NMST434 - Ex. session 11 - EM algorithm

Ex. Mixture of (univariate) normal distributions

$$X_1, \dots, X_m \sim (1-\pi) N(\mu_0, \sigma_0^2) \oplus \pi N(\mu_1, \sigma_1^2)$$

$\theta = (\pi, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2)$ unknown

observed likelihood:

$$L_{obs,m}(\theta) = \prod_{i=1}^{m} \left[(1-\pi) \varphi_0(x_i) + \pi \varphi_1(x_i) \right], \quad \varphi_j(x) = \varphi\left(\frac{x-\mu_j}{\sigma_j}\right)\frac{1}{\sigma_j}$$

We know that for (e.g.) $\mu_0 = x_1$ and $\sigma_0 \to 0$, $L_{obs,m} \to \infty$ but there still exists a solution to the MLE equations that is a consistent estimator of the true θ.

Numerical optimization is not a good method:

- L has singularities ($L \to \infty$)
- the solution we search for is not unique (symmetry of the problem, i.e. $L(\hat\pi, \hat\mu_0, \hat\mu_1, \hat\sigma_0^2, \hat\sigma_1^2) = L(1-\hat\pi, \hat\mu_1, \hat\mu_0, \hat\sigma_1^2, \hat\sigma_0^2)$) and there will be saddle points where $\nabla L = 0$ but we might not have ~~a root of $L=0$~~ an extreme
- $\sigma_j > 0$ (or Σ_j pos. def. may be difficult to enforce)
- the number of parameters is large (for d-variate distributions and k mixing components $dk + kd(d-1)/2 = O(d^2 k)$)

EM-algorithm: augmented dataset

This is how we generate from X_i

$$\binom{X_1}{Z_1}, \dots, \binom{X_m}{Z_m} \text{ iid}, \quad Z_i \sim \text{alt}(\pi), \quad X_i | Z_i \sim \begin{cases} N(\mu_0, \sigma_0^2) & \text{if } Z_i = 0 \\ N(\mu_1, \sigma_1^2) & \text{if } Z_i = 1 \end{cases}$$

$\binom{X_1}{Z_1}$ has a density w.r.t. $\lambda \otimes$ counting measure δ_{N_0}

$$f_{X,Z}(x,z) = f_{X|Z}(x|z) P(Z=z) = \begin{Bmatrix} \varphi_0(x)(1-\pi) & \text{if } z=0 \\ \varphi_1(x)\pi & \text{if } z=1 \end{Bmatrix} =$$

$$= [\varphi_0(x)(1-\pi)]^{1-z} [\varphi_1(x)\pi]^z \qquad x \in \mathbb{R}, z \in \{0,1\}$$

$$f_X(x) = \sum_{z=0}^{1} f_{X,Z}(x,z) = \varphi_0(x)(1-\pi) + \varphi_1(x)\pi = \underline{L_{obs,1}(\theta)}$$

z_i are unobserved. Complete likelihood

$$L_m^c(\theta) = \prod_{i=1}^m \left[\varphi_0(x_i)(1-\pi)\right]^{1-z_i} \left[\varphi_1(x_i)\pi\right]^{z_i}$$

$$\ell_m^c(\theta) = \sum(1-z_i)\log\varphi_0(x_i) + \sum(1-z_i)\cdot\log(1-\pi)$$
$$+ \sum z_i \log\varphi_1(x_i) + \sum z_i \log \pi$$

E-step: for $\tilde\theta$ an initial estimate given

$$E_{\tilde\theta}[\ell_m^c(\theta) \mid X_1,\dots,X_m] = \text{"}\ell_m^c(\theta) \text{ with } \tilde z_i \text{ in place of } z_i\text{"}$$

Bayes in $\lambda\otimes\delta_{\mathbb{N}_0}$

$$E_{\tilde\theta}[z_i \mid X_1,\dots,X_m] = E_{\tilde\theta}[z_i \mid X_i] = P_{\tilde\theta}(z_i=1 \mid X_i) =$$
$$= \frac{f_{X|Z}(X_i \mid 1)\,P(Z_i=1)}{f_X(X_i)} = \frac{\tilde\varphi_1(X_i)\tilde\pi}{\tilde\varphi_0(X_i)(1-\tilde\pi)+\tilde\varphi_1(X_i)\tilde\pi} =: \tilde z_i$$

for $\varphi_j(x) = \frac{1}{\tilde\sigma_j}\varphi\left(\frac{x-\tilde\mu_j}{\tilde\sigma_j}\right)$ with the initial estimated $\tilde\theta$ in place of θ.

M-step: $Q(\theta,\tilde\theta) := E_{\tilde\theta}[\ell_m^c(\theta) \mid X_1,\dots,X_m]$

$$\frac{\partial Q}{\partial\pi} = -\frac{\sum(1-\tilde z_i)}{1-\pi} + \frac{\sum\tilde z_i}{\pi} \overset{!}{=} 0 \Rightarrow \hat\pi = \frac{1}{m}\sum_{i=1}^m \tilde z_i$$

Binom. (Bernoulli)

$$\frac{\partial Q}{\partial\mu_0} = -\sum(1-\tilde z_i)\frac{(x_i-\mu_0)}{2\sigma_0^2}\cdot 2 \overset{!}{=} 0 \Rightarrow \hat\mu_0 = \frac{\sum(1-\tilde z_i)X_i}{\sum(1-\tilde z_i)}$$

$$\frac{\partial Q}{\partial\sigma_0^2} = \sum(1-\tilde z_i)\left[-\frac{1}{2\sigma_0^2} + \frac{(x_i-\mu_0)^2}{2\sigma_0^4}\right] \overset{!}{=} 0 \Rightarrow \hat\sigma_0^2 = \frac{\sum(1-\tilde z_i)(X_i-\hat\mu_0)^2}{\sum(1-\tilde z_i)}$$

$\hat\mu_1$ and $\hat\sigma_1^2$ analogously.

now set new $\tilde\theta$ to be $\hat\theta$ from the M-step, and repeat until convergence.

0. $\tilde\theta$ = init.
1. $\tilde z_i := \tilde\varphi_1(x_i)\tilde\pi / (\tilde\varphi_0(x_i)(1-\tilde\pi)+\tilde\varphi_1(x_i)\tilde\pi)$
2. $\hat\pi = \frac{1}{m}\sum\tilde z_i$, $\hat\mu_j = \frac{\sum(1-\tilde z_i)X_i}{\sum(1-\tilde z_i)}$, $\hat\sigma_j^2 = \dots$
3. $\tilde\theta := \hat\theta$ and go to step 1.

NMST434 Cvičenie 11 - Missing data, EM algorithm

$\binom{X_i}{Y_i} \sim N(\mu, \Sigma)$ random sample of size m $\quad X_i$ missing for $i = m_0+1 \dots m_1+m_0$
Y_i missing for $i = m_0+m_1+1, \dots m$

CCA: complete case analysis - consider only i such that both X_i and Y_i are known

ACA: available case analysis - for each estimator use all available observations

Simple Imputation: if X_i is known but Y_i not, build model $Y \sim \beta_0 + \beta_1 X_i$ based on fully observed data and estimate Y_i by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \sim E[Y_i | X_i]$

Parametric Imputation: estimate also ε_i in $Y = \beta_0 + \beta_1 X_i + \varepsilon_i$ and generate resamples of $\hat{\varepsilon}_i \rightarrow \hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$

Multiple Imputation: Repeat an imputation procedure M times, estimate $\hat{\Theta}_j$ in the j-th imputation, and set $\hat{\Theta}_{MI} = \overline{\hat{\Theta}}_M$ $\sim$ Bootstrap - get parameter estimates out of resamples $\rightarrow$ possible to approximate also the ~~var~~ variance of $\hat{\Theta}_{MI}$:

$$\operatorname{var} \hat{\Theta}_{MI} = E\left(\operatorname{var}(\hat{\Theta}_{MI} \mid \text{imputed data})\right) + \operatorname{var}\left(E(\hat{\Theta}_{MI} \mid \text{imp. d})\right)$$

$$\approx \frac{1}{M} \sum_{j=1}^{M} \widehat{\operatorname{var}}(\hat{\Theta}_{MI} \mid \text{imp. d}_j) + \frac{1}{n-1} \sum (\hat{\Theta}_j - \hat{\Theta}_{MI})(\hat{\Theta}_j - \hat{\Theta}_{MI})^T$$

but this is not a good estimator - needs to be adjusted for estimation of the sampling distribution

$$m_2 = m - m_1 - m_0 \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \qquad z_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$$

EM-algorithm: complete log-likelihood: $\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$

$$\ell_m^c(\theta) = c - \frac{m}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{m} (z_i - \mu)' \Sigma^{-1} (z_i - \mu) \qquad (*)$$

want to maximize observed log-likelihood:

$$\ell_{obs}(\theta) = c - \frac{m_0}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{m_0} (z_i - \mu)' \Sigma^{-1} (z_i - \mu) \qquad \text{complete pairs of } x_i, y_i$$

$$- \frac{m_1}{2} \log \sigma_{22} - \frac{1}{2\sigma_{22}} \sum_{i=m_0+1}^{m_0+m_1} (y_i - \mu_2)^2 \qquad \text{only } y_i \text{ observed}$$

$$- \frac{m_2}{2} \log \sigma_{11} - \frac{1}{2\sigma_{11}} \sum_{i=m_0+m_1+1}^{m} (x_i - \mu_1)^2 \qquad \text{only } x_i \text{ observed}$$

$\rightarrow$ no closed solution to $\max \ell_{obs}(\theta)$

EM-algorithm: i) E-step: $E_{\tilde{\theta}}\left[\ell_m^c(\theta) \mid \text{observed d.}\right] = Q(\theta, \tilde{\theta})$

$\tilde{\theta}$ given $\qquad$ depends only on $E_{\tilde{\theta}}[X_i \mid Y_i]$, $E_{\tilde{\theta}}[Y_i \mid X_i]$ from (*)
θ, and obs. data. $\quad E_{\tilde{\theta}}[X_i^2 \mid Y_i]$, $E_{\tilde{\theta}}[Y_i^2 \mid X_i]$

or by the fact that we have an exponential family of densities.

$$X_i \mid Y_i \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(Y_i - \mu_2);\ \sigma_{11} - \sigma_{12}\sigma_{22}^{-1}\sigma_{21}\right)$$

$$E_{\tilde\theta}[X_i|Y_i] = \tilde\mu_1 + \tilde\sigma_{12}\tilde\sigma_{22}^{-1}(Y_i - \tilde\mu_2)$$

$$E_{\tilde\theta}[X_i^2|Y_i] = \mathrm{var}_{\tilde\theta}(X_i|Y_i) + \left[E_{\tilde\theta}(X_i|Y_i)\right]^2$$

$$= \underbrace{\tilde\sigma_{11} - \tilde\sigma_{12}\tilde\sigma_{22}^{-1}\tilde\sigma_{21}}_{\tilde\sigma_{11}\left(1-\frac{\tilde\sigma_{12}}{\tilde\sigma_{22}\tilde\sigma_{11}}\right) = \tilde\sigma_{11}(1-\tilde\rho^2)} + \left[E_{\tilde\theta}(X_i|Y_i)\right]^2$$

for $\mathcal{L}(Y_i|X_i)$ analogous.

ii) M-step: maximize $Q(\theta, \tilde\theta)$ over θ

- for μ this leads to the usual MLE with replaced values of unobs. data.

$\tilde z_i := (\tilde X_i, \tilde Y_i)$ for $\tilde X_i = \begin{cases} X_i & \text{if } X_i \text{ is observed} \\ E_{\tilde\theta}[X_i|Y_i] & \text{if } X_i \text{ is not observed} \end{cases}$ and $\tilde Y_i$ anal.

$$\tilde\mu_{new} = \frac{1}{m}\sum_{i=1}^m \tilde z_i = \frac{1}{m}\sum_{i=1}^m E[z_i|obs]$$

$$\left[\ \frac{\partial}{\partial\mu}\left[-\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum(\tilde z_i - \mu)'\Sigma^{-1}(\tilde z_i - \mu)\right] = \sum(\tilde z_i - \mu)'\Sigma^{-1} \overset{!}{=} 0\right.$$

$$\left.\hat\mu = \frac{1}{m}\sum \tilde z_i\ \right]$$

for Σ: $$\frac{\partial}{\partial\Sigma}\left[-\frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum\underbrace{(\tilde z_i - \hat\mu)'\Sigma^{-1}(\tilde z_i - \hat\mu)}_{\mathrm{Tr}((\tilde z_i-\mu)'\Sigma^{-1}(\tilde z_i-\mu)) = \mathrm{Tr}(\Sigma^{-1}(\tilde z_i-\mu)(\tilde z_i-\mu)^T)}\right]$$

$$\left.\begin{array}{l} \frac{\partial}{\partial A}\mathrm{Tr}(A) = I \\ \\ \frac{\partial |A|}{\partial A} = |A|A^{-1} \quad \text{Jacobi's formula} \end{array}\right\} \Rightarrow = -\frac{m}{2}\frac{|\Sigma|\Sigma^{-1}}{|\Sigma|} + \frac{1}{2}\sum \Sigma^{-1}(\tilde z_i-\mu)(\tilde z_i-\mu)^T\Sigma^{-1} = 0$$

$$\Sigma = \frac{1}{m}\sum(\tilde z_i - \mu)(\tilde z_i - \mu)^T$$

$$\frac{\partial\,\mathrm{Tr}(A^{-1}B)}{\partial A} = -A^{-1}BA^{-1}$$

$$\tilde\Sigma_{new} = \frac{1}{m}\sum E\left[(z_i - \hat\mu_{new})(z_i - \hat\mu_{new})^T\,\middle|\,obs\right]$$

$$= \frac{1}{m}\sum E[z_i z_i'|obs] - \hat\mu_{new}\hat\mu_{new}'$$

$$= \frac{1}{m}\sum E\left[\begin{pmatrix} x_i^2 & x_i y_i \\ x_i y_i & y_i^2\end{pmatrix}\,\middle|\,obs\right] - \hat\mu_{new}\hat\mu_{new}'$$

+ iterate EM-steps until convergence

Reweighting: if $P(X_i \text{ is not observed}) = \pi_i$ can be estimated, assign to each observed X_i a weight proportional to $\frac{1}{\pi_i}$ and perform analysis.