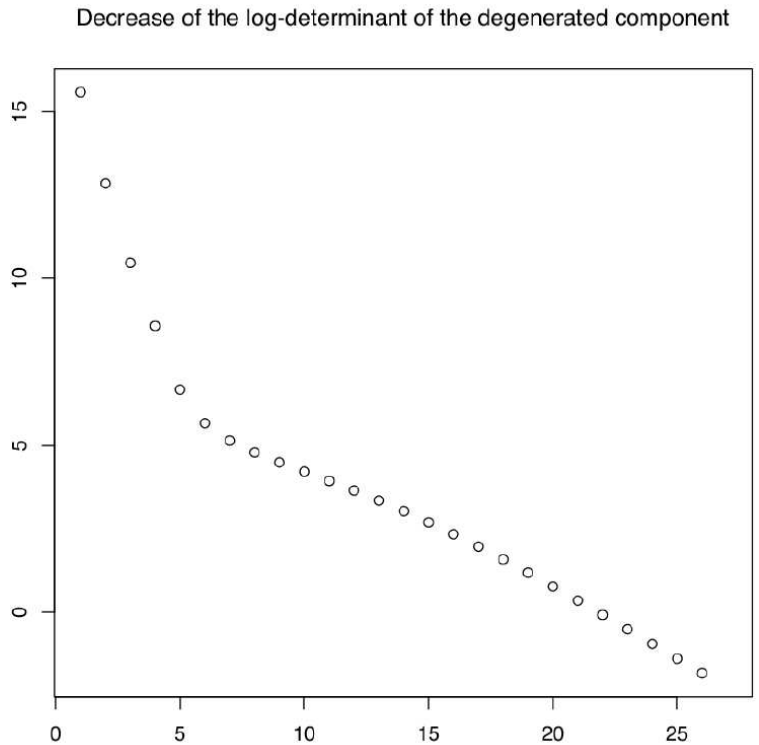
Interpretation of $\boldsymbol{\Sigma}_{ik}^{\mathrm{M}^+}$

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





Detail on the example

	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)

Intermediate conclusion on missing data

Risks

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

Missing data: an intermediary framework between complete and binned data

- Unbounded likelihood like complete data
- Slow degeneracy like binned data (but geometrical, not linear)

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

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

This leads to a linear grow of the log-likelihood (have a look also when n increases!):

$$\ell(\boldsymbol{\theta}^{(q)}; x) \sim -0.5q \log \frac{n-1}{n}$$

and geometrical convergence rate towards 0 for the variance:

$$\Sigma^{(q)} \sim \Sigma^{(0)} \left(\frac{n-1}{n}\right)^q$$

Influence of the missing data rate

% missing data	0	5	10	15	20	25	30
% deg.	16	4	12	11	46	51	100
Average nb 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 iterations before degeneracy decreases

Outline

Existing strategies for avoiding degeneracy

- Constraining the covariance matrices (*e.g.* num. tol.):

$$\forall k, |\boldsymbol{\Sigma}_k| \geq \alpha_{(n)} > 0$$

[Tanaka and Takemura, 2006]

- Relative constraints between covariance matrices:

$$\forall k \neq j, |\boldsymbol{\Sigma}_k| \geq \beta |\boldsymbol{\Sigma}_j| \quad (0 < \beta \leq 1)$$

[Hathaway, 1985] [Ingrassia and Rocci, 2007]

- Bayesian approach: With a well-behaved prior γ, maximise

$$\ln \ell(\boldsymbol{\theta}; \mathbf{x}) + \ln \gamma(\boldsymbol{\theta})$$

[Snoussi and Mahammad-Djafari, 2001] [Ciuperca et al., 2003]

Common difficulty

Additional information α, β or γ is difficult to fix.

A meaningful decomposition of the likelihood

- $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ = a partition of $\mathbf{x}$ in binary notation
- $n_k = \sum_{i=1}^{n} z_{ik}$ = nb. indiv. in class k from $\mathbf{z}$
- $\mathcal{Z}^* = \{\mathbf{z} : \forall k, n_k \geq d + 1\}$ = at least $d + 1$ elements by class

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

$$\Downarrow$$

Degeneracy in $\ell(\boldsymbol{\theta}; \mathbf{x})$ only occurs through $\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z} \notin \mathcal{Z}^*)$

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

$$\mathbf{z} \notin \mathcal{Z}^* \Rightarrow \begin{cases} \text{If } \exists k, n_k = 0\text{: } \hat{\boldsymbol{\theta}} \text{ is partially non-identifiable} \\ \text{If } \exists k, 1 \leq n_k < d+1\text{: Degeneracy in } \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z} \notin \mathcal{Z}^*) \end{cases}$$

$$\Downarrow$$

$\mathbf{z} \notin \mathcal{Z}^*$ has to be naturally discarded

$$\Downarrow$$

Strategy for avoiding degeneracy: Discarding $\mathbf{z} \notin \mathcal{Z}^*$

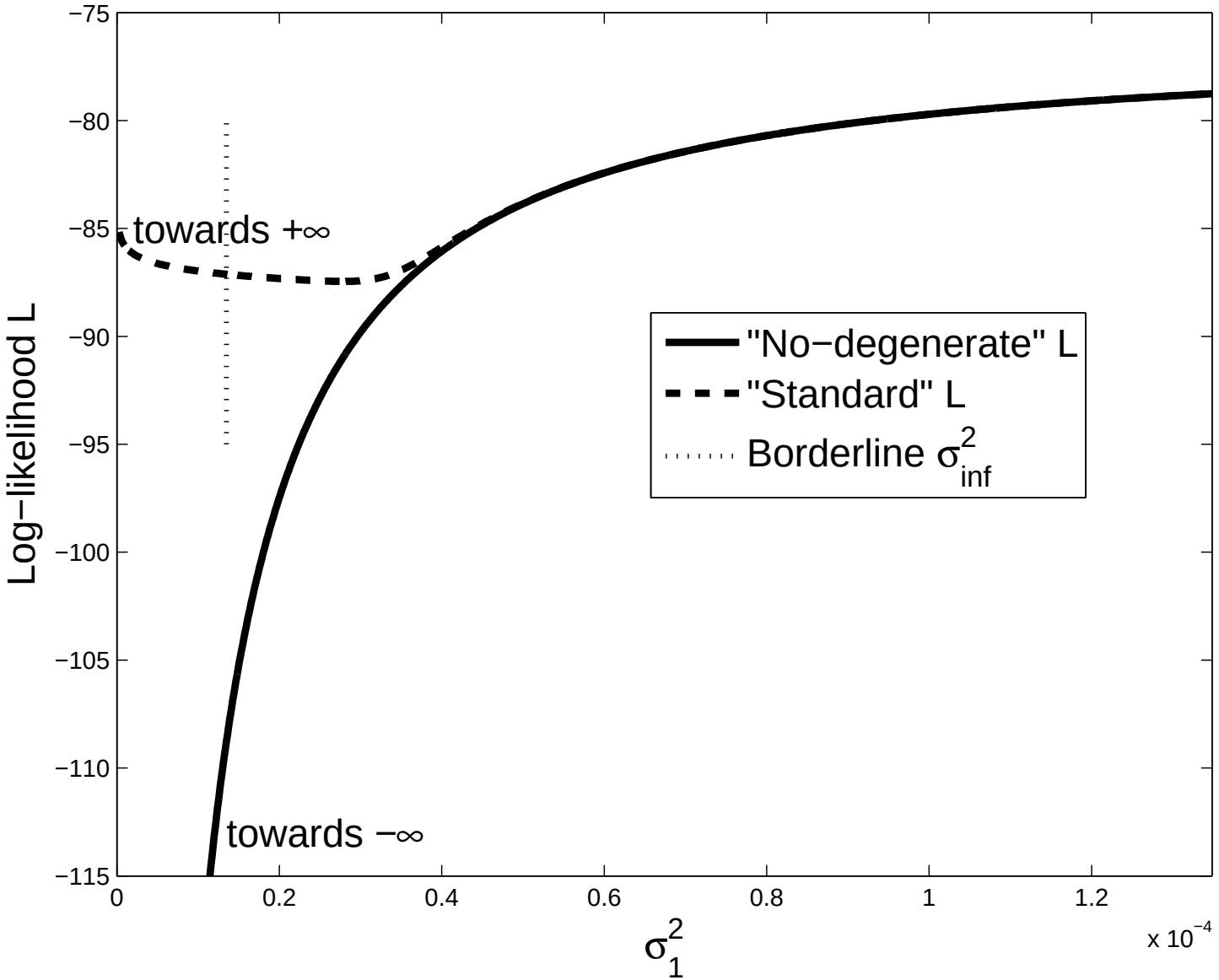
$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z} \in \mathcal{Z}^*)^a$$

[a] Adapt it with missing data: $\mathbf{z} \notin \mathcal{Z}^*$ corresponding to only observed data $\mathbf{x}^{\mathrm{O}}$

Remarks

- $\mathbf{z} \in \mathcal{Z}^*$ natural in the supervised setting to obtain non-singular cov. matrices
- $\hat{\boldsymbol{\theta}}$ approaches the ML estimator as the number of data increases [Policello, 1981]

Effect of $\mathcal{Z}^*$ on the log-likelihood



Specific EM algorithm ('EMgood'): Definition

- E step: $\tilde{z}_{ik}^{+} \propto p(\mathbf{Z} \in \mathcal{Z}^* | \mathbf{x}, z_{ik} = 1; \boldsymbol{\theta}) \overbrace{p(z_{ik} = 1 | \mathbf{x}; \boldsymbol{\theta})}^{z_{ik}^{+}}$
- M step: Standard formulas where z_{ik}^{+} is replaced by $\tilde{z}_{ik}^{+}$

Detail of E step for $g = 2$

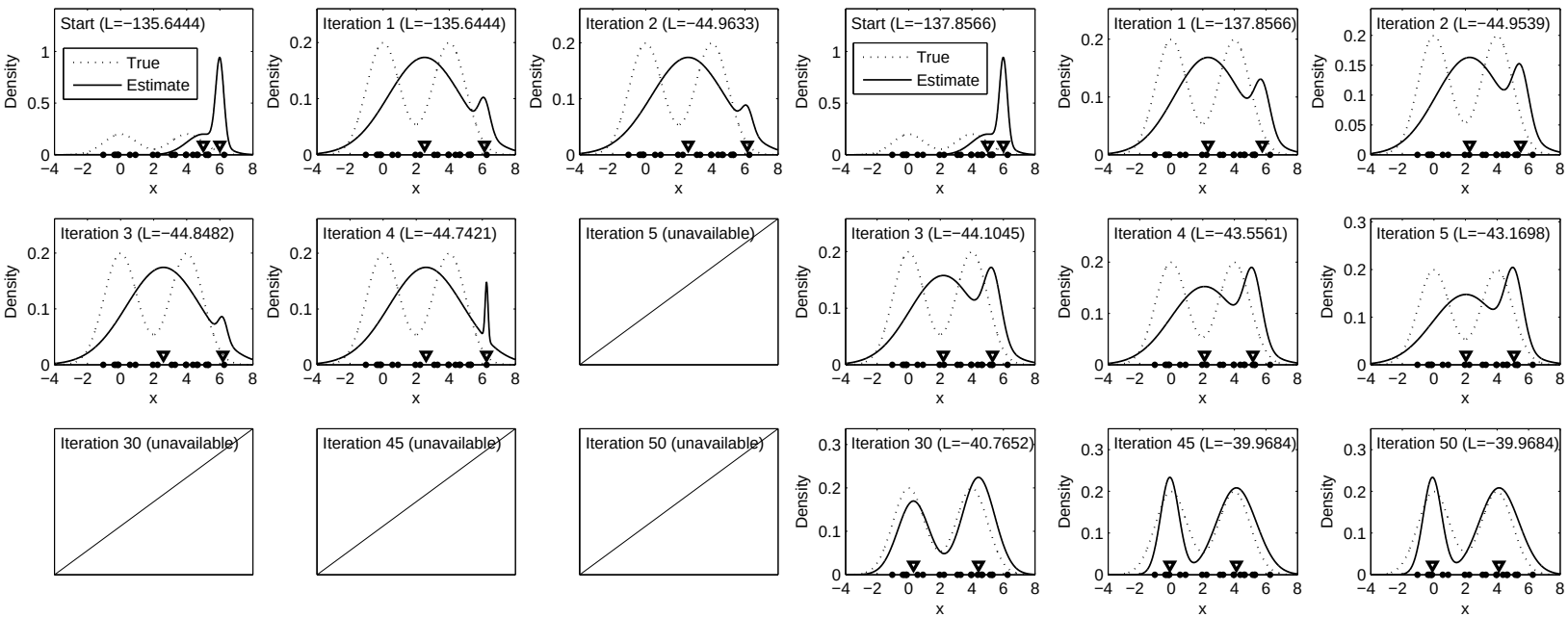
$$p(\mathbf{Z} \in \mathcal{Z}^* | \mathbf{x}, Z_{i1} = 1; \boldsymbol{\theta}) = 1 - \left(\prod_{j \neq i} t_{j2} + \prod_{j \neq i} t_{j1} + \sum_{j \neq i} t_{j2} \prod_{h \neq i, j} t_{h1} \right)$$

Combinatorial problem for $g > 2$ (Stirling nb of 2nd kind involved)

Calculus of E step becomes infeasible for most situations...

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

Example of EMgood on individual data



Standard EM EMgood

Example of EMgood on missing data

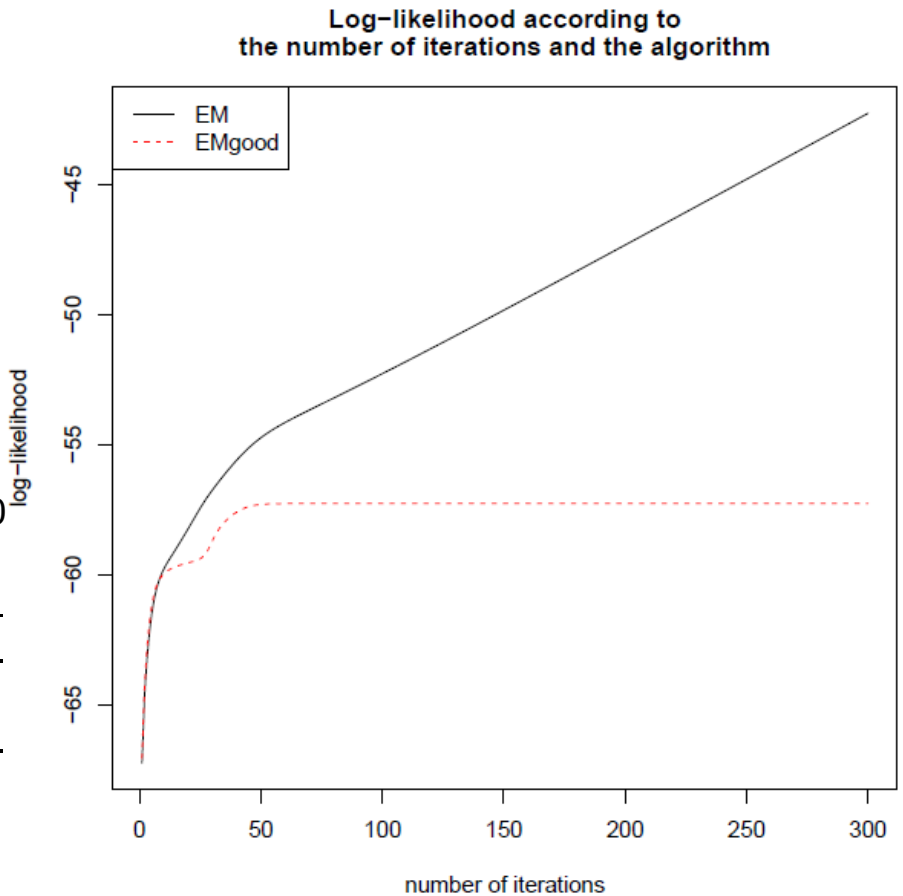
$\pi_1 = \pi_2 = 0.5$

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

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

$n = 30$ 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)



The by-product question

How to use natural information $\mathbf{Z} \in \mathcal{Z}^*$ in a more efficient way than EMgood?

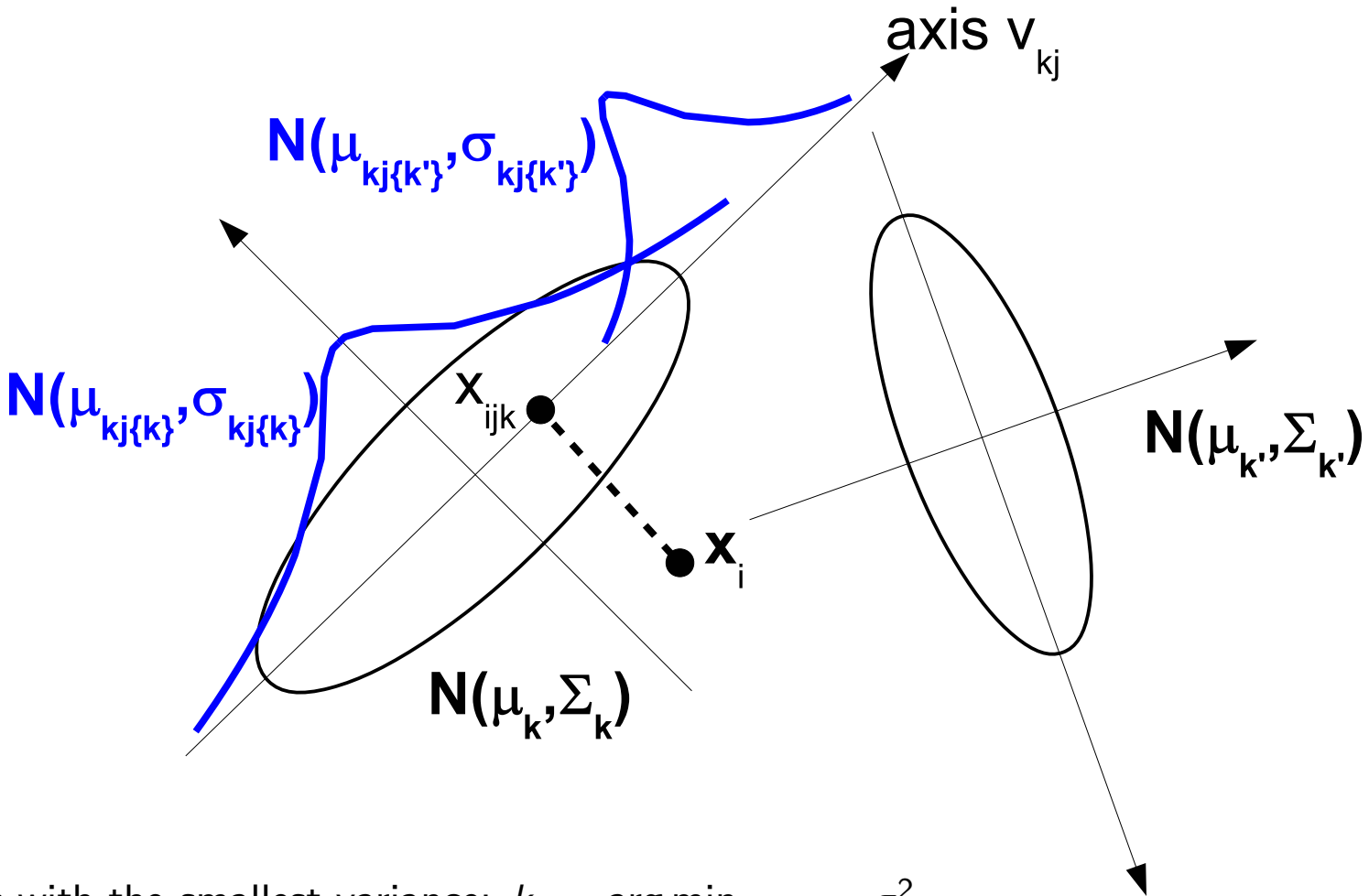
$\Downarrow$

Two strategies

- Strategy 1: Return to a lower bound on variances... but by using now additional information $\mathbf{Z} \in \mathcal{Z}^*$!
- Strategy 2: Design an approximate EMgood

Outline

Multivariate towards univariate mixtures



Class with the smallest variance: $k_0 = \arg\min_{1 \leq k' \leq g} \sigma^2_{kj\{k'\}}$

A non-asymptotic stochastic lower bound on variances

Proposition 3: The bound

For any $\alpha \in (0,1)$, we have,

$$p\left(\forall k \in \{1,\ldots,g\} \ , \ \sigma^2_{kj\{k_0\}} \geq B^d_{jk}(\alpha) \mid \mathbf{Z} \in \mathcal{Z}\right) \geq 1-\alpha,$$

where

$$B^d_{jk}(\alpha) = S^d_{jk}/\chi^2_d(1-\alpha)$$

with S^d_{jk} the minimum non-normalized variance among all subsamples of size $d+1$ in the whole sample $\{X_{ijk}\}_{i\in\{1,\ldots,n\}}$:

$$S^d_{jk} = \min_{\{\mathcal{I}:\#\mathcal{I}=d+1\}} S_{\mathcal{I}jk}.$$

Empirical variance and mean of the subsample $\{X_{ijk}\}_{i\in\mathcal{I}}$ ($\mathcal{I} \subset \{1,\ldots,n\}$)

$$S_{\mathcal{I}jk} = \sum_{i\in\mathcal{I}} (X_{ijk} - \bar{X}_{\mathcal{I}jk})^2, \quad \bar{X}_{\mathcal{I}jk} = \frac{1}{\#\mathcal{I}} \sum_{i\in\mathcal{I}} X_{ijk}$$

Sketch of proof

The proof is straightforward.

1. Axis j of component k
2. Project multivariate into univariate mixture on this axis
3. Conditionally to $\mathbf{Z} \in \mathcal{Z}^*$, there exists $d+1$ distinct random variables $\{X_{ijk}\}_{i \in \{\mathcal{I}\}}$ which belong to the class k_0
4. Classical result from a univariate Gaussian

$$p\left(\sigma^2_{kj\{k_0\}} \geq \frac{S_{\mathcal{I}jk}}{\chi^2_d((1-\alpha))} \middle| \{i \in \mathcal{I} : Z_{i,k_0} = 1\}, \mathbf{z} \in \mathcal{Z}^*\right) = 1 - \alpha.$$

5. We conclude since $S^d_{jk} \leq S_{\mathcal{I}jk}$.

Properties

- Easy and fast to compute from the order statistics
- Not very sharp since it is likely verified with far higher probability than $1 - \alpha$
- EM_α: Stop standard EM run overstepping the lower bound

Proposition 4: Consistency

$\hat{\boldsymbol{\theta}}(\alpha) = \arg\max_{\boldsymbol{\theta} \in \Theta(\alpha)} L(\boldsymbol{\theta}; \mathbf{x})$ is a consistent estimate of $\boldsymbol{\theta}$ where

$$\Theta(\alpha) = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta, \sigma^2_{kj\{k_0\}} \geq B^d_{jk}(\alpha)\}.$$

Sketch of proof

- Univariate: Rely on the result of [Tanaka and Takemura, 2006]
- Multivariate: In progress

Numerical comparison of EM_0 and EM_α: Counting runs

- $g = 2$ Gaussians, 1000 samples of size $n = 10d$
- $\boldsymbol{\theta}^{[0]}$ choosen at random
- Classical EM (EM_0): Stop either when relative increase of the log-likelihood is smaller than a standard threshold $\varepsilon = 10^{-6}$ ("normal stop") or if the numerical tolerance of the computer is reached when estimating covariance matrices ("crash stop"; indicating probably degeneracy)
- New strategy (EM_α): Stop either with a "normal stop" or a "crash stop" (the same "normal stop" and "crash stop" as EM_0), or when our bound on singular matrices is reached with $\alpha = 0.01$ (our so-called "degeneracy stop")

EM_0 stop:	crash		normal	
EM_α stop:	degeneracy	crash or normal	normal	degeneracy or crash
$d = 1$	189/189	0/189	811/811	0/811
$d = 2$	57/57	0/57	943/943	0/943
$d = 4$	34/34	0/34	966/966	0/966
$d = 8$	37/37	0/37	963/963	0/963

And about the missing data case?

This bound is expected to be inefficient because of the slow variance decrease. . .

Outline

The SEMgood algorithm

Stochastic EMgood

Introduces a stochastic step between the E and the M step of the EM algorithm:

- S step : $\mathbf{z}^+ \sim \mathbf{Z}|\mathbf{x}, \mathbf{Z} \in \mathcal{Z}^*; \boldsymbol{\theta}$
- Partition constraints easy to include: Rejection sampling, Gibbs sampling...
- Generate a sequence $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$
- Estimated parameter: $\hat{\boldsymbol{\theta}}^{\text{SEMGOOD}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}} \in \ell(\boldsymbol{\theta}; \mathbf{x})$

Numerical comparison design between EM and SEMgood

- Start both algo. from 10 random values, for each initialization iterate 300 times
- Keep the parameter associated to the best likelihood $\ell(\boldsymbol{\theta}; \mathbf{x})$
- Compute the rand index between the estimated and the true partition

SEMgood 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
- SEMgood never degenerates, the solution with the higher likelihood has an adjusted rand index of 0.30 $\Rightarrow$ SEMgood has good behavior?

SEMgood on simulated data: Spurious maxima

$\pi_1 = \pi_2 = 0.5$

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

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

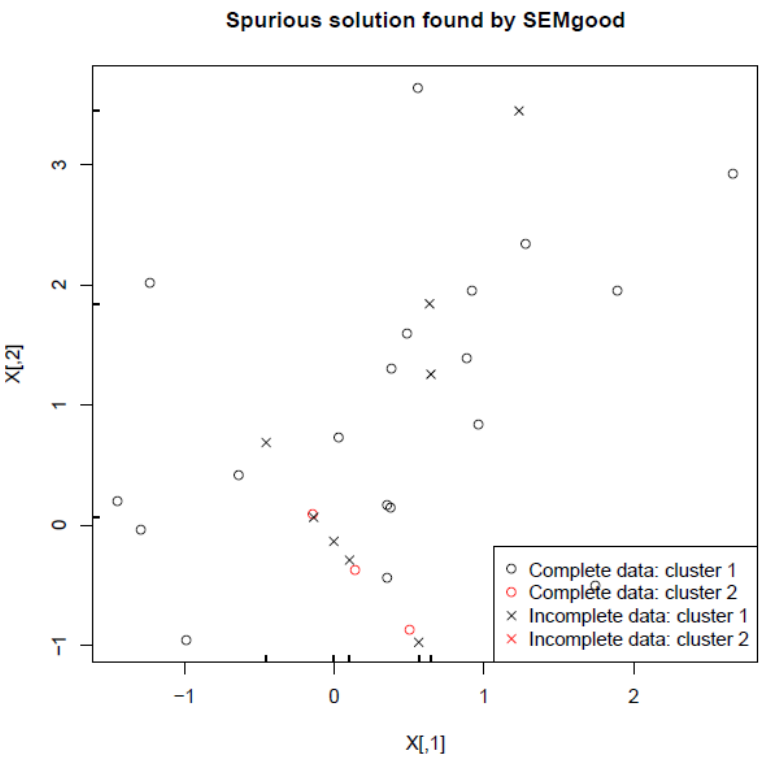
$n = 50$ data, $p = 10\%$ of missing data.

Results on 100 simulations, 10 starting values, 300 iterations by starting value.

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

Problem

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



Alternative to EMgood and SEMgood: $\overline{\text{EMgood}}$

Summary

- EMgood: combinatorial problem
- SEMgood: spurious problem (too efficient scan of the parameter space...)

Initial optimization pb

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

where

$$\mathcal{Z}^* = \{\mathbf{z} : \forall k, n_k \geq d+1)\}$$

New (and easier) optimization pb

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{x}, E[\textstyle\sum_{i=1}^n \mathbf{Z}_i] \in \bar{\mathcal{Z}}^*)$$

where

$$\bar{\mathcal{Z}}^* = \{(n_1, \ldots, n_g) : \forall k, n_k \geq d+1)\}$$

$\overline{\text{EMgood}}$

- The constraint $E[\sum_{i=1}^n \mathbf{Z}_i] \in \bar{\mathcal{Z}}^*$ is easy to satisfy
- At each E step of EM, just verify that $n_k \geq d+1$!
- If not, just stop EM (deg. situation) and start it again from another position

Numerical experiments with $\overline{\text{EMgood}}$ on simulated data

- $\pi_1 = \pi_2 = 0.5$, $d = \{2, ..., 13\}$, $\delta = 6/\sqrt{d}$, $\mu_1 = (0, \ldots, 0)$, $\mu_2 = (\delta, \ldots, \delta)$, $\Sigma_1 = \Sigma_2 = I_d$.
- 20% of missing data
- $n = 150$, $niter = 300$, $nbStart = 1$, $nrep = 100$

	2	3	4	5	6	7	8	9	10	11	12	13
EM	0.97	0.94	0.93	0.89	0.82	0.74	0.79	0.75	0.76	0.70	0.67	0.68
$\overline{\text{EMgood}}$	0.97	0.94	0.94	0.90	0.86	0.85	0.91	0.82	0.85	0.79	0.83	0.80

Table : Mean ARI for each dimension d

	2	3	4	5	6	7	8	9	10	11	12	13
EM	0.00	0.00	0.00	0.01	0.00	0.01	0.02	0.04	0.09	0.13	0.09	0.12
$\overline{\text{EMgood}}$	0.00	0.00	0.01	0.04	0.13	0.77	1.19	2.58	3.82	6.46	8.63	9.43

Table : Mean number of restarts for each dimension d

Thus $\overline{\text{EMgood}}$ seems to detect deg., allowing welcomed restartings

Outline

What is label switching?

A useful notation

- $\mathcal{P}_g$ permutation set of $\{1, \ldots, g\}$
- $\sigma(\boldsymbol{\theta}) = (\boldsymbol{\theta}_{\sigma(1)}, \ldots, \boldsymbol{\theta}_{\sigma(g)})$ with $\sigma \in \mathcal{P}_g$

Posterior invariant to label permutation

$$\left\{ \begin{array}{ll} \text{Label invariant mixture distribution} & p(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}|\sigma(\boldsymbol{\theta})) \\ \text{Label invariant prior} & p(\boldsymbol{\theta}) = p(\sigma(\boldsymbol{\theta})) \end{array} \right\}$$

$$\Downarrow$$

Label invariant posterior $p(\boldsymbol{\theta}|\mathbf{x}) = p(\sigma(\boldsymbol{\theta})|\mathbf{x})$

Consequences

Many ponctual estimates are useless: Posterior mean ($E[\boldsymbol{\theta}_1|\mathbf{x}] = E[\boldsymbol{\theta}_2|\mathbf{x}]$), ...

Gibbs algorithm in mixtures

Principle (iteration q)

- $\mathbf{z}^q \sim p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{q-1})$
- $\boldsymbol{\theta}^q \sim p(\boldsymbol{\theta}|\mathbf{x}, \mathbf{z}^q)$

Convergence towards invariant distributions

- $(\boldsymbol{\theta}^q, \mathbf{z}^q) \xrightarrow{d} p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$
- $\Rightarrow \boldsymbol{\theta}^q \xrightarrow{d} p(\boldsymbol{\theta}|\mathbf{x})$
- $\Rightarrow \mathbf{z}^q \xrightarrow{d} p(\mathbf{z}|\mathbf{x})$

A toy example (to be continued)

Mixture model

- Two univariate Gaussians ($g = 2$): $p(\cdot|\mu_k) = \mathcal{N}(\mu_k, \Sigma_k)$
- Known proportions ($\pi_k = 0.5$) and variances ($\Sigma_k = 1$)
- Unknown centers: μ_1 and μ_2 ($\mu_1 = 0$, $\mu_2 = 0.25$)

Prior

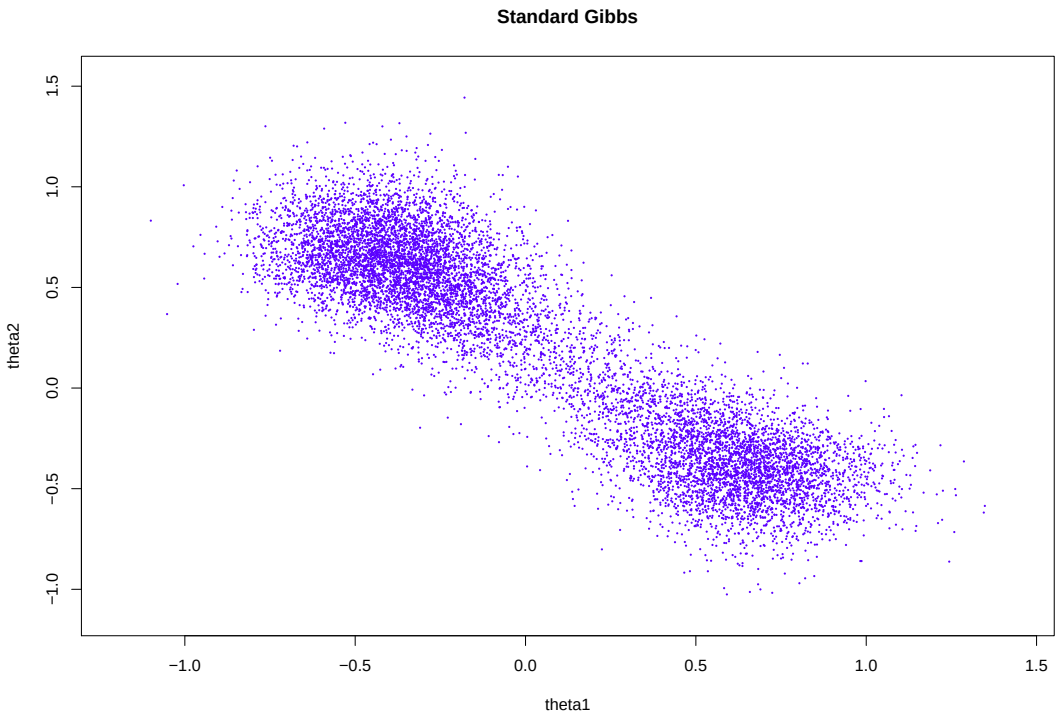
- $\mu_k \sim \mathcal{N}(0, 1)$ with $\mu_1 \perp \mu_2$

Posterior sampling from Gibbs

- $\mu_k|\mathbf{z}, \mathbf{x} \sim \mathcal{N}(n_k \bar{x}_k/(n_k + 1), 1/(n_k + 1))$
- $z_i|\mu_1, \mu_2, \mathbf{x} \sim \mathcal{M}_2(1, t_{i1}(\mu_1, \mu_2), t_{i2}(\mu_1, \mu_2))$

with $n_k = \sum_{i=1}^n \mathbb{I}_{z_i=k}$, $\bar{x}_k = \sum_{i=1}^n \mathbb{I}_{z_i=k} x_i/n_k$, $t_{ik}(\mu_1, \mu_2) = p(z_i = k|\mathbf{x}, \mu_1, \mu_2)$

$p(\boldsymbol{\theta}|\mathbf{x})$: Two modes!



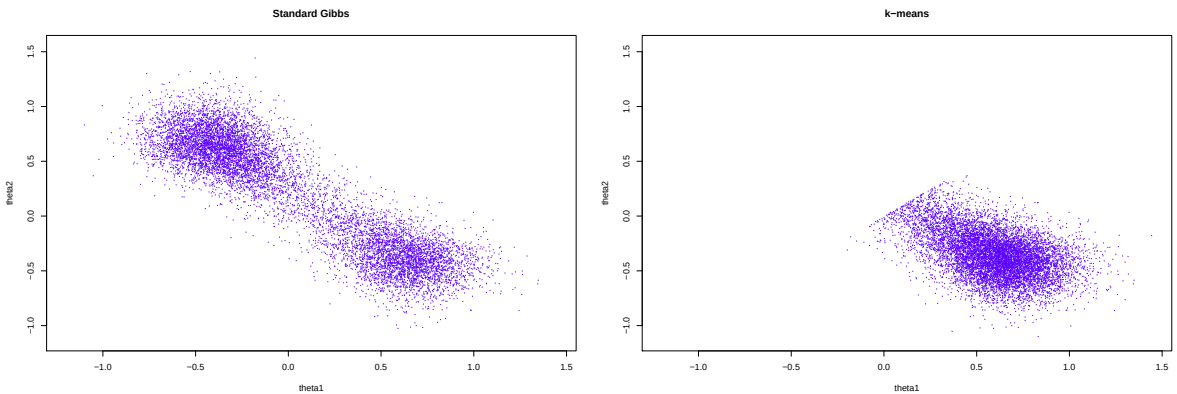
Outline

Constraining the prior

- Artificial identifiability constraints on $\boldsymbol{\theta}$
  [Diebolt & Robert '94]
- Ordering constraints: $\mu_1 < \mu_2$
- The new prior becomes proportional to $p(\boldsymbol{\theta})\mathbb{I}_{\mu_1<\mu_2}$
- Fail to solve the problem
  [Celeux *et al.* '00], [Jasra *et al.* '05]

k-means algorithm on Θ

- Relabeling algorithms on generated $\boldsymbol{\theta}$
  [Stephens '97], [Celeux '98]
- Search for a permutation minimizing a loss function
- k-means like algorithm on Θ
- Variability underestimation of the posterior $p(\boldsymbol{\theta}|\mathbf{x})$
  [Celeux '97]



Invariant loss function

- Invariant loss function to a permutation of $\boldsymbol{\theta}$ (ex.: MAP)
  [Celeux *et al.* '00]
- Require to choose a loss function related to the problem at hand
- Optimization of this function
- Many standard loss functions are not label invariant...

Probabilistic relabeling

- Take into account uncertainty on parameter permutation
  [Jasra *et al.* '05]
- Model on a noswitch posterior learned from a noswitched sequence
- Probability of each parameter permutation arising from Gibbs sampling
- Allow standard loss functions as posterior mean
- What is a noswitched sequence? Which model to choose?

Restricting the latent partition

- Use a Bernoulli mixture model for modeling $\mathbf{z}^q$
- Then, retain a particular permutation on $\mathbf{z}^q$
  [Puolamäki & Kaski '09]
- Justification of this ad hoc approach?

Outline

Main idea

Ascertainment

- The label switching is inherent to the mixture model
- Thus, there is no theoretical solution to "unswitch" $p(\boldsymbol{\theta}|\mathbf{x})$
  (at least without an external and new information but we have not)

An algorithmic (and pragmatic) idea

- Consider a sequence $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ from the Gibbs sampler for a n sample $\mathbf{x}$, thus

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q \sim p_Q(\boldsymbol{\theta}|\mathbf{x}) \overset{Q \to \infty}{\longrightarrow} p(\boldsymbol{\theta}|\mathbf{x})$$

- We know that infinite sampler $p(\boldsymbol{\theta}|\mathbf{x})$ is "bad" for some tasks because switch
- We expect that finite sampler $p_Q(\boldsymbol{\theta}|\mathbf{x})$ could be "better" for such tasks

We say "pragmatic" since many practitioners use $p_Q(\boldsymbol{\theta}|\mathbf{x})$ as it. . . we no real problems

Example of theoretical guarantees we could expect

Let $\hat{\boldsymbol{\theta}}_Q^{\text{MEAN}}$ be the mean of the Gibbs sample:

$$\hat{\boldsymbol{\theta}}_Q^{\text{MEAN}} = \frac{1}{Q} \sum_{q=1}^{Q} \boldsymbol{\theta}_q$$

Classical result

$$\lim_{n \to \infty} \left(\lim_{Q \to \infty} \hat{\boldsymbol{\theta}}_Q^{\text{MEAN}} \right) \neq \boldsymbol{\theta}$$

Result we expect

With Q_n an increasing function of n (to be defined)

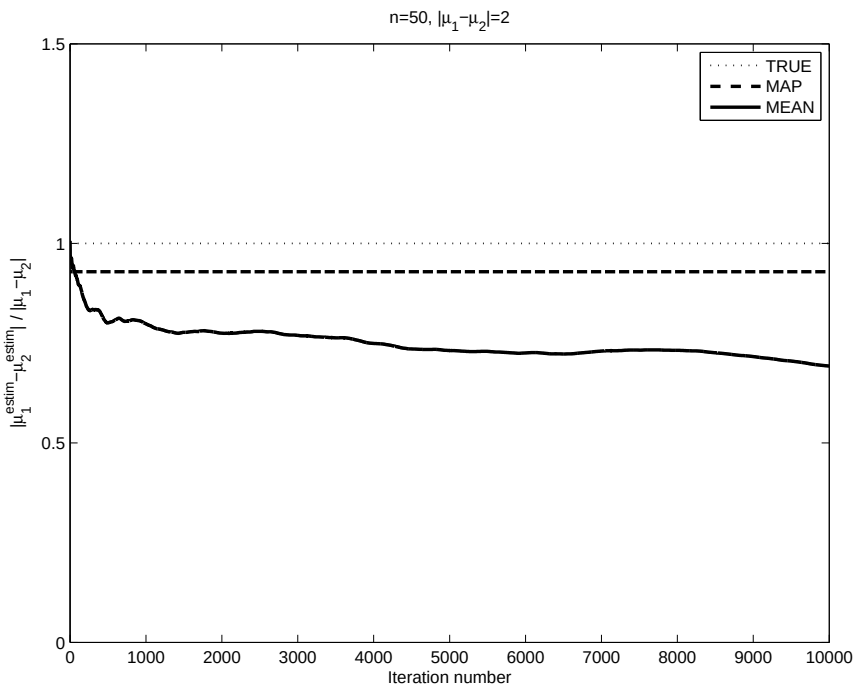
$$\lim_{n \to \infty} \hat{\boldsymbol{\theta}}_{Q_n}^{\text{MEAN}} = \boldsymbol{\theta}$$

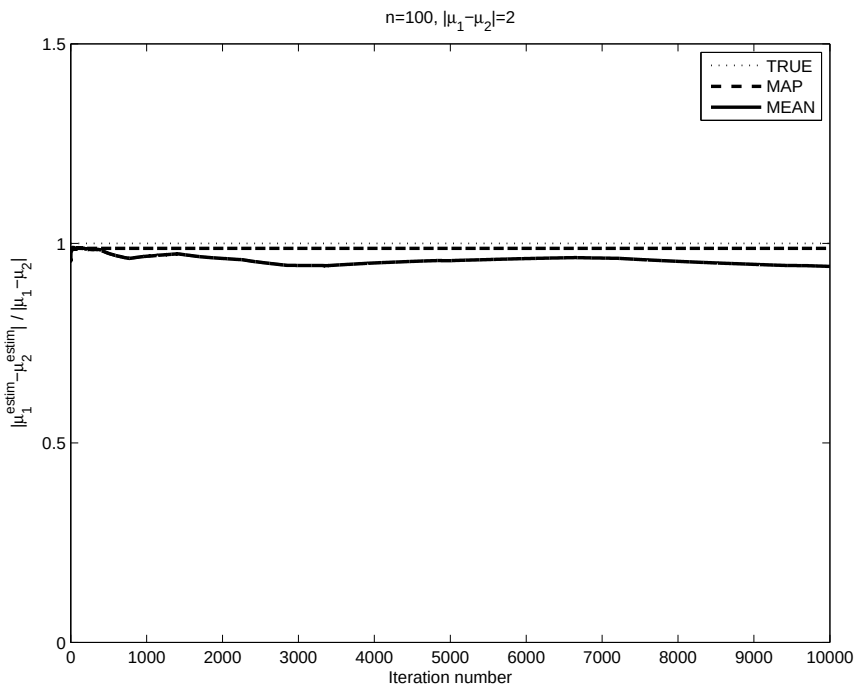
Thus Q_n plays the role of a stopping time in the Gibbs sampler

Gibbs simulation (ex. continued)

Effect of overlapping $|\mu_1 - \mu_2|$ and sample size n on $\dfrac{|\hat{\mu}_1 - \hat{\mu}_2|}{|\mu_1 - \mu_2|}$

"High" overlapp

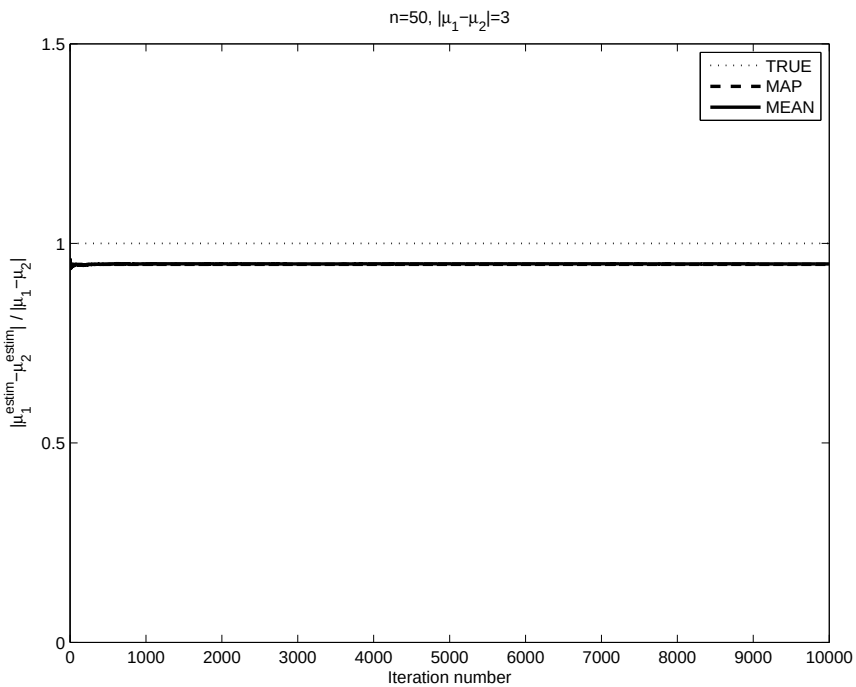


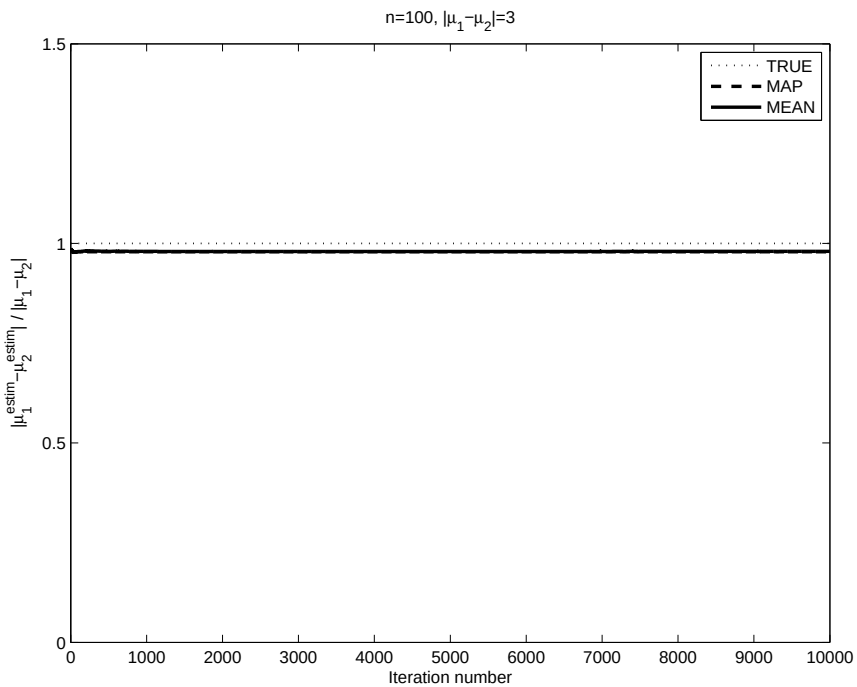


Gibbs simulation (ex. continued)

Effect of overlapping $|\mu_1 - \mu_2|$ and sample size n on $\dfrac{|\hat{\mu}_1 - \hat{\mu}_2|}{|\mu_1 - \mu_2|}$

"Low" overlapp





First theoretical attempt

A necessary condition to obtain a "good" stopping time Q_n is to have guarantee to vanish label switching in $p_{Q_n}(\boldsymbol{\theta}|\mathbf{x})$, thus

$$p_{Q_n}(\boldsymbol{\theta}|\mathbf{x}) \neq p_{Q_n}(\sigma(\boldsymbol{\theta})|\mathbf{x})$$

Our way

It implies to control the switch probability during the Gibbs dynamics

Simplified theoretical example in Gaussian mixtures

Two homoscedastic Gaussian components and $\boldsymbol{\theta}$ known up to a permutation
Probability of switch for one iteration is given by

$$p^{switch} = \frac{p(\mathbf{x}, \mathbf{z}; \sigma(\boldsymbol{\theta}))}{p(\mathbf{x}, \mathbf{z}; \sigma(\boldsymbol{\theta})) + p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})}$$

After some algebra, we asymptotically have on n

$$p^{switch} \propto \exp\left(-\frac{n}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}}\right)$$

We deduce the (asymptotic) probability of no switch during Q Gibbs iterations

$$p_Q^{noswitch} = \left(1 - p^{switch}\right)^Q \simeq \left(1 + \exp\left(-\frac{n}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}}\right)\right)^{-Q}$$

Simplified theoretical example in Gaussian mixtures (continued)

And thus, for n and/or $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|$ large enough

$$p_Q^{noswitch} \geq 1 - \varepsilon \quad \Leftrightarrow \quad Q \leq \ln(1-\varepsilon) \exp\left(\frac{n}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2_{\boldsymbol{\Sigma}^{-1}}\right)$$

So, we recognize the previous numerical results:

- Q is an increasing (fast!) function of n
- Q is also an increasing (fast!) function of the component separation

It could also explain why, in (co-)clustering (separated components), practitioners use Gibbs sampler as it and without dramatic label switching problems

Conclusion

Degeneracy

- Better undestanding, some hidden but dramatic difficulties
- Some solutions by playing on t (clustering) or A (dynamics)

Label switching

- Definitively present for m and (some) $\hat{\theta}$
- But again some (early) solutions by playing on t (clustering) or A (dynamics)

Spurious

- We have seen it is very present through a SEMgood for instance
- Still open question to solve it...