EM Algorithm

Outline:

- Expectation-maximization (EM) algorithm.
- Examples.

Reading:

A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977

which you can download through the library's web site. You can also read

T.K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Mag.,* vol. 13, pp. 47–60, Nov. 1996

available at IEEE Xplore.

EM Algorithm

EM algorithm provides a systematic approach to finding ML estimates in cases where our model can be formulated in terms of "observed" and "unobserved" (missing) data. Here, "missing data" refers to quantities that, if we could measure them, would allow us to easily estimate the parameters of interest.

EM algorithm can be used in both classical and Bayesian scenarios to maximize likelihoods or posterior probability density/mass functions (pdfs/pmfs), respectively. Here, we focus on the classical scenario; the Bayesian version is similar and will be discussed in handout # 4.

To illustrate the EM approach, we derive it for a class of models known as *mixture models*. These models are specified through the distributions (pdfs, say)[1]

$$f_{\boldsymbol{Y} \,|\, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{y} \,|\, \boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\varphi})$$

and

$$f_{\boldsymbol{U} \,|\, \boldsymbol{\varphi}}(\boldsymbol{u} \,|\, \boldsymbol{\varphi})$$

with unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$. (Note that $\boldsymbol{\varphi}$ parametrizes $f_{\boldsymbol{U} \,|\, \boldsymbol{\varphi}}(\boldsymbol{u} \,|\, \boldsymbol{\varphi})$ and $f_{\boldsymbol{Y} \,|\, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{y} \,|\, \boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\varphi})$, whereas $\boldsymbol{\theta}$ parametrizes $f_{\boldsymbol{Y} \,|\, \boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{y} \,|\, \boldsymbol{u}, \boldsymbol{\theta}, \boldsymbol{\varphi})$ only.) This leads to the marginal

[1] We focus on pdfs without loss of generality.

distribution of $\boldsymbol{y}$ (given the parameters):

$$f_{\boldsymbol{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}) = \int_{\mathcal{U}} \overbrace{f_{\boldsymbol{Y}\,|\,\boldsymbol{U},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\varphi})\, f_{\boldsymbol{U}\,|\,\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{\varphi})}^{f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})}\, d\boldsymbol{u} \tag{1}$$

where $\mathcal{U}$ is the support of $f_{\boldsymbol{U}\,|\,\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{\varphi})$.

Notation:

- $\boldsymbol{y} \equiv$ observed data,
- $\boldsymbol{u} \equiv$ unobserved (missing) data,
- $(\boldsymbol{u}, \boldsymbol{y}) \equiv$ complete data,
- $f_{\boldsymbol{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}) \equiv$ marginal observed data density,
- $f_{\boldsymbol{U}\,|\,\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{\varphi}) \equiv$ marginal unobserved data density,
- $f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}) \equiv$ complete-data density,
- $f_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\varphi}) \equiv$ conditional unobserved-data (missing-data) density.
- $$\mathrm{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}[\cdot\,|\,\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\varphi}]$$

means, in the pdf case,

$$\mathrm{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}[\cdot\,|\,\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\varphi}] = \int_{\mathcal{U}} \cdot f_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\varphi})\,d\boldsymbol{u}$$

and, therefore, at the pth estimate of the parameters:

$$\mathrm{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}[\cdot\,|\,\boldsymbol{y},\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p] = \int_{\mathcal{U}} \cdot f_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{y},\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p)\,d\boldsymbol{u}.$$

Goal: Estimate $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ by maximizing the marginal log-likelihood function of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, i.e. find $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ that maximize

$$L(\boldsymbol{\theta},\boldsymbol{\varphi}) = \ln f_{\boldsymbol{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}). \qquad (2)$$

We assume that

(i) U_i are conditionally independent given $\boldsymbol{\varphi}$, where $\boldsymbol{U} = [U_1,\ldots,U_N]^T$:

$$f_{\boldsymbol{U}\,|\,\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{\varphi}) = \prod_{i=1}^{N} f_{U_i\,|\,\boldsymbol{\varphi}}(u_i\,|\,\boldsymbol{\varphi}) \qquad (3)$$

$\mathcal{U}_i$ is the support of $f_{U_i\,|\,\boldsymbol{\varphi}}(u_i\,|\,\boldsymbol{\varphi})$, and

(ii) Y_i are conditionally independent given the missing data $\boldsymbol{u}$:

$$f_{\boldsymbol{Y}\,|\,\boldsymbol{U},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{u},\boldsymbol{\theta},\boldsymbol{\varphi}) = \prod_{i=1}^{N} f_{Y_i\,|\,U_i,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i\,|\,u_i,\boldsymbol{\theta},\boldsymbol{\varphi}). \tag{4}$$

Thus

$$f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}) = \prod_{i=1}^{N} \underbrace{f_{Y_i\,|\,U_i,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i\,|\,u_i,\boldsymbol{\theta},\boldsymbol{\varphi})\, f_{U_i\,|\,\boldsymbol{\varphi}}(u_i\,|\,\boldsymbol{\varphi})}_{f_{Y_i,U_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i,u_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})} \tag{5}$$

$$f_{\boldsymbol{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$$

$$\overset{\text{see (1)}}{=} \prod_{i=1}^{N} \underbrace{\int_{\mathcal{U}} f_{Y_i\,|\,U_i,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i\,|\,u,\boldsymbol{\theta},\boldsymbol{\varphi})\, f_{U_i\,|\,\boldsymbol{\varphi}}(u\,|\,\boldsymbol{\varphi})\, du}_{f_{Y_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})} \tag{6}$$

$$f_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\varphi}) = \frac{f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})}{f_{\boldsymbol{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})} \tag{7}$$

$$\overset{\text{see (5) and (6)}}{=} \prod_{i=1}^{N} f_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}(u_i\,|\,y_i,\boldsymbol{\theta},\boldsymbol{\varphi}). \tag{8}$$

Comments:

- When we cover graphical models, it will become obvious from the underlying graph that the conditions **(i)** and **(ii)** in

(3) and (4) imply (8). The above derivation of (8) is fairly straightforward as well.

- We have two blocks of parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, because we wish to separate the parameters modeling the distribution of $\boldsymbol{u}$ from the parameters modeling the distribution of $\boldsymbol{y}$ given $\boldsymbol{u}$. Of course, we can also lump the two blocks together into one big vector.

Now,

$$\underbrace{\ln f_{\boldsymbol{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})}_{L(\boldsymbol{\theta},\boldsymbol{\varphi})}$$

$$= \underbrace{\ln f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})}_{\text{complete}} \quad - \underbrace{\ln f_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\varphi})}_{\substack{\text{conditional unobserved}\\ \text{given observed}}}$$

or, summing across observations,

$$L(\boldsymbol{\theta},\boldsymbol{\varphi}) = \sum_i \ln f_{Y_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}) = \sum_i \ln f_{Y_i,U_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i,u_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$$

$$- \sum_i \ln f_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}(u_i\,|\,y_i,\boldsymbol{\theta},\boldsymbol{\varphi}).$$

If we choose values of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, $\boldsymbol{\theta}_p$ and $\boldsymbol{\varphi}_p$ say, we can take the expected value of the above expression with respect to

$f_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{u}\,|\,\boldsymbol{y},\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p)$ to get:

$$\begin{aligned}
&\sum_i \mathrm{E}_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}[\ln f_{Y_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})\,|\,y_i,\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p] \\
&\quad = \sum_i \mathrm{E}_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}[\ln f_{Y_i,U_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i,U_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})\,|\,y_i,\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p] \\
&\qquad - \sum_i \mathrm{E}_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}[\ln f_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}(U_i\,|\,y_i,\boldsymbol{\theta},\boldsymbol{\varphi})\,|\,y_i,\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p].
\end{aligned}$$

Since $L(\boldsymbol{\theta},\boldsymbol{\varphi}) = \ln f_{\boldsymbol{Y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{y}\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})$ in (2) *does not* depend on $\boldsymbol{u}$, it is constant for this expectation. Hence,

$$\begin{aligned}
L(\boldsymbol{\theta},\boldsymbol{\varphi}) &= \sum_i \mathrm{E}_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}[\ln f_{Y_i,U_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi}}(y_i,U_i\,|\,\boldsymbol{\theta},\boldsymbol{\varphi})\,|\,y_i,\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p] \\
&\quad - \sum_i \mathrm{E}_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}[\ln f_{U_i\,|\,Y_i,\boldsymbol{\theta},\boldsymbol{\varphi}}(U_i\,|\,y_i,\boldsymbol{\theta},\boldsymbol{\varphi})\,|\,y_i,\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p]
\end{aligned}$$

which can be written as

$$L(\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{y}) = Q(\boldsymbol{\theta},\boldsymbol{\varphi}\,|\,\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p) - H(\boldsymbol{\theta},\boldsymbol{\varphi}\,|\,\boldsymbol{\theta}_p,\boldsymbol{\varphi}_p). \qquad (9)$$

To clarify, we explicitly write out the Q and H functions (for the pdf case):

$$
\begin{aligned}
& Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \\
& = \sum_i \int_{\mathcal{U}_i} \ln f_{Y_i, U_i \,|\, \boldsymbol{\theta}, \boldsymbol{\varphi}}(y_i, u \,|\, \boldsymbol{\theta}, \boldsymbol{\varphi}) f_{U_i \,|\, Y_i, \boldsymbol{\theta}, \boldsymbol{\varphi}}(u \,|\, y_i, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)\, du \\
& H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \\
& = \sum_i \int_{\mathcal{U}_i} \ln f_{U_i \,|\, Y_i, \boldsymbol{\theta}, \boldsymbol{\varphi}}(u \,|\, y_i, \boldsymbol{\theta}, \boldsymbol{\varphi})\, f_{U_i \,|\, Y_i, \boldsymbol{\theta}, \boldsymbol{\varphi}}(u \,|\, y_i, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)\, du.
\end{aligned}
$$

Recall that our goal is to maximize $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$. The key to the missing information principle is that $H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ is maximized (with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$) by $\boldsymbol{\theta} = \boldsymbol{\theta}_p$ and $\boldsymbol{\varphi} = \boldsymbol{\varphi}_p$:

$$
H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \leq H(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \tag{10}
$$

for any $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ in the parameter space.

Proof. Consider the function

$$
\frac{f_{\boldsymbol{U} \,|\, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{u} \,|\, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\varphi})}{f_{\boldsymbol{U} \,|\, \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{u} \,|\, \boldsymbol{y}, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)}.
$$

By Jensen's inequality (introduced in homework assignment #

1), we have

$$
\begin{aligned}
&\mathrm{E}_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}\Big\{ \ln \Big[\frac{f_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{U} \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\varphi})}{f_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{U} \mid \boldsymbol{y}, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)}\Big] \,\Big|\, \boldsymbol{y}, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p\Big]\Big\} \\
&\quad \leq \ln \mathrm{E}_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}\Big[\frac{f_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{U} \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\varphi})}{f_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{U} \mid \boldsymbol{y}, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)} \,\Big|\, \boldsymbol{y}, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p\Big] \\
&= \ln \int_{\mathcal{U}} \frac{f_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\varphi})}{f_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)} \, f_{\boldsymbol{U} \mid \boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\varphi}}(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \, d\boldsymbol{u} = 0.
\end{aligned}
$$

□

Digression

The result (10) is fundamental to both information theory and statistics. Here is its special case that is perhaps familiar to those in (or who took) information theory.

Proposition 1. **Special case of (10).** *For two pmfs* $\boldsymbol{p} = (p_1, \dots, p_K)$ *and* $\boldsymbol{q} = (q_1, \dots, q_K)$ *on* $\{1, \dots K\}$*, we have*

$$\sum_k p_k \ln p_k \geq \sum_k p_k \ln q_k.$$

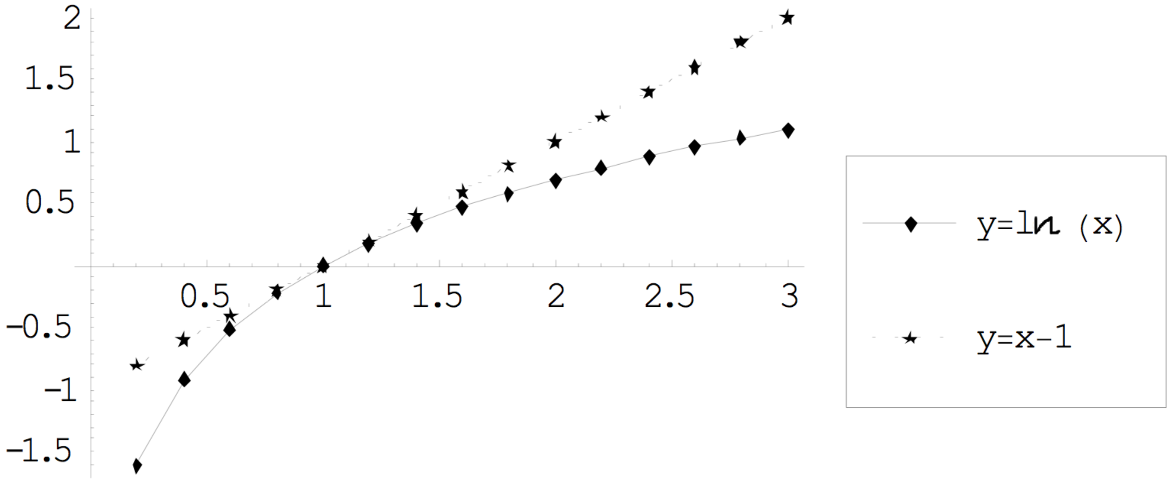
Proof. First, note that

- both sums in the above expression do not change if we restrict them to $\{k : p_k > 0\}$ (as, by convention $0 \ln 0 = 0$) and
- the above result is automatically true if there is some k such that $p_k > 0$ and $q_k = 0$.

Hence, without loss of generality, we assume that all elements of $\boldsymbol{p}$ and $\boldsymbol{q}$ are strictly positive. We wish to show that

$$\sum_k p_k \ln(q_k / p_k) \leq 0.$$

The ln function satisfies the inequality $\ln x \le x - 1$ for all $x > 0$; see the picture and note that the slope of $\ln x$ is 1 at $x = 1$.



So,

$$\sum_k p_k \ln \frac{q_k}{p_k} \le \sum_k p_k \left(\frac{q_k}{p_k} - 1 \right) = \sum_k (q_k - p_k) = 1 - 1 = 0.$$

□

Why is this result so important? It leads to the definition of the *Kullback-Leibler distance* $D(\boldsymbol{p} \,\|\, \boldsymbol{q})$ from one pmf $(\boldsymbol{p})$ to another $(\boldsymbol{p})$:

$$D(\boldsymbol{p} \,\|\, \boldsymbol{q}) = \sum_k p_k \ln \frac{p_k}{q_k}.$$

The above proposition shows that this "distance" is always nonnegative, and it can be shown that $D(\boldsymbol{p} \,\|\, \boldsymbol{q}) = 0$ if and only if $\boldsymbol{p} = \boldsymbol{q}$. An extension to continuous distributions and (10) is straightforward.

EM ALGORITHM (cont.)

We now continue with the EM algorithm derivation. Equation (9) may be written as

$$Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) = L(\boldsymbol{\theta}, \boldsymbol{\varphi}) + \underbrace{H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)}_{\leq H(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)}. \tag{11}$$

Note: If we maximize $Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ for given $\boldsymbol{\theta}_p$ and $\boldsymbol{\varphi}_p$, we are effectively finding a transformation $\mathcal{T}$ that can be written as

$$(\boldsymbol{\theta}_{p+1}, \boldsymbol{\varphi}_{p+1}) = \mathcal{T}(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$$

where $(\boldsymbol{\theta}_{p+1}, \boldsymbol{\varphi}_{p+1})$ is the value of the pair $(\boldsymbol{\theta}, \boldsymbol{\varphi})$ that maximizes $Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$.

In this context, we define a *fixed-point transformation* as a transformation that maps $(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ onto itself, i.e. $\mathcal{T}_\star$ is a fixed-point transformation if

$$\mathcal{T}_\star(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) = (\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p).$$

Proposition 2. ***Missing Information Principle.*** *For a transformation defined by maximizing*

$$Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$$

on the right-hand side of (11) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ and if $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$ and $H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ are differentiable,

(i) *the pair $(\boldsymbol{\theta}, \boldsymbol{\varphi})$ that maximizes $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$ [i.e. the ML estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$] constitutes a fixed-point transformation.*

(ii) *any fixed-point transformation is either the ML estimate or a stationary point of $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$.*

Proof. (heuristic)

(i) If the ML estimate $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\varphi}}) = (\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$, i.e.

$$Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\varphi}}) = L(\boldsymbol{\theta}, \boldsymbol{\varphi}) + H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\varphi}})$$

then $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\varphi}})$ maximizes both terms on the right-hand side of the above expression simultaneously (and hence also their sum, Q).

(ii) For *any* $(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$,

$$\mathcal{T}(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) = (\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \qquad (12)$$

maximizes $H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$. By the assumptions, both $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$ and $H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ are differentiable, implying that (12) cannot maximize

$$Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) = L(\boldsymbol{\theta}, \boldsymbol{\varphi}) + H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$$

unless it is also a maximum (a local maximum, in general) or a stationary point of $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$. □

To summarize:

E Step (Expectation). Compute $Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ where, for given $\boldsymbol{\theta}_p$ and $\boldsymbol{\varphi}_p$,

$$\begin{aligned} & Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \\ & \qquad = \sum_i \mathrm{E}_{U_i \,|\, Y_i, \boldsymbol{\theta}, \boldsymbol{\varphi}} \Big[\ln f_{Y_i, U_i}(y_i, U_i \,|\, \boldsymbol{\theta}, \boldsymbol{\varphi}) \,\Big|\, y_i, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p \Big]. \end{aligned}$$

M Step (Maximization). Maximize $Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$:

$$(\boldsymbol{\theta}_{p+1}, \boldsymbol{\varphi}_{p+1}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\varphi}} Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p).$$

Iterate between the E and M steps until convergence i.e. until reaching a fixed-point transformation. Then, based on

Proposition 2, we have reached either the ML estimate or a stationary point of $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$, if $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$ and $H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ are differentiable.

An important property of the EM algorithm is that the observed-data log-likelihood function $L(\boldsymbol{\theta}, \boldsymbol{\varphi})$ increases at each iteration or, more precisely, does not decrease. The likelihood-climbing property, however, does not guarantee convergence, in general. (This property carries over to the Bayesian scenario, where it can be called posterior climbing, since, in the Bayesian scenario, the goal is to maximize the posterior pdf or pmf.)

We now show the likelihood-climbing property. Recall (9):

$$L(\boldsymbol{\theta}, \boldsymbol{\varphi}) = Q(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) - H(\boldsymbol{\theta}, \boldsymbol{\varphi} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p).$$

The previous values $(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)$ and updated values $(\boldsymbol{\theta}_{p+1}, \boldsymbol{\varphi}_{p+1})$ satisfy

$$\begin{aligned} & L(\boldsymbol{\theta}_{p+1}, \boldsymbol{\varphi}_{p+1}) - L(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) \\ & \quad = \underbrace{Q(\boldsymbol{\theta}_{p+1}, \boldsymbol{\varphi}_{p+1} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) - Q(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)}_{\geq 0, \text{ since } Q \text{ is maximized}} \\ & \quad + \underbrace{H(\boldsymbol{\theta}_p, \boldsymbol{\varphi}_p \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p) - H(\boldsymbol{\theta}_{p+1}, \boldsymbol{\varphi}_{p+1} \,|\, \boldsymbol{\theta}_p, \boldsymbol{\varphi}_p)}_{\geq 0, \text{ by (10)}} \geq 0. \end{aligned}$$

Another important property of the EM algorithm is that it handles parameter constraints automatically, e.g. estimates of probabilities will be between zero and one and estimates of variances will be nonnegative. This is because each M step produces an ML-type estimate (for the complete data).

Main Disadvantages of the EM Algorithm, compared with the competing Newton-type algorithms:

- The convergence can be very slow. (Hybrid and other approaches have been proposed in the literature to improve convergence speed. One such approach is called PX-EM.)

- There is no immediate way to assess accuracy of EM-based estimators, e.g. to compute CRB.

Canonical Exponential Family for I.I.D. Complete Data

Consider the scenario where the complete data (Y_i, U_i) are independent given the canonical parameters $\boldsymbol{\eta}$, following pdfs (or pmfs) $f_{Y_i,U_i \,|\, \boldsymbol{\eta}}(y_i, u_i \,|\, \boldsymbol{\eta})$ that belong to the canonical exponential family:

$$f_{Y_i,U_i \,|\, \boldsymbol{\eta}}(y_i, u_i \,|\, \boldsymbol{\eta}) = h(y_i, u_i) \exp[\boldsymbol{\eta}^T \boldsymbol{T}_i(y_i, u_i) - A(\boldsymbol{\eta})] \quad (13)$$

and, therefore,

$$\begin{aligned} & f_{\boldsymbol{Y},\boldsymbol{U} \,|\, \boldsymbol{\eta}}(y_i, u_i \,|\, \boldsymbol{\eta}) \\ &= \Big[\prod_i h(y_i, u_i)\Big] \cdot \exp\Big\{\boldsymbol{\eta}^T \,[\sum_i \boldsymbol{T}_i(y_i, u_i)] - N\, A(\boldsymbol{\eta})\Big\} \end{aligned}$$

and

$$Q(\boldsymbol{\eta} \,|\, \boldsymbol{\eta}_p) = \sum_i \mathrm{E}_{U_i \,|\, Y_i, \boldsymbol{\eta}}\{\boldsymbol{\eta}^T \boldsymbol{T}_i(y_i, U_i) - A(\boldsymbol{\eta}) + \ln h(y_i, U_i) \,|\, y_i, \boldsymbol{\eta}_p\} \quad (14)$$

which is maximized with respect to $\boldsymbol{\eta}$ for

$$\sum_i \underbrace{\mathrm{E}_{U_i \,|\, Y_i, \boldsymbol{\eta}}[\boldsymbol{T}_i(y_i, U_i) \,|\, y_i, \boldsymbol{\eta}_p]}_{\mathcal{T}_i^{(p+1)}} = N \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}}.$$

To obtain this equation, take the derivative of (14) with respect to $\boldsymbol{\eta}$ and set it to zero. But, solving the system

$$\sum_i \boldsymbol{T}_i(y_i, u_i) = N \frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \qquad (15)$$

yields the ML estimates of $\boldsymbol{\eta}$ under the complete-data model. (For $N = 1$, (15) reduces to Theorem 2.3.1 in Bickel & Doksum. Of course, the regularity conditions stated in this theorem also need to be satisfied.)

To summarize: When complete data fits the canonical exponential-family model, the EM algorithm is easily derived as follows.

- The expectation (E) step is reduced to computing conditional expectations of the complete-data natural sufficient statistics $[\sum_{i=1}^N \boldsymbol{T}_i(y_i, u_i)]$, given the observed data and parameter estimates from the previous iteration.

- The maximization (M) step is reduced to finding the expressions for the complete-data ML estimates of the unknown parameters $(\boldsymbol{\theta})$, and replacing the *complete-data natural sufficient statistics* $[\sum_{i=1}^N \boldsymbol{T}_i(y_i, u_i)]$ that occur in these expressions with their conditional expectations computed in the E step.

Gaussian Toy Example: EM Solution

Suppose we wish to find $\widehat{\theta}_{\mathrm{ML}}$ for the model

$$f_{Y \mid \theta}(y \mid \theta) = \mathcal{N}(y \mid \theta, 1 + \sigma^2) \tag{16}$$

where $\sigma^2 > 0$ is a known constant and θ is the unknown parameter. We know the answer:

$$\widehat{\theta}_{\mathrm{ML}} = y \quad \text{(trivial)}.$$

However, it is instructional to derive the EM algorithm for this toy problem.

We invent the missing data U as follows:

$$Y = U + W$$

where

$$\{U \mid \theta\} \sim \mathcal{N}(\theta, \sigma^2), \quad W \sim \mathcal{N}(0, 1)$$

and U and W are conditionally independent given θ. Hence,

$$f_{Y \mid U, \theta}(y \mid u, \theta) = \mathcal{N}(y \mid u, 1), \quad f_{U \mid \theta}(u \mid \theta) = \mathcal{N}(u \mid \theta, \sigma^2)$$

and the complete-data log-likelihood function is

$$
\begin{aligned}
f_{Y,U\,|\,\theta}(y,u\,|\,\theta) &= \frac{1}{\sqrt{2\,\pi}}\exp[-(y-u)^2/2] \\
&\quad \cdot\frac{1}{\sqrt{2\,\pi\,\sigma^2}}\exp[-(u-\theta)^2/(2\,\sigma^2)] \qquad (17)
\end{aligned}
$$

which is a joint Gaussian pdf for Y and U. The pdf of Y (given θ) is Gaussian:

$$f_{Y\,|\,\theta}(y\,|\,\theta)=\mathcal{N}(\theta,1+\sigma^2)$$

the same as the original model in (16). Note that

$$(\widehat{\theta}_{\mathrm{ML}})_{\mathrm{complete\,data}}=u$$

is the complete-data ML estimate of θ obtained by maximizing (17). Now, the complete-data model (17) belongs to the one-parameter exponential family and the complete-data natural sufficient statistic is u, so our EM iteration reduces to updating u:

$$\theta_{p+1}=u_{p+1}=\mathrm{E}_{\,U\,|\,Y,\theta}[U\,|\,y,\theta_p].$$

Here, finding $f_{U\,|\,Y,\theta}(u\,|\,y,\theta_p)$ is a basic Bayesian exercise (which we will do for a more general case at the very beginning of handout # 4) and we practiced it in homework assignment

1:

$$f_{U\,|\,Y,\theta}(u\,|\,y,\theta_p) = \mathcal{N}\Big(u\,\Big|\,\frac{y+\theta_p/\sigma^2}{1+1/\sigma^2}, \frac{1}{1+1/\sigma^2}\Big). \qquad (18)$$

Therefore

$$\mathrm{E}_{\,U\,|\,Y,\theta}[U\,|\,y,\theta_p] = \frac{y+\theta_p/\sigma^2}{1+1/\sigma^2}$$

and, consequently, the EM iteration is

$$\theta_{p+1} = \frac{y+\theta_p/\sigma^2}{1+1/\sigma^2}.$$

A (more) detailed EM algorithm derivation for our Gaussian toy example: Note that we have only one parameter here, θ. We first write down the logarithm of the joint pdf of the observed and unobserved data: from (17), we have

$$\begin{aligned}
\ln f_{Y,U\,|\,\theta}(y,u\,|\,\theta) &= \underbrace{\mathrm{const}}_{\text{not a function of } \theta} - \frac{(u-\theta)^2}{2\,\sigma^2} \\
&= \underbrace{\mathrm{const}}_{\text{not a function of } \theta} - \frac{1}{2\,\sigma^2}\,\theta^2 + \frac{u}{\sigma^2}\,\theta. \qquad (19)
\end{aligned}$$

Why can we ignore terms in the above expression that do not contain θ? Because the maximization in the M step will be with respect to θ. Now, we need the find the *conditional pdf* *of the unobserved data given the observed data, evaluated at $\theta = \theta_p$* [see (17)]:

$$
\begin{aligned}
f_{U\,|\,Y,\theta}(u\,|\,y,\theta_p) &\propto f_{Y,U\,|\,\theta}(y,u\,|\,\theta_p)\\
&\propto \exp[-(y-u)^2/2]\cdot\exp[-(u-\theta_p)^2/(2\,\sigma^2)]\\
&\propto \exp(y\,u-\tfrac{1}{2}\,u^2)\cdot\exp\Big(\frac{\theta_p}{\sigma^2}\,u-\frac{1}{2\,\sigma^2}\,u^2\Big)\\
&\overset{\substack{\text{combine the linear}\\ \text{and quadratic terms}}}{=} \exp\Big[\Big(y+\frac{\theta_p}{\sigma^2}\Big)\,u-\tfrac{1}{2}\,\Big(1+\frac{1}{\sigma^2}\Big)\,u^2\Big]\\
&\overset{\substack{\text{(look up the table}\\ \text{of distributions)}}}{\text{is the kernel of}} \mathcal{N}\Big(u\,\Big|\,\frac{y+\theta_p/\sigma^2}{1+1/\sigma^2},\frac{1}{1+1/\sigma^2}\Big).
\end{aligned}
\tag{20}
$$

We can derive this result using the conditional pdf expressions for Gaussian pdfs in handout # 0b. The derivation presented above uses the Bayesian machinery and $\propto$ notation, which we will master soon, in the next few lectures.

Hence, *knowing the joint log pdf of the observed and unobserved data in (17) is key to all steps of our EM algorithm derivation*. We now use (19) and (20) to derive the E and M steps for this example:

$$
\begin{aligned}
Q(\theta \mid \theta_p) &= \mathrm{E}_{U \mid Y,\theta}\big[\ln f_{Y,U \mid \theta}(y, U \mid \theta) \mid y, \theta_p\big] \\
&= \underbrace{\mathrm{const}}_{\text{not a function of } \theta} \\
&\qquad \underbrace{-\frac{1}{2\,\sigma^2}\,\theta^2}_{\text{no } U\text{, expectation disappears}} \qquad + \frac{\mathrm{E}_{U \mid Y,\theta}[U \mid y, \theta_p]}{\sigma^2} \cdot \theta
\end{aligned}
$$

which is a quadratic form of θ, and is easily maximized with respect to θ, yielding the following M step:

$$
\theta_{p+1} = \mathrm{E}_{U \mid Y,\theta}[U \mid y, \theta_p] = \frac{y + \theta_p/\sigma^2}{1 + 1/\sigma^2}.
$$

A More Useful Gaussian Example: EM Solution

Consider the model

$$\boldsymbol{Y} = a\,\boldsymbol{U} + \boldsymbol{W}$$

where

$$\{\boldsymbol{U} \,|\, a\} \sim \mathcal{N}(\mathbf{0}, C), \quad \boldsymbol{W} \sim \mathcal{N}(\mathbf{0}, \sigma^2\, I)$$

$\boldsymbol{U}$ and $\boldsymbol{W}$ are independent given $\boldsymbol{\theta}$, C is a known covariance matrix, and

$$\boldsymbol{\theta} = [a, \sigma^2]^T$$

is the vector of unknown parameters. Therefore,

$$f_{\boldsymbol{Y} \,|\, \boldsymbol{U}, \boldsymbol{\theta}}(\boldsymbol{y} \,|\, \boldsymbol{u}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y} \,|\, a\,\boldsymbol{u}, \sigma^2\, I)$$

and the marginal likelihood function of $\boldsymbol{\theta}$ is

$$f_{\boldsymbol{Y} \,|\, \boldsymbol{\theta}}(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y} \,|\, \mathbf{0}, a^2\, C + \sigma^2\, I).$$

We wish to find the marginal ML estimate of $\boldsymbol{\theta}$ by maximizing $f_{\boldsymbol{Y} \,|\, \boldsymbol{\theta}}(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. There is no closed-for solution to this problem. A Newton-type iteration is an option, but requires good matrix-differentiation skills and making sure that the estimate of σ^2 is nonnegative.

Note that a is not identifiable: we cannot uniquely determine its sign. But, a^2 is identifiable.

For missing data $\boldsymbol{U}$, we have the complete-data log-likelihood function:

$$f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta}) = \frac{1}{(2\,\pi\,\sigma^2)^{N/2}}\exp[-\|\boldsymbol{y}-a\,\boldsymbol{u}\|_{\ell_2}^2/(2\,\sigma^2)] \cdot \frac{1}{\sqrt{|2\,\pi\,C|}}\exp(-\tfrac{1}{2}\,\boldsymbol{u}^T\,C^{-1}\,\boldsymbol{u}) \qquad (21)$$

and, therefore,

$$\begin{aligned} f_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{u}\,|\,\boldsymbol{y},\boldsymbol{\theta}_p) &\propto f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta}_p) \\ &\propto \frac{1}{(2\,\pi\,\sigma_p^2)^{N/2}}\exp[-\|\boldsymbol{y}-a_p\,\boldsymbol{u}\|_{\ell_2}^2/(2\,\sigma_p^2)]\,\exp(-\tfrac{1}{2}\,\boldsymbol{u}^T\,C\,\boldsymbol{u}) \\ &\propto \exp[(a_p/\sigma_p^2)\,\boldsymbol{y}^T\,\boldsymbol{u}]\,\exp\{-\tfrac{1}{2}\,\boldsymbol{u}^T\,[(a_p^2/\sigma_p^2)\,I+C^{-1}]\,\boldsymbol{u}\} \end{aligned}$$

is the kernel of $\mathcal{N}\Big(\boldsymbol{u}\,|\,\underbrace{(a_p/\sigma_p^2)\,\Sigma_p\,\boldsymbol{y}}_{\triangleq\,\boldsymbol{\mu}_p},\,\Sigma_p\Big)$

where

$$\Sigma_p = \mathrm{cov}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{U}\,|\,\boldsymbol{y},\boldsymbol{\theta}_p) = [(a_p^2/\sigma_p^2)\,I+C^{-1}]^{-1}.$$

Here,

$$\|\boldsymbol{x}\|_{\ell_2}^2 \triangleq \boldsymbol{x}^T\,\boldsymbol{x} \quad \ell_2 \text{ (Euclidean) norm}$$

for an arbitrary real-valued vector $\boldsymbol{x}$.

Now,

$$\ln f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{u}\,|\,\boldsymbol{\theta}) = \underbrace{\text{const}}_{\text{not a function of } \theta} -\tfrac{1}{2}\,N\,\ln(\sigma^2) - \boldsymbol{y}^T\,\boldsymbol{y}/(2\,\sigma^2)$$

$$+(a/\sigma^2)\,\boldsymbol{y}^T\,\boldsymbol{u} - \tfrac{1}{2}\,\boldsymbol{u}^T\,[(a^2/\sigma^2)\,I + C^{-1}]\,\boldsymbol{u}$$

$$= \underbrace{\text{const}}_{\text{not a function of } \theta} -\tfrac{1}{2}\,N\,\ln(\sigma^2) - \boldsymbol{y}^T\,\boldsymbol{y}/(2\,\sigma^2) + (a/\sigma^2)\,\boldsymbol{y}^T\,\boldsymbol{u}$$

$$-\tfrac{1}{2}\,\text{tr}\{[(a^2/\sigma^2)\,I + C^{-1}]\,\boldsymbol{u}\,\boldsymbol{u}^T\}$$

$$= \underbrace{\text{const}}_{\text{not a function of } \theta} -\tfrac{1}{2}\,N\,\ln(\sigma^2) - \boldsymbol{y}^T\,\boldsymbol{y}/(2\,\sigma^2) + (a/\sigma^2)\,\boldsymbol{y}^T\,\boldsymbol{u}$$

$$-\tfrac{1}{2}\,\text{tr}\{[(a^2/\sigma^2)\,I]\,\boldsymbol{u}\,\boldsymbol{u}^T\}$$

$$= \underbrace{\text{const}}_{\text{not a function of } \theta} -\tfrac{1}{2}\,N\,\ln(\sigma^2) - \boldsymbol{y}^T\,\boldsymbol{y}/(2\,\sigma^2) + (a/\sigma^2)\,\boldsymbol{y}^T\,\boldsymbol{u}$$

$$-\tfrac{1}{2}\,(a^2/\sigma^2)\,\text{tr}(\boldsymbol{u}\,\boldsymbol{u}^T)$$

and

$$Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_p) = \text{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}\big[\ln f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{U}\,|\,\boldsymbol{\theta})\,|\,\boldsymbol{y},\boldsymbol{\theta}_p\big]$$

$$= \underbrace{\text{const}}_{\text{not a function of } \theta} -\tfrac{1}{2}\,N\,\ln(\sigma^2)$$

$$-\boldsymbol{y}^T\,\boldsymbol{y}/(2\,\sigma^2) + (a/\sigma^2)\,\boldsymbol{y}^T\,\text{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}[\boldsymbol{U}\,|\,\boldsymbol{y},\boldsymbol{\theta}_p]$$

$$-\tfrac{1}{2}\,(a^2/\sigma^2)\,\text{tr}\{\text{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{U}\,\boldsymbol{U}^T\,|\,\boldsymbol{y},\boldsymbol{\theta}_p)\}.$$

Since

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{U}\,|\,\boldsymbol{y},\boldsymbol{\theta}_p) &= (a_p/\sigma_p^2)\,\Sigma_p\,\boldsymbol{y} \overset{\triangle}{=} \boldsymbol{\mu}_p \\
\mathrm{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{U}\,\boldsymbol{U}^T\,|\,\boldsymbol{y},\boldsymbol{\theta}_p) &= \boldsymbol{\mu}_p^T\,\boldsymbol{\mu}_p + \underbrace{\mathrm{cov}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{U}\,|\,\boldsymbol{y},\boldsymbol{\theta}_p)}_{\Sigma_p}
\end{aligned}$$

we have

$$\begin{aligned}
Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_p) &= \mathrm{E}_{\boldsymbol{U}\,|\,\boldsymbol{Y},\boldsymbol{\theta}}\big[\ln f_{\boldsymbol{Y},\boldsymbol{U}\,|\,\boldsymbol{\theta}}(\boldsymbol{y},\boldsymbol{U}\,|\,\boldsymbol{\theta})\,|\,\boldsymbol{y},\boldsymbol{\theta}_p\big] \\
&= \underbrace{\mathrm{const}}_{\text{not a function of } \boldsymbol{\theta}} - \tfrac{1}{2}\,N\,\ln(\sigma^2) - \boldsymbol{y}^T\,\boldsymbol{y}/(2\,\sigma^2) + (a/\sigma^2)\,\boldsymbol{y}^T\,\boldsymbol{\mu}_p \\
&\quad - \tfrac{1}{2}\,(a^2/\sigma^2)\,\mathrm{tr}[(\boldsymbol{\mu}_p\,\boldsymbol{\mu}_p^T + \Sigma_p)] \\
&= \underbrace{\mathrm{const}}_{\text{not a function of } \boldsymbol{\theta}} - \tfrac{1}{2}\,N\,\ln(\sigma^2) - \boldsymbol{y}^T\,\boldsymbol{y}/(2\,\sigma^2) + (a/\sigma^2)\,\boldsymbol{y}^T\,\boldsymbol{\mu}_p \\
&\quad - \tfrac{1}{2}\,(a^2/\sigma^2)\,\boldsymbol{\mu}_p^T\,\boldsymbol{\mu}_p - \tfrac{1}{2}\,(a^2/\sigma^2)\,\mathrm{tr}(\Sigma_p) \\
&= \underbrace{\mathrm{const}}_{\text{not a function of } \boldsymbol{\theta}} - \tfrac{1}{2}\,N\,\ln(\sigma^2) \\
&\quad - \tfrac{1}{2}\,\|\boldsymbol{y} - a\,\boldsymbol{\mu}_p\|_{\ell_2}^2/\sigma^2 - \tfrac{1}{2}\,(a^2/\sigma^2)\,\mathrm{tr}(\Sigma_p)
\end{aligned}$$

yielding

$$\begin{aligned}
a_{p+1} &= \frac{\boldsymbol{y}^T\,\boldsymbol{\mu}_p}{\boldsymbol{\mu}_p^T\,\boldsymbol{\mu}_p + \mathrm{tr}(\Sigma_p)} = \frac{\boldsymbol{y}^T\,\boldsymbol{\mu}_p}{\|\boldsymbol{\mu}_p\|_{\ell_2}^2 + \mathrm{tr}(\Sigma_p)} \\
\sigma_{p+1}^2 &= \frac{\|\boldsymbol{y} - a_p\,\boldsymbol{\mu}_p\|_{\ell_2}^2 + a_p^2}{N}
\end{aligned}$$

or, perhaps,

$$\begin{aligned} a_{p+1} &= \frac{\boldsymbol{y}^T\,\boldsymbol{\mu}_p}{\boldsymbol{\mu}_p^T\,\boldsymbol{\mu}_p + \mathrm{tr}(\Sigma_p)} = \frac{\boldsymbol{y}^T\,\boldsymbol{\mu}_p}{\|\boldsymbol{\mu}_p\|_{\ell_2}^2 + \mathrm{tr}(\Sigma_p)} \\ \sigma_{p+1}^2 &= \frac{\|\boldsymbol{y} - a_{p+1}\,\boldsymbol{\mu}_p\|_{\ell_2}^2 + a_{p+1}^2}{N}. \end{aligned}$$

Example: Semi-blind Channel Estimation

This example is a special case of

A. Dogandžić, W. Mo, and Z.D. Wang, "Semi-blind SIMO flat-fading channel estimation in unknown spatially correlated noise using the EM algorithm," *IEEE Trans. Signal Processing,* vol. 52, pp. 1791–1797, Jun. 2004

see also references therein. It also fits the mixture model that we chose to illustrate the EM algorithm.

Measurement Model:

We observe $y(t)$, modeled as

$$Y(t) = h\,u(t) + W(t) \quad t = 1, 2, \ldots, N$$

where

- h is unknown channel,
- $u(t)$ is an unknown symbol received by the array at time t (missing data),
- $W(t)$ is zero-mean additive white Gaussian noise with unknown variance σ^2,

- N is the number of snapshots (block size).

The symbols $u(t),\ t = 1, 2, \ldots, N$

- belong to a *known* M-ary constant-modulus constellation $\{u_1, u_2, \ldots, u_M\}$, with

$$|u_m| = 1 \quad m = 1, 2, \ldots, M$$

- are modeled as i.i.d. random variables with probability mass function

$$p_U(u(t)) = \frac{1}{M}\, i(u(t))$$

where

$$i(u) = \begin{cases} 1, & u \in \{u_1, u_2, \ldots, u_M\} \\ 0, & \text{otherwise} \end{cases} .$$

Note: an extension to arbitrary known prior symbol probabilities is trivial.

Training Symbols

To allow unique estimation of the channel h, assume that $N_{\rm T}$ *known* (training) symbols

$$u_{\rm T}(\tau) \quad \tau = 1, 2, \ldots, N_{\rm T}$$

are embedded in the transmission scheme and denote the corresponding snapshots received by the array as

$$y_{\mathrm{T}}(\tau) \quad \tau = 1, 2, \ldots, N_{\mathrm{T}}.$$

Then

$$y_{\mathrm{T}}(\tau) = h\, u_{\mathrm{T}}(\tau) + W(\tau) \quad \tau = 1, 2, \ldots, N_{\mathrm{T}}.$$

Summary of the Model

We know

- snapshots $y(t)$ $t = 1, 2, \ldots, N$ and $y_{\mathrm{T}}(\tau)$ $\tau = 1, 2, \ldots, N_{\mathrm{T}}$, and
- training symbols $u_{\mathrm{T}}(\tau)$ $\tau = 1, 2, \ldots, N_{\mathrm{T}}$.

The *unknown* symbols $u(t)$ $t = 1, 2, \ldots, N$

- belong to a (known) constant-modulus constellation with

  $$|u_m| = 1 \quad m = 1, 2, \ldots, M$$

  and

- are equiprobable.

Goal: Using the above information, estimate the channel h and noise variance σ^2. Hence, the unknown parameter vector is

$$\boldsymbol{\theta} = [h, \sigma^2]^T.$$

ML Estimation

We treat the unknown symbols $u(t)\, t = 1, 2, \ldots, N$ as *unobserved* (missing) data and apply the EM algorithm. Given $u(t)$ and h, the measurements $y(t)$ are distributed as

$$f_{Y \,|\, U, \boldsymbol{\theta}}(y(t) \,|\, u(t), \boldsymbol{\theta}) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left\{ - \frac{[y(t) - h\,u(t)]^2}{2\,\sigma^2} \right\}$$

for $t = 1, \ldots, N$. Similarly, the measurements $y_{\mathrm{T}}(\tau)$ containing the training sequence are distributed as

$$f_{y_{\mathrm{T}} \,|\, U_{\mathrm{T}}, \boldsymbol{\theta}}(y_{\mathrm{T}}(\tau) \,|\, u_{\mathrm{T}}(\tau), \boldsymbol{\theta}) \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left\{ - \frac{[y_{\mathrm{T}}(\tau) - h\,u_{\mathrm{T}}(\tau)]^2}{2\,\sigma^2} \right\}$$

for $\tau = 1, \ldots, N_{\mathrm{T}}$.

Complete-data Likelihood and Sufficient Statistics

The *complete-data likelihood function* is the joint distribution of $\boldsymbol{y}(t)$, $u(t)$ (for $t = 1, 2, \ldots, N$), and $y_{\mathrm{T}}(\tau)$ (for $\tau = 1, 2, \ldots, N_{\mathrm{T}}$) given $\boldsymbol{\theta}$:

$$\Big[\prod_{t=1}^{N} p_U(u(t))\, f_{Y\,|\,U,\boldsymbol{\theta}}(y(t)\,|\,u(t),\boldsymbol{\theta})\Big] \cdot \prod_{\tau=1}^{N_{\mathrm{T}}} p_{y_{\mathrm{T}}\,|\,u_{\mathrm{T}},\boldsymbol{\theta}}(\boldsymbol{y}_{\mathrm{T}}(\tau)|u_{\mathrm{T}}(\tau),\boldsymbol{\theta})$$

$$\begin{aligned} = \quad & [\prod_{t=1}^{N} p_U(u(t))] \cdot (2\,\pi\,\sigma^2)^{-N/2} \cdot \exp(-\frac{N\,|h|^2}{2\,\sigma^2}) \\ & \cdot \exp\Big[\frac{N+N_{\mathrm{T}}}{\sigma^2} T_1(\boldsymbol{y},\boldsymbol{U})\,h - \frac{N+N_{\mathrm{T}}}{2\,\sigma^2} T_2(\boldsymbol{y})\Big] \end{aligned}$$

which belongs to the exponential family of distributions. Here,

$$\begin{aligned} T_1(\boldsymbol{y},\boldsymbol{U}) &= \frac{1}{N+N_{\mathrm{T}}} \Big\{[\sum_{t=1}^{N} y(t)\,U(t)] + [\sum_{\tau=1}^{N_{\mathrm{T}}} y_{\mathrm{T}}(\tau)\,u_{\mathrm{T}}(\tau)]\Big\} \\ T_2(\boldsymbol{y}) &= \frac{1}{N+N_{\mathrm{T}}} \Big\{[\sum_{t=1}^{N} y^2(t)] + [\sum_{\tau=1}^{N_{\mathrm{T}}} y_{\mathrm{T}}^2(\tau)]\Big\}. \end{aligned}$$

are the *complete-data natural sufficient statistics* for $\boldsymbol{\theta}$. We also have

$$p_{U(t)\,|\,y(t),\boldsymbol{\theta}}(u_m\,|\,y(t),\boldsymbol{\theta}_p) = \frac{\exp\{-\frac{1}{2\,\sigma_p^2}\,[y(t)-h_p\,u_m]^2\}}{\sum_{n=1}^{N}\exp\{-\frac{1}{2\,\sigma_p^2}\,[y(t)-h_p\,u_n]^2\}}.$$

EM Algorithm: Apply the recipe for exponential family.

- **The Expectation (E) Step:**
  - compute conditional expectations of the complete-data natural sufficient statistics, given $\boldsymbol{\theta} = \boldsymbol{\theta}_p$ and the observed data $y(t),\ t = 1,\ldots,N$ and $y_{\mathrm{T}}(\tau),\ \tau = 1,\ldots,N_{\mathrm{T}}$.
- **The Maximization (M) Step:**
  - find the expressions for the complete-data ML estimates of h and σ^2, and
  - replace the complete-data natural sufficient statistics that occur in these expressions with their conditional expectations computed in the E step.

Note: the complete-data ML estimates of h and σ^2 are

$$\begin{aligned}\widehat{h}(\boldsymbol{y},\boldsymbol{U}) &= T_1(\boldsymbol{y},\boldsymbol{U})\\ \widehat{\sigma}^2(\boldsymbol{y},\boldsymbol{U}) &= T_2(\boldsymbol{y}) - \widehat{h}^2(\boldsymbol{y},\boldsymbol{U}).\end{aligned}$$

EM Algorithm

EM Step 1:

$$h^{(k+1)} = \frac{1}{N + N_{\mathrm{T}}}$$

$$\cdot \left[\sum_{t=1}^{N} \boldsymbol{y}(t) \overbrace{\frac{\sum_{m=1}^{M} u_m \exp[y(t)\, h^{(k)}\, u_m/(\sigma^2)^{(k)}]}{\sum_{n=1}^{M} \exp[y(t)\, h^{(k)}\, u_n/(\sigma^2)^{(k)}]}}^{\frac{\sum_{m=1}^{M} u_m \exp\{-\frac{1}{2}[y(t)-h^{(k)}u_m]^2/(\sigma^2)^{(k)}\}}{\sum_{n=1}^{M} \exp\{-\frac{1}{2}[y(t)-h^{(k)}u_n]^2/(\sigma^2)^{(k)}\}}} + \sum_{\tau=1}^{N_{\mathrm{T}}} y_{\mathrm{T}}(\tau) u_{\mathrm{T}}(\tau)^* \right]$$

EM Step 2:

$$(\sigma^2)^{k+1} = T_2(\boldsymbol{y}) - (h^{(k)})^2 \quad \text{or} \quad T_2(\boldsymbol{y}) - (h^{(k+1)})^2.$$