Chapter 4: Expectation-Maximization

Maarten Jansen

Example: Mixture distribution

1. **Latent (hidden) variable** X
   $X \sim \text{bernoulli}(p)$
2. **Observed variable** Y, dependent on X
   $Y|X=0 \sim f_0;\boldsymbol{\theta}_0$ and $Y|X=1 \sim f_1;\boldsymbol{\theta}_1$

   Marginal pdf
   So, with $q = 1-p$: $\boxed{f_Y(y) = qf_0(y;\boldsymbol{\theta}_0) + pf_1(y;\boldsymbol{\theta}_1)}$

Objective: estimate p, $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$, using maximum likelihood

Maximum likelihood for mixture models

log-likelihood expression

$$\boxed{\text{LL}_Y(p,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) = \sum_{i=1}^n \log\left[qf_0(y_i;\boldsymbol{\theta}_0) + pf_1(y_i;\boldsymbol{\theta}_1)\right]}$$

Marginal likelihood, given the observations $\boldsymbol{Y}$

Problems: Optimization over several parameters, hard to find initial values, uncertain numerical convergence (no guarantee of convexity, etc)

Part of the remedy: Profile likelihood: find expressions for part of the parameters once other parameters are estimated.

Profile likelihood

Example:Box-Cox-transform Suppose positive observations and
$\boxed{Y = (X^\lambda - 1)/\lambda}$

Find λ such that Y is well described by a normal model.

(Remark: since $Y > -1/\lambda$, the normal model is always an approximation. We will assume that an appropriate, sufficiently large μ can be found)

Suppose $Y \sim N(\mu,\sigma^2)$,

then $f_X(x) = f_Y(y(x))\left|\dfrac{dy(x)}{dx}\right| = f_Y(y(x)) \cdot |x|^{\lambda-1}$

hence

$$\text{LL}(\mu,\sigma^2,\lambda) = \sum_{i=1}^n -\frac{(y_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\left[\log(\sigma^2) + \log(2\pi)\right] + (\lambda-1)\sum_{i=1}^n \log(|x_i|)$$

Finding the optimal λ is difficult, as the y_i's depend on λ

But $\dfrac{\partial \text{LL}}{\partial \mu} = 0 \Leftrightarrow \widehat{\mu} = \dfrac{1}{n}\sum_{i=1}^n y_i = \overline{y}$

and $\dfrac{\partial \mathrm{LL}}{\partial \sigma^2} = 0 \Leftrightarrow \widehat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu})^2$

Filling in these variables, we have

$\mathrm{LL}_{\widehat{\mu},\widehat{\sigma}^2}(\lambda) = -\dfrac{n}{2} - \dfrac{n}{2}\log(\widehat{\sigma}^2(\lambda)) - \dfrac{n}{2}\log(2\pi) + (\lambda-1)\sum_{i=1}^{n}\log(|x_i|)$ Minimizing $\mathrm{LL}_{\widehat{\mu},\widehat{\sigma}^2}(\lambda)$ is a one-dimensional problem, which is a fairly easy numerical routine.

Back to log-likelihood for mixture models

1. **Marginal density and log-likelihood**

$\boxed{f_Y(y) = qf_0(y;\boldsymbol{\theta}_0) + pf_1(y;\boldsymbol{\theta}_1)}$ (See slide 1)

$\boxed{\mathrm{LL}_Y(p,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) = \sum_{i=1}^{n}\log\left[qf_0(y_i;\boldsymbol{\theta}_0) + pf_1(y_i;\boldsymbol{\theta}_1)\right]}$ (See slide 2)

2. **Joint density and log-likelihood**

(Suppose we observe X)

$f_{XY}(0,y) = f_0(y;\boldsymbol{\theta}_0)P(X=0)$ and $f_{XY}(1,y) = f_1(y;\boldsymbol{\theta}_1)P(X=1)$

$$\boxed{f_{XY}(x,y) = (1-x)\cdot qf_0(y;\boldsymbol{\theta}_0) + x\cdot pf_1(y;\boldsymbol{\theta}_1) = [qf_0(y;\boldsymbol{\theta}_0)]^{(1-x)}\cdot[pf_1(y;\boldsymbol{\theta}_1)]^x}$$

$$\boxed{\begin{array}{l}\mathrm{LL}_{X,Y}(p,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) = \\ \sum_{i=1}^{n}(1-x_i)\left\{\log[f_0(y_i;\boldsymbol{\theta}_0)] + \log(1-p)\right\} + x_i\left\{\log[f_1(y;\boldsymbol{\theta}_1)] + \log(p)\right\}\end{array}}$$

Optimizing the joint log-likelihood

$$\boxed{\mathrm{LL}_{X,Y}(p,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) = \sum_{i|x_i=0}\left\{\log[f_0(y_i;\boldsymbol{\theta}_0)] + \log(1-p)\right\} + \sum_{i|x_i=1}\left\{\log[f_1(y;\boldsymbol{\theta}_1)] + \log(p)\right\}}$$

Let $N_1 = \#\{i|x_i=1\}$, then this is

$\mathrm{LL}_{X,Y}(p,\boldsymbol{\theta}_0,\boldsymbol{\theta}_1) =$
$\sum_{i|x_i=0}\log[f_0(y_i;\boldsymbol{\theta}_0)] + \sum_{i|x_i=1}\log[f_1(y;\boldsymbol{\theta}_1)] + (n-N_1)\log(1-p) + N_1\log(p)$

Then $\frac{\partial \mathrm{LL}_{X,Y}}{\partial p} = 0 \Leftrightarrow \widehat{p} = N_1/n$

And $\frac{\partial \mathrm{LL}_{X,Y}}{\partial \theta_k}$ involves $\sum_{i|x_i=0}$ only if θ_k belongs to $\boldsymbol{\theta}_0$ (similar for $\boldsymbol{\theta}_1$)

$\rightarrow$ Observing X leads to an optimization which splits into easier problems in a natural way.

Unfortunately, we cannot observe X.
How can we incorporate the benefits from the joint log-likelihood?

More than two possible states of the latent variable

1. **Marginal density and log-likelihood** with $\boxed{p_s = P(X_i = s)}$

$\boxed{f_Y(y) = \sum_{s=1}^{S} p_s f_s(y;\boldsymbol{\theta}_s)}$ $\boxed{\mathrm{LL}_Y(\boldsymbol{p},\boldsymbol{\theta}) = \sum_{i=1}^{n}\log\left[\sum_{s=1}^{S} p_s f_s(y_i;\boldsymbol{\theta}_s)\right]}$

2. **Joint density and log-likelihood**

$$\boxed{f_{XY}(x,y) = \sum_{s=1}^{S}[p_s f_s(y;\boldsymbol{\theta}_s)]^{I(x_i=s)}}$$

The log-likelihood: $\mathrm{LL}_{X,Y}(\boldsymbol{p},\boldsymbol{\theta}) = \sum_{s=1}^{S}\sum_{i|X_i=s}\log(p_s) + \log\left[f_s(y_i;\boldsymbol{\theta}_s)\right]$

becomes: $\boxed{\mathrm{LL}_{X,Y}(\boldsymbol{p},\boldsymbol{\theta}) = \sum_{s=1}^{S}\left\{N_s\log(p_s) + \sum_{i|X_i=s}\log\left[f_s(y_i;\boldsymbol{\theta}_s)\right]\right\}}$

The principle of EM

Objective: Find maximum (log-)likelihood estimator for parameter $\boldsymbol{\theta}$, given observations $\boldsymbol{Y}$

$\mathrm{LL}_Y(\boldsymbol{\theta}) = \log\left[f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\right]$ (marginal log-likelihood)

Latent variable $\boldsymbol{X}$, joint log-likelihood

We assume that the joint log-likelihood is easier to maximize

As X is unobserved, we can consider $\mathrm{LL}_{X,Y}$ as a **random variable**

$\mathrm{LL}_{X,Y}(\boldsymbol{\theta}) = \log\left[f_{\boldsymbol{XY}}(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta})\right]$

We compute the **conditional expectation**, given that the observation $\boldsymbol{Y}=\boldsymbol{y}$:

Marginal ↔ expected joint log-likelihood

$$
\begin{aligned}
E\left[\mathrm{LL}_{X,Y}(\boldsymbol{\theta};\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right] &= E\left\{\log\left[f_{\boldsymbol{XY}}(\boldsymbol{X},\boldsymbol{y};\boldsymbol{\theta})\right]|\boldsymbol{Y}=\boldsymbol{y}\right\} \\
&= E\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta})\cdot f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\right]|\boldsymbol{Y}=\boldsymbol{y}\right\} \\
&= E\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta})\right]+\log\left[f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\right]|\boldsymbol{Y}=\boldsymbol{y}\right\} \\
&= E\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta})\right]|\boldsymbol{Y}=\boldsymbol{y}\right\}+\log\left[f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\right]
\end{aligned}
$$

Conclusion

$$\boxed{E\left[\mathrm{LL}_{X,Y}(\boldsymbol{\theta};\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right] = \mathrm{LL}_Y(\boldsymbol{\theta}) + E\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta})\right]|\boldsymbol{Y}=\boldsymbol{y}\right\}}$$

Interpretation of the result on slide 9

1. The equality contains a term $\mathrm{LL}_Y(\boldsymbol{\theta})$, which is exactly the log-likelihood that we want to maximize
2. The expectation depends on the unknown parameter $\boldsymbol{\theta}$, but the result holds also if we define expectations with <u>any</u> other, incorrect value of $\boldsymbol{\theta}$
3. For the term $E\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta})\right]|\boldsymbol{Y}=\boldsymbol{y}\right\}$ we have the following result

Result from Jensen's inequality

Suppose $\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}\sim f_1$, where $f_1(\boldsymbol{x}) = f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}_1)$, then

$$\boxed{E\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}_2)\right]|\boldsymbol{Y}=\boldsymbol{y}\right\} \le E\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}_1)\right]|\boldsymbol{Y}=\boldsymbol{y}\right\}}$$

Indeed, denote $f_2(\boldsymbol{x}) = f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}_2)$ then we have to prove that

$\boxed{\text{for } X\sim f_1,\ E\left\{\log[f_2(X)]\right\}\le E\left\{\log[f_1(X)]\right\}}$

We have

$$E\left\{\log[f_2(X)]\right\} - E\left\{\log[f_1(X)]\right\} = E\left\{\log\left[\frac{f_2(X)}{f_1(X)}\right]\right\} \le \log\left\{E\left[\frac{f_2(X)}{f_1(X)}\right]\right\}$$

As $X\sim f_1$, and $f_2(x)$ is a density or probability mass function, $E\left[\frac{f_2(X)}{f_1(X)}\right] = 1$

Hence,

$E\left\{\log[f_2(X)]\right\} - E\left\{\log[f_1(X)]\right\} \le 0$

Using the inequality in an iterative scheme

- Suppose we have a current estimator value $\boldsymbol{\theta}_1$
- 1. Compute **expectation** $\boxed{E_{\boldsymbol{\theta}_1}\left[\mathrm{LL}_{X,Y}(\boldsymbol{\theta};\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right]}$
     - for any parameter value $\boldsymbol{\theta}$
     - using $\boldsymbol{\theta}_1$ in the definition of the expectation operator
  2. Find the **new** parameter $\boldsymbol{\theta}_2$ that **maximizes** the expression

Properties of the algorithm

- Because $\boldsymbol{\theta}_2$ maximizes, we have
  $E_{\boldsymbol{\theta}_1}\left[\mathrm{LL}_{X,Y}(\boldsymbol{\theta}_2;\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right] \geq E_{\boldsymbol{\theta}_1}\left[\mathrm{LL}_{X,Y}(\boldsymbol{\theta}_1;\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right]$
- Given the result on slide 11:
  $E_{\boldsymbol{\theta}_1}\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}_2)\right]|\boldsymbol{Y}=\boldsymbol{y}\right\} \leq E_{\boldsymbol{\theta}_1}\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta}_1)\right]|\boldsymbol{Y}=\boldsymbol{y}\right\}$
- Given the result on slide 9:
  $\mathrm{LL}_Y(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}_1}\left[\mathrm{LL}_{X,Y}(\boldsymbol{\theta};\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right] - E_{\boldsymbol{\theta}_1}\left\{\log\left[f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{X}|\boldsymbol{y};\boldsymbol{\theta})\right]|\boldsymbol{Y}=\boldsymbol{y}\right\}$
- All together: $\boxed{\mathrm{LL}_Y(\boldsymbol{\theta}_2) \geq \mathrm{LL}_Y(\boldsymbol{\theta}_1)}$
- Each iteration step produces a new estimator with higher likelihood, but there is **no guarantee** for convergence towards the maximum likelihood

The EM-algorithm for mixture distributions

Suppose $\left\{\begin{array}{rcl} P(X=s) &=& p_s \\ f_Y(y|X=s) &=& f_s(y;\boldsymbol{\theta}_s) \\ f_Y(y) &=& \sum_{s=1}^S p_s f_s(y;\boldsymbol{\theta}_s) \end{array}\right.$

and suppose that after j iterations, we have estimators $\widehat{p}_{s,j}$ and $\widehat{\boldsymbol{\theta}}_{s,j}$

Then we use these parameters to compute $E_j\left[\mathrm{LL}_{X,Y}(\boldsymbol{p},\boldsymbol{\theta};\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right]$

We use the expression on slide 7, which we first rewrite as:

$$\mathrm{LL}_{X,Y}(\boldsymbol{p},\boldsymbol{\theta};\boldsymbol{X}) = \sum_{s=1}^S \left\{N_s\log(p_s) + \sum_{i=1}^n I(X_i=s)\log\left[f_s(y_i;\boldsymbol{\theta}_s)\right]\right\}$$

Then

$$E\left[\mathrm{LL}_{X,Y}(\boldsymbol{p},\boldsymbol{\theta};\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right] = \sum_{s=1}^S \left\{E(N_s|\boldsymbol{Y}=\boldsymbol{y})\log(p_s) + \sum_{i=1}^n P(X_i=s|\boldsymbol{Y}=\boldsymbol{y})\log\left[f_s(y_i;\boldsymbol{\theta}_s)\right]\right\}$$

The Expectation step

The previous expression contains $P(X_i=s|\boldsymbol{Y}=\boldsymbol{y})$ and

$$E(N_s|\boldsymbol{Y}=\boldsymbol{y}) = \sum_{i=1}^n P(X_i=s|\boldsymbol{Y}=\boldsymbol{y})$$

We assume independent observations, so
$P(X_i=s|\boldsymbol{Y}=\boldsymbol{y}) = P(X_i=s|Y_i=y_i)$

Using **Bayes' rule** we find $P(X_i=s|Y_i=y_i) = \dfrac{p_s f_s(y_i;\boldsymbol{\theta}_s)}{\sum_{s=1}^S p_s f_s(y_i;\boldsymbol{\theta}_s)}$

In all these expressions, we use $\widehat{p}_{s,j}$ and $\widehat{\boldsymbol{\theta}}_{s,j}$

The Maximization step (1)

Find p_s and $\boldsymbol{\theta}_s$ that maximize $E\left[\mathrm{LL}_{X,Y}(\boldsymbol{p},\boldsymbol{\theta};\boldsymbol{X})|\boldsymbol{Y}=\boldsymbol{y}\right]$ The expression for the expected joint likelihood (slide 14) has separated terms for p_s and for each of the $\boldsymbol{\theta}_s$

The p_s's are constrained by $\sum_{s=1}^S p_s = 1$, but the $\boldsymbol{\theta}_s$ are independent from each other, unless they contain common parameters

We thus optimize $\frac{\partial}{\partial p_s}\left\{\sum_{s=1}^S E(N_s|\boldsymbol{Y}=\boldsymbol{y})\log(p_s) + \lambda\sum_{s=1}^S p_s\right\}$

from which: $\widehat{p}_{s,j} = \frac{E(N_s|\boldsymbol{Y}=\boldsymbol{y})}{\sum_{s=1}^S E(N_s|\boldsymbol{Y}=\boldsymbol{y})}$

The Maximization step (2)

For $\boldsymbol{\theta}_s$, we can optimize for each s separately $\sum_{i=1}^n I(X_i = s)\log\left[f_s(y_i;\boldsymbol{\theta}_s)\right]$

Unlike the full joint log-likelihood on slide 7, this sum

- involves all observations
- is a weighted sum over the observations

but at least, it only involves one value of s

Alternatives for EM

Another example: absolute observations

Suppose $X \sim N(\mu,\sigma^2)$ but we observe $Y_i = |X_i|$.

Then $X_i = S_iY_i$ where $S_i \sim 2\mathrm{Bernoulli}(p) - 1$ and $p = 1-\Phi(-\mu/\sigma)$

$$\begin{aligned}
&\mathrm{LL}_Y(\mu,\sigma^2)\\
&= \sum_{i=1}^n \log\left\{\exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] + \exp\left[-\frac{(-y_i-\mu)^2}{2\sigma^2}\right]\right\} - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi)
\end{aligned}$$

$$\begin{aligned}
&\mathrm{LL}_{Y,S}(\mu,\sigma^2;\boldsymbol{S}) = \mathrm{LL}_X(\mu,\sigma^2;\boldsymbol{S})\\
&= -\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi)\\
&= -\sum_{i|S_i=1}\frac{(y_i-\mu)^2}{2\sigma^2} - \sum_{i|S_i=-1}\frac{(-y_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi)
\end{aligned}$$

Expectation step

$$E\left[\mathrm{LL}_{Y,S}(\mu,\sigma^2;\boldsymbol{S})|\boldsymbol{Y}=\boldsymbol{y}\right]$$
$$= -\sum_{i=1}^{n} P(S_i=1|Y_i=y_i)\frac{(y_i-\mu)^2}{2\sigma^2} - \sum_{i=1}^{n} P(S_i=-1|Y_i=y_i)\frac{(-y_i-\mu)^2}{2\sigma^2}$$
$$-\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi)$$

Where
$$\begin{aligned} & P(S_i=1|Y_i=y_i) \\ &= \frac{P(S_i=1)\, f_{Y|S_i=1}(y_i;\mu,\sigma^2)}{P(S_i=1)\, f_{Y|S_i=1}(y_i;\mu,\sigma^2) + P(S_i=-1)\, f_{Y|S_i=-1}(y_i;\mu,\sigma^2)} \\ &= \frac{P(S_i=1)\, f_{X|S_i=1}(y_i;\mu,\sigma^2)}{P(S_i=1)\, f_{X|S_i=1}(y_i;\mu,\sigma^2) + P(S_i=-1)\, f_{X|S_i=-1}(-y_i;\mu,\sigma^2)} \end{aligned}$$

or: $P(S_i=1|Y_i=y_i) = \dfrac{f_X(y_i;\mu,\sigma^2)}{f_X(y_i;\mu,\sigma^2)+f_X(-y_i;\mu,\sigma^2)}$

Note: sign probability depends only on local densities $f_X(y_i;\mu,\sigma^2)$ and $f_X(-y_i;\mu,\sigma^2)$; not on global prior probability $P(S_i=1)$.