Homework 6 & Mixture models

Gabriela Malenová

May 27, 2020

Problem 2: Question 7

Let $D = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\}$, where $\mathbf{x}^{(i)} \in \mathbb{N}_0^d$ and $y^{(i)} \in \{0, 1\}$. Here, $\mathbf{x} = [x_1, \ldots, x_d]$ and the class conditional distributions, $P(x_i|y)$, are given by independent Poisson distributions. What is the joint distribution $P(\mathbf{x}, y)$?

Solution:

- Definition of joint distribution: $P(A, B) := P(B)P(A|B)$.
- In our case:

$$\begin{aligned} P(\mathbf{x}, y) &= P(x_1, \ldots, x_d, y) \\ &= P(x_2, \ldots, x_d, y)P(x_1|x_2, \ldots, x_d, y). \end{aligned}$$

- Since x_i:s are independent, $P(x_1|x_2, \ldots, x_d, y) = P(x_1|y)$.

Problem 2: Question 7

- Hence,
$$\begin{aligned} P(\mathbf{x}, y) &= P(x_2, \ldots, x_d, y) P(x_1|y) \\ &= P(x_3, \ldots, x_d, y) P(x_1|y) P(x_2|y) \\ &= \ldots = P(y) P(x_1|y) \ldots P(x_d|y) \\ &= P(y) \prod_{j=1}^{d} P(x_j|y). \end{aligned}$$

- Let $\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1 \in \mathbb{R}^d$ be the parameters of the Poisson distributions for $y = 0$ and $y = 1$ respectively. Then

$$\begin{aligned} P(\mathbf{x}, y) &= P(y) \prod_{j=1}^{d} Poisson(\lambda_{y,j}) \\ &= P(y) \prod_{j=1}^{d} \frac{e^{-\lambda_{y,j}} \lambda_{y,j}^{x_j}}{x_j!}. \end{aligned}$$

Problem 2: Question 8

Use MLE to optimize the parameters $p_y := P(Y = y)$ and $\lambda_{y,j}$.

Solution:

- Define, $n_1 = \sum_{i=1}^{n} y_i$ and $n_0 = n - n_1$. The probability of observing y is $p_y = P(Y = y) = \frac{n_y}{n}$.
- From Question 6, the MLE for $\lambda_{y,j}$ is the empirical mean of x_j (j denotes dimension, not sample) labeled as y. More precisely,

$$\lambda_{y,j} = \frac{1}{n_y} \sum_{i=1}^{n} x_j^{(i)} I_{Y_j = y}.$$

Problem 2: Question 9

New observation, $\mathbf{x} \in \mathcal{X}$, predict $y_{\text{pred}} = \arg\max_{y \in \mathcal{Y}} P(y|X = \mathbf{x})$.
Find the hyperplane that determines the label prediction.

Solution:

- Decision boundary: $P(y = 0|X = \mathbf{x}) = P(y = 1|X = \mathbf{x})$.
- Joint distribution: $P(y, X = \mathbf{x}) = P(\mathbf{x})P(y|X = \mathbf{x})$.
- Hence

$$\begin{aligned} & P(y = 0|X = \mathbf{x}) = P(y = 1|X = \mathbf{x}) \\ \iff & P(y = 0, X = \mathbf{x}) = P(y = 1, X = \mathbf{x}). \end{aligned} \tag{1}$$

- From Question 7: $P(y, X = \mathbf{x}) = p_y \prod_{j=1}^{d} \frac{e^{-\lambda_{y,j}} \lambda_{y,j}^{x_j}}{x_j!}$
- Then

$$(1) \iff p_0 \prod_{j=1}^{d} \frac{e^{-\lambda_{0,j}} \lambda_{0,j}^{x_j}}{x_j!} = p_1 \prod_{j=1}^{d} \frac{e^{-\lambda_{1,j}} \lambda_{1,j}^{x_j}}{x_j!}$$

Problem 2: Question 9

- We cancel out $x_j!$ and do log:

$$p_0 \prod_{j=1}^{d} e^{-\lambda_{0,j}} \lambda_{0,j}^{x_j} = p_1 \prod_{j=1}^{d} e^{-\lambda_{1,j}} \lambda_{1,j}^{x_j}$$

$$\iff \log p_0 + \sum_{j=1}^{d} \log\left(e^{-\lambda_{0,j}} \lambda_{0,j}^{x_j}\right) = \log p_1 \sum_{j=1}^{d} \log\left(e^{-\lambda_{1,j}} \lambda_{1,j}^{x_j}\right)$$

$$\iff \log p_0 + \sum_{j=1}^{d} (x_j \log \lambda_{0,j} - \lambda_{0,j}) = \log p_1 + \sum_{j=1}^{d} (x_j \log \lambda_{1,j} - \lambda_{1,j})$$

$$\iff \log \frac{p_1}{p_0} + \sum_{j=1}^{d} (x_j \log \frac{\lambda_{1,j}}{\lambda_{0,j}} + \lambda_{0,j} - \lambda_{1,j}) = 0.$$

- Define $a_j := \log \frac{\lambda_{1,j}}{\lambda_{0,j}},\ b := -\log \frac{p_1}{p_0} + \sum_{j=1}^{d} (\lambda_{1,j} - \lambda_{0,j})$, then

$$\iff \mathbf{a}^T \mathbf{x} = b.$$

Problem 2: Question 9

Inequality (arbitrary assignment on the boundary):

$$
\begin{aligned}
y_{\text{pred}} = 1 &\iff P(y = 0 | X = \mathbf{x}) \leq P(y = 1 | X = \mathbf{x}) \\
&\iff \mathbf{a}^T \mathbf{x} \geq b, \\
y_{\text{pred}} = 0 &\iff \mathbf{a}^T \mathbf{x} < b.
\end{aligned}
$$

To sumarize: $y_{\text{pred}} = [\mathbf{a}^T \mathbf{x} \geq b]$.

Problem 2: Question 10

One can define a cost function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such that $c(y_{\text{pred}}, y_{\text{true}})$ is the cost of predicting y_{pred} given the true label is y_{true}. What is the Bayes optimal decision rule for c wrt a distribution $P(X, Y)$.

Solution: According to Bayesian Decision Theory, the best action (from $\mathcal{A}$) to take is the one that minimizes the cost

$$a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_Y[c(a, Y)|X].$$

In our case, $\mathcal{A} = \mathcal{Y}$ (the decision corresponds to picking a label), so we conclude that

$$y^* = \arg\min_{y \in \mathcal{Y}} \mathbb{E}_Y[c(y, Y)|X].$$

Note: Answer (c) is correct, not (a) as previously stated.

Problem 3: Question 12

Posterior probablities for multiclass logistic regression $P(y=k|X=\mathbf{x}) = \frac{\exp(\mathbf{a}_k^T\mathbf{x})}{\sum_i \exp(\mathbf{a}_i^T\mathbf{x})}$. The cross entropy error reads

$$E(\mathbf{a}_1, \ldots, \mathbf{a}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \log P(y=k|X=\mathbf{x}_n),$$

where $t_{nk} := \delta_{\text{label of } \mathbf{x}_n, k}$. Compute $\nabla_{\mathbf{a}_j} E$.

Solution: First, define $y_{kn} := P(y=k|X=\mathbf{x}_n)$. Note that

$$\nabla_{\mathbf{a}_j} E = -\sum_n \sum_k t_{nk} \frac{\nabla_{\mathbf{a}_j} y_{kn}}{y_{kn}}.$$

So we need to expand $\nabla_{\mathbf{a}_j} y_{kn}$. Two cases: (a) $j = k$, (b) $j \neq k$.

Problem 3: Question 12

Recall: $y_{kn} = \frac{\exp(\mathbf{a}_k^T \mathbf{x}_n)}{\sum_i \exp(\mathbf{a}_i^T \mathbf{x}_n)}$, $\mathbf{a}_j = [a_{j1}, \ldots, a_{jd}]$, $\mathbf{x}_n = [x_{n1}, \ldots, x_{nd}]$.

- (a) $j = k$

$$\frac{\partial y_{kn}}{\partial a_{j\ell}} = \frac{x_{n\ell} \exp(\mathbf{a}_k^T \mathbf{x}_n) \sum_i \exp(\mathbf{a}_i^T \mathbf{x}_n) - x_{n\ell} \exp(\mathbf{a}_k^T \mathbf{x}_n) \exp(\mathbf{a}_j^T \mathbf{x}_n)}{\left(\sum_i \exp(\mathbf{a}_i^T \mathbf{x}_n)\right)^2}$$

$$= x_{n\ell} y_{kn} (1 - y_{jn}).$$

- (b) $j \neq k$, the first term disappears,

$$\frac{\partial y_{kn}}{\partial a_{j\ell}} = -\frac{x_{n\ell} \exp(\mathbf{a}_k^T \mathbf{x}_n) \exp(\mathbf{a}_j^T \mathbf{x}_n)}{\left(\sum_i \exp(\mathbf{a}_i^T \mathbf{x}_n)\right)^2} = -x_{n\ell} y_{kn} y_{jn}.$$

Altogether (vector-wise):

$$\nabla_{\mathbf{a}_j} y_{kn} = \mathbf{x}_n y_{kn} (\delta_{jk} - y_{jn}).$$

Problem 3: Question 12

Then

$$
\begin{aligned}
\nabla_{\mathbf{a}_j} E &= -\sum_n \sum_k t_{nk} \frac{\nabla_{\mathbf{a}_j} y_{kn}}{y_{kn}} \\
&= -\sum_n \sum_k t_{nk} \frac{\mathbf{x}_n y_{kn}(\delta_{jk} - y_{jn})}{y_{kn}} \\
&= -\sum_n \mathbf{x}_n \left(\underbrace{\sum_k t_{nk}\delta_{jk}}_{t_{nj}} - y_{jn} \underbrace{\sum_k t_{nk}}_{=1} \right) \\
&= \sum_n \mathbf{x}_n (y_{jn} - t_{nj}) \\
&= \sum_n \mathbf{x}_n (P(y = j | X = \mathbf{x}_n) - t_{nj}).
\end{aligned}
$$

Gaussian mixture models

GMM: Tailored to fit multimodal distributions.

- d dimensions
- N observations
- K mixture components, each normal but with different parameters (mean, covariance matrix)

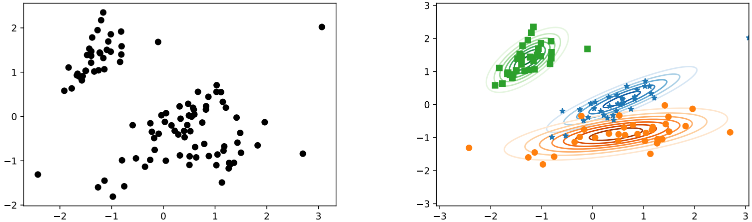

Figure: $d = 2$, $K = 3$, $N = 100$.

Expectation-Maximization (EM)

We assume

$$p(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \pi_k \geq 0, \quad \sum_k \pi_k = 1,$$

where $\theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let us introduce the latent variable $z \in \{1, \ldots, K\}$ to determine the component from which an observation originates. It's prior and conditional distributions are

$$p(z = k) = \pi_k, \qquad p(\mathbf{x}|z = k) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

EM algorithm:

- E step: membership probabilities

$$r_{nk} := p(z_n = k|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- M step: Maximize likelihood function over all sample points

$$p(\mathbf{x}_{1:N}|\theta) = \prod_{n=1}^{N} p(\mathbf{x}_n|\theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

M step

More convenient: maximize the following related loss function instead

$$L(\theta) = \mathbb{E}[p(\mathbf{x}, z|\theta)] = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \log(\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)).$$

Differentiating wrt to θ and setting to 0, the optimal solution reads

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}},$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} r_{nk}},$$

$$\pi_k = \frac{\sum_{n=1}^{N} r_{nk}}{\sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk}}.$$

Moreover $N_k = \sum_{n=1}^{N} r_{nk}$ and $\sum_{k=1}^{K} N_k = N$.

Exam question

You are given a data set $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, $i = 1, \dots, N$. The data points accumulate on m different lines, $\mathbf{a}_j^T \mathbf{x}_i = y_i$, for $\mathbf{a}_j \in \mathbb{R}^d$, $j = 1, \dots, m$.

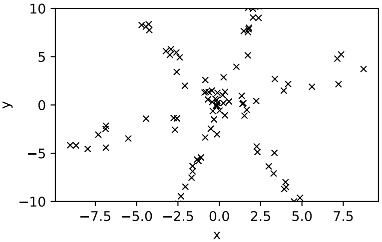

Figure: $d = 1$, $m = 3$.

Then

$$p(\mathbf{x}, y|\theta) = \sum_{j=1}^{m} \pi_j \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{a}_j^T \mathbf{x} - y)^2}{2\sigma^2}\right),$$

where $\theta = (\pi_{1:m}, \mathbf{a}_{1:m})$, $\sum \pi_j = 1$, $\pi_j \geq 0$ and $\sigma > 0$ is given and fixed.

Exam question

- (a) Find the responsibilities in the E-step of Soft EM, $r_{nk}^{(t)} = p(z_n = k | \mathbf{x}_n, y_n, \theta^{(t-1)})$.
- (b) Write down the class predictions $z_n^{(t)}$ for $(\mathbf{x}_n, z_n)$ in the E-step of Hard EM in terms of $r_{nk}^{(t)}$.
- (c) Assume that we observe the true labels $z_1, \ldots, z_\ell$ for the first ℓ datapoints, $\ell < N$. How can we modify the E-step of Soft EM to incorporate the additional information?
- (d) Write down the optimization objective for the M-step in Soft EM for $\pi_j^{(t)}$ and $\mathbf{a}_j^{(t)}$ in the terms of the responsibilities $r_{nk}^{(t)}$.

Exam question

(a) Find the responsibilities in the E-step of Soft EM, $r_{nk}^{(t)} = p(z_n = k|\mathbf{x}_n, y_n, \theta^{(t-1)})$.

Solution: Using Bayes' theorem (omitting θ)

$$p(z_n = k|\mathbf{x}_n, y_n) = \frac{p(\mathbf{x}_n, y_n|z_n = k)p(z_n = k)}{\sum_{j=1}^{m} p(\mathbf{x}_n, y_n|z_n = j)p(z_n = j)}.$$

We have $p(z = j) = \pi_j^{(t-1)}$, and $p(\mathbf{x}, y|z = k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{a}_k^{(t-1)^T}\mathbf{x} - y)^2}{2\sigma^2}\right)$. Hence

$$r_{nk}^{(t)} = \frac{\pi_k^{(t-1)} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{a}_k^{(t-1)^T}\mathbf{x}_n - y_n)^2}{2\sigma^2}\right)}{\sum_{j=1}^{m} \pi_j^{(t-1)} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{a}_j^{(t-1)^T}\mathbf{x}_n - y_n)^2}{2\sigma^2}\right)}.$$

Exam question

(b) Write down the class predictions $z_n^{(t)}$ for $(\mathbf{x}_n, z_n)$ in the E-step of Hard EM in terms of $r_{nk}^{(t)}$.

Solution

$$z_n^{(t)} = \arg\max_k r_{nk}.$$

(c) Assume that we observe the true labels $z_1, \ldots, z_\ell$ for the first ℓ datapoints, $\ell < N$. How can we modify the E-step of Soft EM to incorporate the additional information?

Solution: We change responsibilities of the corresponding point-cluster pairs to 1, and set to 0 otherwise, i.e.

$$r_{nk}^{(t)} = \delta_{z_n k}, \quad \text{for } n \leq \ell, \ \forall k.$$

Exam question

(d) Write down the optimization objective for the M-step in Soft EM for $\pi_j^{(t)}$ and $\mathbf{a}_j^{(t)}$ in the terms of the responsibilities $r_{nk}^{(t)}$.

Solution: Loss function:

$$L(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{m} r_{nk}^{(t)} \log \left(\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(\mathbf{a}_k{}^T \mathbf{x}_n - y_n)^2}{2\sigma^2} \right) \right).$$

We select

$$\pi_{1:m}^{(t)}, \mathbf{a}_{1:m}^{(t)} = \arg \max_{\pi_{1:m}, \mathbf{a}_{1:m}} L(\theta),$$

such that $\pi_j \geq 0$, $\sum_j \pi_j = 1$.

Exam question

(e) Derive the explicit update rules for $\pi_j^{(t)}$ and $\mathbf{a}_j^{(t)}$ used in the M-step of Soft EM. Hint: You can assume that the matrix $\sum_{n=1}^{N} r_{nk}^{(t)} \mathbf{x}_n \mathbf{x}_n^T$ is invertible for all $k = 1, \ldots, m$.

Solution: First, denote $y_{kn} := \pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{a}_k{}^T \mathbf{x}_n - y_n)^2}{2\sigma^2}\right)$. Then differentiate $L(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{m} r_{nk}^{(t)} \log\left(y_{kn}\right)$ wrt $\mathbf{a}_j$:

$$\frac{\partial L}{\partial a_{j\ell}} = \sum_n r_{nj}^{(t)} \frac{\partial_{a_{j\ell}} y_{jn}}{y_{jn}} = \sum_n r_{nj}^{(t)} \frac{-y_{jn}\left(\frac{1}{\sigma^2}\right)(\mathbf{a}_j^T \mathbf{x}_n - y_n) x_{n\ell}}{y_{jn}}.$$

We want $\frac{\partial L}{\partial a_{j\ell}} = 0$ hence (vector-wise)

$$\sum_n r_{nj}^{(t)} (\mathbf{a}_j^T \mathbf{x}_n - y_n) \mathbf{x}_n^T = 0$$

$$\iff \mathbf{a}_j^T \sum_n r_{nj}^{(t)} \mathbf{x}_n \mathbf{x}_n^T = \sum_n r_{nj}^{(t)} y_n \mathbf{x}_n^T.$$

Exam question

Since $\sum_{n=1}^{N} r_{nk}^{(t)} \mathbf{x}_n \mathbf{x}_n^T$ is invertible, we can write

$$\mathbf{a}_j^T = \left(\sum_n r_{nj}^{(t)} y_n \mathbf{x}_n^T \right) \left(\sum_n r_{nj}^{(t)} \mathbf{x}_n \mathbf{x}_n^T \right)^{-1}.$$

Similarly, we want to differentiate L wrt to π_j to obtain the extrema, however, *we must not forget the constraint* $\sum_k \pi_k = 1$! To incorporate it, we will use the Lagrange multipliers. More precisely, we want to maximize

$$\widetilde{L}(\theta) = L(\theta) + \lambda \left(\sum_k \pi_k - 1 \right).$$

Then differentiation wrt both λ and π_j gives

$$\frac{\partial \widetilde{L}}{\partial \lambda} = \sum_k \pi_k - 1 \qquad \stackrel{!}{=} 0,$$

$$\frac{\partial \widetilde{L}}{\partial \pi_j} = \sum_n r_{nj}^{(t)} \frac{1}{\pi_j} + \lambda \qquad \stackrel{!}{=} 0.$$

Exam question

Let us define $d_j := \sum_n r_{nj}$. We obtained the following equations

$$\lambda = -\frac{d_j}{\pi_j}, \qquad \sum_j \pi_j = 1.$$

Combining these, we conclude that

$$\pi_j = -\frac{d_j}{\lambda} \Rightarrow 1 = \sum_j \pi_j = -\sum_j \frac{d_j}{\lambda} \Rightarrow \lambda = -\sum_j d_j,$$

and finally

$$\pi_j = \frac{d_j}{\sum_j d_j}.$$

Note that indeed $\pi_j \geq 0$ and $\sum_j \pi_j = 1$.

Exercise 2

Suppose the lifetime of lightbulbs follows exponential distribution with unknown mean θ. In an experiment with N bulbs, the exact lifetimes $Y_1, \ldots, Y_N$ are recorded. In another experiment with M bulbs, we enter the lab at time $t > 0$, and register which of the lightbulbs are still burning (indicator $E_i = 1$), and which have expired ($E_i = 0$). What is the MLE of θ?

Solution: Let $X_1, \ldots, X_M$ be the (unobserved) lifetimes associated with the second experiment, and $Z = \sum_{i=1}^{M} E_i$ the number of lightbulbs in the second experiment that are still alive at time t. Thus, the observed data from both the experiments combined is

$$\mathcal{Y} = (Y_1, \ldots, Y_N, E_1, \ldots, E_M),$$

and the unobserved data is

$$\mathcal{X} = (X_1, \ldots, X_M).$$

Exercise 2

Exponential distribution $p(Y|\theta) = \frac{1}{\theta}\exp(-Y/\theta)$. The log-likelihood is

$$L(\theta) = \log\left(\prod_{j=1}^{N} p(Y_j|\theta) \prod_{j=1}^{M} p(X_j|\theta)\right)$$
$$= -N\log(\theta) - \frac{1}{\theta}\sum_{j=1}^{N} Y_j - M\log(\theta) - \frac{1}{\theta}\sum_{j=1}^{M} X_j.$$

But X is not observed. We replace it with its expected value $\mathbb{E}[X_i|\mathcal{Y}]$,

$$\mathbb{E}[X_i|\mathcal{Y}] = \mathbb{E}[X_i|E_i] = \begin{cases} t+\theta, & \text{for } E_i = 1, \\ \theta - tp, & \text{for } E_i = 0, \end{cases}$$

where $p := \frac{\exp(-t/\theta)}{1-\exp(-t/\theta)}$.

Exercise 2

Then using the current numerical parameter $\theta^{(i-1)}$

$$\begin{aligned}\widetilde{L}^{(i)}(\theta) &= -\log(\theta)(N+M) - \frac{1}{\theta}\sum_{j=1}^{N} Y_j - \frac{1}{\theta}\sum_{j=1}^{M}\mathbb{E}[X_j|E_j] \\ &= -\log(\theta)(N+M) - \frac{1}{\theta}N\overline{Y} \\ &- \frac{1}{\theta}\left(Z(t+\theta^{(i-1)}) + (M-Z)(\theta^{(i-1)} - tp^{(i-1)})\right)\end{aligned}$$

The solution to $(\widetilde{L}^{(i)}(\theta))' = 0$ gives the M-step

$$\theta^{(i)} = \frac{N\overline{Y} + Z(t+\theta^{(i-1)}) + (M-Z)(\theta^{(i-1)} - tp^{(i-1)})}{M+N}.$$

References

- Lecture slides & videos
- *Bernard Flury and Alice Zoppe*. Exercises in EM.
- *Victor Lavrenko*. EM algorithm: how it works.