Solutions to Practice Problems
ECE 271A
Electrical and Computer Engineering
University of California San Diego
Nuno Vasconcelos

1. We have seen, in problem set **3** that the least squares problem corresponds to a probabilistic model of the form

$$P_{Z|X}(z|x;\theta) = \mathcal{G}(z, f(x,\theta), \sigma^2).$$

a) The main difference with respect to problem set **3** is that the z_i are missing for $i > m+1$. To compute the likelihood of $\mathcal{D}$ we have to marginalize over these missing values,

$$\begin{aligned}
P_{Z|X}(\{z_1,\ldots,z_m\}|\{x_1,\ldots,x_n\};\theta) &= \int \prod_{i=1}^{n} \mathcal{G}(z_i, f(x_i,\theta), \sigma^2) dz_{m+1}\ldots dz_n \\
&= \prod_{i=1}^{m} \mathcal{G}(z_i, f(x_i,\theta), \sigma^2) \prod_{i=m+1}^{n} \int \mathcal{G}(z_i, f(x_i,\theta), \sigma^2) dz_i \\
&= \prod_{i=1}^{m} \mathcal{G}(z_i, f(x_i,\theta), \sigma^2) \prod_{i=m+1}^{n} 1 \\
&= \prod_{i=1}^{m} \mathcal{G}(z_i, f(x_i,\theta), \sigma^2)
\end{aligned}$$

This is the same as ignoring $x_{m+1},\ldots,x_n$. It makes sense because we do not have information on their z_i. If knowing the x_i alone produced a benefit, we could make that benefit arbitrarily large by just including more x_i. This would make no sense, since the x_i do not cost anything and add no information (they are deterministic).

b) The likelihood of the complete data is the same as above, but including all (x_i, z_i), i.e.

$$P_{Z|X}(\{z_1,\ldots,z_n\}|\{x_1,\ldots,x_n\};\theta) = \prod_{i=1}^{n} \mathcal{G}(z_i, f(x_i,\theta), \sigma^2)$$

leading to the log-likelihood

$$\begin{aligned}
L_c(\theta) &= \sum_{i=1}^{n} -\frac{(z_i - f(x_i,\theta))^2}{2\sigma^2} - \frac{n}{2}\log 2\pi\sigma^2 \\
L_c(\theta) &= \sum_{i=1}^{n} -\frac{z_i^2 - 2z_i f(x_i,\theta) + f(x_i,\theta)^2}{2\sigma^2} - \frac{n}{2}\log 2\pi\sigma^2.
\end{aligned}$$

Hence, in the E-step we need to compute, for $i \in \{m+1,\ldots,n\}$,

$$\begin{aligned}
\hat{z}_i &= E_{Z_i|\{Z_1,\ldots,Z_m\},\{X_1,\ldots,X_N\}}[z_i|z_1,\ldots,z_m,x_1,\ldots,x_n] = E_{Z_i|X_i}[z] = f(x_i,\theta^{(n)}) \\
\hat{z}_i^2 &= E_{Z_i|\{Z_1,\ldots,Z_m\},\{X_1,\ldots,X_N\}}[z_i^2|z_1,\ldots,z_m,x_1,\ldots,x_n] = E_{Z_i|X_i}[z^2] = \sigma^2 + f(x_i,\theta^{(n)}).
\end{aligned}$$

The M-step consists of solving

$$\begin{aligned}\theta^{(n+1)} &= \arg\min_\theta \sum_{i=1}^{m}(z_i - f(x_i,\theta))^2 + \sum_{i=m+1}^{n}(\hat{z}_i - f(x_i,\theta))^2 \qquad (1)\\ &= \arg\min_\theta ||\hat{\mathbf{z}} - \mathbf{\Phi}\theta||^2\end{aligned}$$

where $\hat{\mathbf{z}} = (z_1, \dots, z_m, \hat{z}_{m+1}, \dots, \hat{z}_n)$ and

$$\mathbf{\Phi} = \left[\begin{array}{ccc} 1 & \dots & x_1^K \\ \vdots & & \vdots \\ 1 & \dots & x_n^K \end{array}\right].$$

This is a standard least squares problem for which the solution, as seen in problem set 3, is

$$\theta^{(n+1)} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\hat{\mathbf{z}}.$$

Note that it does not really require the values $\hat{z}_i^2$ and, so, we do not have to compute them in the E-step. Hence, the EM algorithm consists of

E-step:

$$\hat{z}_i = f(x_i, \theta^{(n)}) \qquad \forall i > m$$

M-step:

$$\theta^{(n+1)} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\hat{\mathbf{z}}$$

with $\hat{\mathbf{z}} = (z_1, \dots, z_m, \hat{z}_{m+1}, \dots, \hat{z}_n)$ and

$$\mathbf{\Phi} = \left[\begin{array}{ccc} 1 & \dots & x_1^K \\ \vdots & & \vdots \\ 1 & \dots & x_n^K \end{array}\right].$$

c) Note that, upon convergence,

$$\theta^{(n+1)} = \theta^{(n)}$$

and so, for $i > m$,

$$\hat{z}_i = f(x_i, \theta^{(n)}) = f(x_i, \theta^{(n+1)}).$$

It follows that, in equation 1, the term

$$\sum_{i=m+1}^{n}(\hat{z}_i - f(x_i,\theta))^2$$

will be zero, i.e. the $\hat{z}_i$ converge to the values that satisfy the regression with zero error. This will happen independently of the value of $\theta^{(n)}$ to which EM converges. The truth is that we have too many degrees of freedom, it is always possible to set the missing z_i so that they meet the regression exactly, and EM does just that. Hence, the $\hat{z}_i$ end up having no influence in the solution. It follows that this is the one that maximizes the likelihood function of **a**), which is equivalent to minimizing the first summation of equation 1.

d) Using the EM algorithm seems unnecessary at first, since the cost function of **a)** has closed form solution

$$\theta^* = (\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{y}.$$

where $\mathbf{y} = (z_1, \ldots, z_m)$ and

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & \ldots & x_1^K \\ \vdots & & \vdots \\ 1 & \ldots & x_m^K \end{bmatrix}.$$

Usually this is indeed the case, and there is no benefit to EM. On the other hand, the EM solution can be very beneficial if 1) the regression has to be repeated many times (e.g. an industrial process that has to be computed every day, or the vision system of a robot that has to compute the regression thirty times a second), 2) the missing values may vary (i.e. z_i is not always missing for the same i), and 3) the matrices are large, making the matrix inversion required by the least squares solution a complex operation. Note that the matrix $(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ depends only on the x_i and can therefore be pre-computed once and stored. The M-step thus consists of a single matrix-vector multiplication and is not heavy, making EM computationally appealing (even though it is an iterative solution). On the other hand, because the values of x_i required by $(\mathbf{\Gamma}^T\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T$ keep changing, the direct solution of **a)** always requires matrix inversion. Hence, depending on the number of iterations required for EM convergence, the frequency with which the regression needs to be updated, and the matrix size, EM may be the best solution. In practice, this is not a rare situation.

2. We start by writing the expression for the joint pdf

$$P_{\mathbf{W},U}(\mathbf{w},u;\mu,\mathbf{\Sigma}) = P_{\mathbf{W}|U}(\mathbf{w}|u;\mu,\mathbf{\Sigma})P_U(u) = \mathcal{G}\left(\mathbf{w},\mu,\frac{1}{u}\mathbf{\Sigma}\right)\frac{\beta^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}u^{\frac{\nu}{2}-1}e^{-\frac{\nu}{2}u}$$

and the log-likelihood of the complete data $\{\mathcal{D},\mathcal{U}\}$

$$\begin{aligned} L_c(\mu,\mathbf{\Sigma}) &= \sum_{i=1}^{n}\left[-\frac{u_i}{2}(\mathbf{w}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}_i-\mu) - \frac{1}{2}\log[(2\pi)^d\frac{1}{u_i^d}|\mathbf{\Sigma}|] + \left(\frac{\nu}{2}-1\right)\log u_i - \frac{\nu}{2}u_i\right] \\ &\quad + n\log\frac{\beta^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} \end{aligned}$$

Since the end-goal is to maximize with respect to $(\mu,\mathbf{\Sigma})$ we can drop all terms that do not depend on these parameters to obtain

$$L_c'(\mu,\mathbf{\Sigma}) = \sum_{i=1}^{n}\left[-\frac{u_i}{2}(\mathbf{w}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}_i-\mu) - \frac{1}{2}\log|\mathbf{\Sigma}|\right].$$

This is linear in the hidden variables u_i and so, in the E-step, we simply need to compute

$$E_{U_i|\mathbf{W};\mu^{(n)},\mathbf{\Sigma}^{(n)}}[u_i|\mathcal{D}].$$

For this we start by computing

$$\begin{aligned} P_{U_i|\mathbf{W}}(u_i|\mathcal{D}) &= \frac{P_{\mathbf{W}|U_i}(\mathcal{D}|u_i)P_{U_i}(u_i)}{P_{\mathbf{W}}(\mathcal{D})} \\ &= \frac{\prod_{k\neq i}P_{\mathbf{W}_k}(\mathbf{w}_k)P_{\mathbf{W}_i|U_i}(\mathbf{w}|u_i)P_{U_i}(u_i)}{\prod_{k\neq i}P_{\mathbf{W}_k}(\mathbf{w}_k)P_{\mathbf{W}_i}(\mathbf{w}_i)} \\ &= P_{U_i|\mathbf{W}_i}(u_i|\mathbf{w}_i) \\ &= P_{U|\mathbf{W}}(u|\mathbf{w}). \end{aligned}$$

We next note that

$$P_{\mathbf{W}|U}(\mathbf{w}|u) = \frac{u^d}{\sqrt{(2\pi)^d|\mathbf{\Sigma}|}}\exp\left\{-\frac{u}{2}(\mathbf{w}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}-\mu)\right\}$$

is a member of the exponential family, i.e. of the form

$$P_{\mathbf{W}|U}(\mathbf{w}|u) = f(\mathbf{w})g(u)e^{\phi(u)^T u(\mathbf{w})}$$

with

$$f(\mathbf{w}) = 1,\;\; g(u) = \frac{u^d}{\sqrt{(2\pi)^d\mathbf{\Sigma}}},\;\; \phi(u) = -\frac{u}{2},\;\; u(\mathbf{w}) = (\mathbf{w}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}-\mu),$$

and that

$$P_U(u) \propto u^{\frac{\nu}{2}-1}e^{-\frac{u}{2}\nu} \propto g(u)^{\frac{\nu-2}{2d}}e^{-\frac{u}{2}\nu}.$$

We have seen in problem set 3 that this implies that 1) $P_U(u)$ is a conjugate prior for $P_{\mathbf{W}|U}(\mathbf{w}|u)$ and 2) the posterior is of the form

$$\begin{aligned} P_{U|\mathbf{W}}(u|\mathbf{w}) &\propto g(u)^{\frac{\nu-2}{2d}+1}e^{\phi(u)[u(\mathbf{w})+\nu]} \\ &\propto u^{\frac{\nu-2+2d}{2}}e^{-\frac{1}{2}[\nu+(\mathbf{w}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}-\mu)]u} \\ &\propto u^{\frac{\nu+2d}{2}-1}e^{-\frac{1}{2}[\nu+(\mathbf{w}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}-\mu)]u} \end{aligned}$$

i.e.

$$P_{U|\mathbf{W}}(u|\mathbf{w}) \sim Gamma\left(\frac{\nu+2d}{2}, \frac{1}{2}\left[\nu + (\mathbf{w}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}-\mu)\right]\right).$$

It follows that

$$\hat{u}_i = E_{U|\mathbf{W};\mu^{(n)},\mathbf{\Sigma}^{(n)}}[u_i|\mathcal{D}] = \frac{\nu+2d}{\nu + (\mathbf{w}_i-\mu^{(n)})^T(\mathbf{\Sigma}^{(\mathbf{n})})^{-1}(\mathbf{w}_i-\mu^{(n)})}.$$

The Q function is therefore

$$Q(\{\mu,\mathbf{\Sigma}\}|\{\mu^{(n)},\mathbf{\Sigma}^{(n)}\}) = \arg\max_{\{\mu,\mathbf{\Sigma}\}} -\sum_{i=1}^{n}\frac{\hat{u}_i}{2}(\mathbf{w}_i-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{w}_i-\mu) - \frac{n}{2}\log|\mathbf{\Sigma}|.$$

It is maximized when

$$\frac{\partial Q}{\partial \mu} = -\sum_{i=1}^{n}\hat{u}_i\mathbf{\Sigma}^{-1}(\mathbf{w}_i-\mu) = 0$$

i.e.

$$\mu^{(n+1)} = \frac{\sum_{i=1}^{n}\hat{u}_i\mathbf{w}_i}{\sum_{i=1}^{n}\hat{u}_i}$$

and

$$\frac{\partial Q}{\partial \mathbf{\Sigma^{-1}}} = -\sum_{i=1}^{n}\frac{\hat{u}_i}{2}(\mathbf{w}_i-\mu)(\mathbf{w}_i-\mu)^T + \frac{n}{2}\frac{1}{|\mathbf{\Sigma}^{-1}|}|\mathbf{\Sigma}^{-1}|\mathbf{\Sigma} = 0$$

i.e.

$$\mathbf{\Sigma}^{(n+1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{u}_i}{2}(\mathbf{w}_i-\mu^{(n+1)})(\mathbf{w}_i-\mu^{(n+1)})^T.$$

In summary, the EM algorithm consists of
E-step:

$$\hat{u}_i = \frac{\nu+2d}{\nu + (\mathbf{w}_i-\mu^{(n)})^T(\mathbf{\Sigma}^{(n)})^{-1}(\mathbf{w}_i-\mu^{(n)})}$$

M-step:

$$\mu^{(n+1)} = \frac{\sum_{i=1}^{n}\hat{u}_i\mathbf{w}_i}{\sum_{i=1}^{n}\hat{u}_i} \qquad \mathbf{\Sigma}^{(n+1)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{u}_i}{2}(\mathbf{w}_i-\mu^{(n+1)})(\mathbf{w}_i-\mu^{(n+1)})^T.$$