02471 Machine Learning for Signal Processing

Solution

Exercise 9: Bayesian inference and the EM algorithm

9.1 Cost functions, Maximum Likelihood and Bayesian Inference

Exercise 9.1.1

The multivariate normal distribution (or multivariate Gaussian distribution) is

$$p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu}_{\boldsymbol{y}},\Sigma_{\boldsymbol{y}}) = \frac{1}{(2\pi)^{N/2}|\Sigma_{\boldsymbol{y}}|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})^T\Sigma_{\boldsymbol{y}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{\boldsymbol{y}})\right)$$

The log to this expression, using the rules $\ln ab = \ln a + \ln b$ and $\ln a^b = b\ln a$ becomes

$$\begin{aligned}\ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu}_{\boldsymbol{y}},\Sigma_{\boldsymbol{y}}) &= \ln(2\pi)^{-N/2} + \ln|\Sigma_{\boldsymbol{y}}|^{-1/2} - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})\\ &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_{\boldsymbol{y}}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})^T\Sigma_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})\end{aligned}$$

We need to find the expression for $\boldsymbol{\mu}_{\boldsymbol{y}}$ which is

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{y}} &= \mathbb{E}[\boldsymbol{y}]\\ &= \mathbb{E}[f(X,\boldsymbol{\theta}) + \eta]\\ &= f(X,\boldsymbol{\theta}) + \mathbb{E}[\eta]\end{aligned}$$

If we assume zero-mean noise, $\mathbb{E}[\eta] = 0$, we have $\mathbb{E}[\boldsymbol{y}] = f(X,\boldsymbol{\theta})$. Additionally we need to find the expression for $\Sigma_{\boldsymbol{y}}$:

$$\begin{aligned}\Sigma_{\boldsymbol{y}} &= \mathbb{E}\Big[(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])(\boldsymbol{y}-\mathbb{E}[\boldsymbol{y}])^T\Big]\\ &= \mathbb{E}\Big[(f(X,\boldsymbol{\theta}) + \eta - f(X,\boldsymbol{\theta}))(f(X,\boldsymbol{\theta}) + \eta - f(X,\boldsymbol{\theta}))^T\Big]\\ &= \mathbb{E}\big[\eta\eta^T\big]\\ &= \Sigma_\eta\end{aligned}$$

By substitution we now obtain

$$\begin{aligned}\ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu}_{\boldsymbol{y}},\Sigma_{\boldsymbol{y}}) &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_{\boldsymbol{y}}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})^T\Sigma_{\boldsymbol{y}}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{y}})\\ &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_\eta| - \frac{1}{2}(\boldsymbol{y}-f(X,\boldsymbol{\theta}))^T\Sigma_\eta^{-1}(\boldsymbol{y}-f(X,\boldsymbol{\theta}))\end{aligned}$$

Exercise 9.1.2

If we assume that we have noise that is statistically independent sample to sample (e.g white noise), and assume $\Sigma_\eta = \sigma^2 I$. In that case, we have $|\Sigma_\eta| = |\sigma^2 I| = \sigma^{2N}$, and $\Sigma_\eta^{-1} = (\sigma^2 I)^{-1} =$

$\frac{1}{\sigma^2}I$. Thus we can rewrite

$$\begin{aligned}
\ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu}_{\boldsymbol{y}},\Sigma_{\boldsymbol{y}}) &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_\eta| - \frac{1}{2}(\boldsymbol{y} - f(X,\boldsymbol{\theta}))^T\Sigma_\eta^{-1}(\boldsymbol{y} - f(X,\boldsymbol{\theta})) \\
&= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^{2N} - \frac{1}{2}(\boldsymbol{y} - f(X,\boldsymbol{\theta}))^T\frac{1}{\sigma^2}I(\boldsymbol{y} - f(X,\boldsymbol{\theta})) \\
&= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - f(X,\boldsymbol{\theta}))^T(\boldsymbol{y} - f(X,\boldsymbol{\theta})) \\
&= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2
\end{aligned}$$

Exercise 9.1.3

If we consider this as an optimization problem we have

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta}|\boldsymbol{y};\boldsymbol{\mu}_{\boldsymbol{y}},\Sigma_{\boldsymbol{y}}) \\
&= \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu}_{\boldsymbol{y}},\Sigma_{\boldsymbol{y}}) + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 + \ln p(\boldsymbol{\theta})
\end{aligned}$$

If we consider the prior as constant we can remove that from the optimization problem, thus we get

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2
\end{aligned}$$

Exercise 9.1.4

Reusing the expression from the previous exercise we get

$$\begin{aligned}
\ln p(\boldsymbol{\theta};\mathbf{0},\sigma_\theta^2 I) &= -\frac{K}{2}\ln(2\pi) - \frac{K}{2}\ln\sigma_\theta^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta} - \mathbf{0}\|^2 \\
&= -\frac{K}{2}\ln(2\pi) - \frac{K}{2}\ln\sigma_\theta^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2
\end{aligned}$$

Let us combine this result with the log-posterior we derived in the last exercise

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 - \frac{K}{2}\ln(2\pi) - \frac{K}{2}\ln\sigma_\theta^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 - \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 + \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2
\end{aligned}$$

If we reparameterize with $\sigma_\theta^2 = \frac{\sigma^2}{\lambda} \Leftrightarrow \lambda = \frac{\sigma^2}{\sigma_\theta}$, we get

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 + \frac{1}{2\frac{\sigma^2}{\lambda}}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 + \frac{\lambda}{2\sigma^2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - f(X,\boldsymbol{\theta})\|^2 + \lambda\|\boldsymbol{\theta}\|^2
\end{aligned}$$

Exercise 9.1.5

We consider the log to the univariate Laplacian distribution

$$\begin{aligned}
\ln p(x|\mu, b) &= \ln\left(\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)\right) \\
&= \ln(2b)^{-1} - \frac{1}{b}|x-\mu| \\
&= -\ln 2 - \ln b - \frac{1}{b}|x-\mu|
\end{aligned}$$

Exercise 9.1.6

Let is now consider a weight vector $\boldsymbol{\theta}$ of length l. If we assume each θ_k follows a zero-mean Laplacian distribution, and the individual weights are statistical independent, we get

$$\begin{aligned}
\ln p(\boldsymbol{\theta}|0, b) &= \sum_{i=1}^{l} -\ln 2 - \ln b - \frac{1}{b}|\theta_i| \\
&= -l\ln 2 - l\ln b - \frac{1}{b}\sum_{i=1}^{l}|\theta_i| \\
&= -l\ln 2 - l\ln b - \frac{1}{b}\|\boldsymbol{\theta}\|_1
\end{aligned}$$

Exercise 9.1.7

Combine this with the previous results, and obtain the compete log-likelihood $\boldsymbol{\theta}$

$$\begin{aligned}
\ln p(\boldsymbol{\theta}, \boldsymbol{y}|X) &= \ln p(\boldsymbol{y}|\boldsymbol{\theta}; f(X,\boldsymbol{\theta}), \sigma^2 I) + \ln p(\boldsymbol{\theta}|0, b) \\
&= -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|(\boldsymbol{y} - f(X,\boldsymbol{\theta}))\|^2 - l\ln 2 - l\ln b - \frac{1}{b}\|\boldsymbol{\theta}\|_1
\end{aligned}$$

Exercise 9.1.8

From Bayes formula, we know, given a dataset X, optimizing $\ln p(\boldsymbol{\theta}, \boldsymbol{y}|X)$ is the same as optimizing $\ln p(\boldsymbol{\theta}|\boldsymbol{y}|X)$. Disregarding all terms not related to $\boldsymbol{\theta}$ we get

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|(\boldsymbol{y} - f(X,\boldsymbol{\theta}))\|^2 - \frac{1}{b}\|\boldsymbol{\theta}\|_1 \\
&= \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|(\boldsymbol{y} - f(X,\boldsymbol{\theta}))\|^2 + \frac{1}{b}\|\boldsymbol{\theta}\|_1
\end{aligned}$$

If we reparameterize with $b = 2\sigma^2/\lambda$ we get the Lasso cost function and we can see that LASSO corresponds to having normal likelihood with i.id.d samples and univariate Laplace prior on $\boldsymbol{\theta}$.

9.2 Derive EM updates for Bayesian linear regression

Exercise 9.2.1

We have already derived expressions for these in the previous exercise. Using the previous results we get:

$$\begin{aligned}\ln p(\boldsymbol{y},\boldsymbol{\theta}|\alpha,\beta) &= \ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\theta},\beta) + \ln p(\boldsymbol{\theta};\mathbf{0},\alpha)\\ &= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{\beta}{2}\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2 - \frac{K}{2}\ln(2\pi) + \frac{K}{2}\ln\alpha - \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2\\ &= -\frac{1}{2}(N+K)\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{\beta}{2}\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2 + \frac{K}{2}\ln\alpha - \frac{\alpha}{2}\|\boldsymbol{\theta}\|^2\end{aligned}$$

To compute the expectation we use the following rule $A^TA = \mathrm{trace}(AA^T)$, and use that trace is a linear operator i.e. $\mathrm{trace}(A+B) = \mathrm{trace}(A) + \mathrm{trace}(B)$, and $\mathbb{E}[\mathrm{trace}(A)] = \mathrm{trace}(\mathbb{E}[A])$:

$$\begin{aligned}A := \mathbb{E}\big[\theta^T\theta\big] &= \mathbb{E}\big[\mathrm{trace}(\theta\theta^T)\big]\\ &= \mathrm{trace}\big(\mathbb{E}\big[\theta\theta^T\big]\big)\end{aligned}$$

We recognize $\mathbb{E}\big[\theta\theta^T\big]$ as the structure of the correlation matrix eq (2.33), hence we have, at step j

$$\begin{aligned}\mathbb{E}\big[\theta\theta^T\big] &= \mathrm{Cov}(\theta) + \mathbb{E}[\theta]\mathbb{E}\big[\theta^T\big]\\ &= \Sigma_{\theta|y}^{(j)} + \boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\end{aligned}$$

Inserting into the trace we get $\mathrm{trace}(A+B) = \mathrm{trace}(A) + \mathrm{trace}(B)$, and $\mathbb{E}[\mathrm{trace}(A)] = \mathrm{trace}(\mathbb{E}[A])$, we get

$$\begin{aligned}A &= \mathrm{trace}\Big(\Sigma_{\theta|y}^{(j)} + \boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Big)\\ &= \mathrm{trace}\Big(\Sigma_{\theta|y}^{(j)}\Big) + \mathrm{trace}\Big(\boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Big)\\ &= \mathrm{trace}\Big(\Sigma_{\theta|y}^{(j)}\Big) + \|\boldsymbol{\mu}_{\theta|y}^{(j)}\|^2\end{aligned}$$

Exercise 9.2.2

The other term we need to evaluate is $\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2$. To evaluate, we again use the $\mathrm{trace}(\cdot)$ function and perform the following rewrite

$$\begin{aligned}\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2 &= (\boldsymbol{y}-\Phi\boldsymbol{\theta})^T(\boldsymbol{y}-\Phi\boldsymbol{\theta})\\ &= \boldsymbol{y}^T\boldsymbol{y} - (\Phi\boldsymbol{\theta})^T\boldsymbol{y} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} + (\Phi\boldsymbol{\theta})^T\Phi\boldsymbol{\theta}\\ &= \boldsymbol{y}^T\boldsymbol{y} - (\Phi\boldsymbol{\theta})^T\boldsymbol{y} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} + \mathrm{trace}\big(\Phi\boldsymbol{\theta}(\Phi\boldsymbol{\theta})^T\big)\\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} - \boldsymbol{y}^T\Phi\boldsymbol{\theta} + \mathrm{trace}\big(\Phi\boldsymbol{\theta}\boldsymbol{\theta}^T\Phi^T\big)\\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\boldsymbol{\theta} + \mathrm{trace}\big(\Phi\boldsymbol{\theta}\boldsymbol{\theta}^T\Phi^T\big)\end{aligned}$$

To proceed we now take the expectation to $\|\boldsymbol{y}-\Phi\boldsymbol{\theta}\|^2$, where $\boldsymbol{\theta}$ is the only random variable, and again using that $\mathrm{trace}(\cdot)$ is a linear operator we get

$$\begin{aligned}B := \mathbb{E}\big[\|\boldsymbol{y}-\Phi\theta\|^2\big] &= \mathbb{E}\big[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\theta + \mathrm{trace}\big(\Phi\theta\theta^T\Phi^T\big)\big]\\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\mathbb{E}[\theta] + \mathrm{trace}\big(\Phi\mathbb{E}\big[\theta\theta^T\big]\Phi^T\big)\end{aligned}$$

Exercise 9.2.3

We have already found the expressions for $\mathbb{E}[\theta]$ and $\mathbb{E}\left[\theta\theta^T\right]$ earlier, so by substitution, and again using that trace$(\cdot)$ is a linear operator we get

$$\begin{aligned}
B &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\boldsymbol{\mu}_{\theta|y}^{(j)} + \text{trace}\Big(\Phi\Big(\Sigma_{\theta|y}^{(j)} + \boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Big)\Phi^T\Big) \\
&= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\boldsymbol{\mu}_{\theta|y}^{(j)} + \text{trace}\Big(\Phi\Sigma_{\theta|y}^{(j)}\Phi^T\Big) + \text{trace}\Big(\Phi\boldsymbol{\mu}_{\theta|y}^{(j)}\boldsymbol{\mu}_{\theta|y}^{(j)T}\Phi^T\Big) \\
&= \|\boldsymbol{y} - \Phi\boldsymbol{\mu}_{\theta|y}^{(j)}\|^2 + \text{trace}\Big(\Phi\Sigma_{\theta|y}^{(j)}\Phi^T\Big)
\end{aligned}$$

Exercise 9.2.4

From the book, sec 12.9.4 we have expressions for how to specify the posterior. From eq. (12.135) and eq. (12.136) we have, if

$$\begin{aligned}
p(\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_z, \Sigma_z) \\
p(\boldsymbol{t}|\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{t}|\boldsymbol{z}; A\boldsymbol{z}, \Sigma_{t|z})
\end{aligned}$$

then the posterior is

$$\begin{aligned}
p(\boldsymbol{z}|\boldsymbol{t}) &= \mathcal{N}(\boldsymbol{z}|\boldsymbol{t}; \boldsymbol{\mu}_{z|t}, \Sigma_{z|t}) \\
\boldsymbol{\mu}_{z|t} &= \boldsymbol{\mu}_z + \Sigma_{z|t}A^T\Sigma_{t|z}^{-1})(\boldsymbol{t} - A\boldsymbol{\mu}_z) \\
\Sigma_{z|t} &= (\Sigma_z^{-1} + A^T\Sigma_{t|z}^{-1}A)^{-1}
\end{aligned}$$

In our case, we have $\boldsymbol{z} := \boldsymbol{\theta}$, $\boldsymbol{\mu}_z := \boldsymbol{0}$, $\boldsymbol{t} := \boldsymbol{y}$, $\Sigma_z^{-1} := \alpha I$, $\boldsymbol{t} := \boldsymbol{y}$, $A := \Phi$, and $\Sigma_{t|z}^{-1} := \beta I$. Then we get the following expressions

$$\begin{aligned}
\boldsymbol{\mu}_{\theta|y} &= \beta\Sigma_{\theta|y}\Phi^T\boldsymbol{y} \\
\Sigma_{\theta|y} &= (\alpha I + \beta\Phi^T\Phi)^{-1}
\end{aligned}$$

Exercise 9.2.5

The derivative of $\mathcal{Q}\big(\alpha, \beta; \alpha^{(j)}, \beta^{(j)}\big)$ follows the same structure, so we only show one of them.

$$\begin{aligned}
\frac{\partial}{\partial\alpha}\mathcal{Q}\big(\alpha, \beta; \alpha^{(j)}, \beta^{(j)}\big) = \frac{K}{2}\frac{1}{\alpha} - \frac{1}{2}A &= 0, \quad \Leftrightarrow \\
\frac{1}{\alpha} &= \frac{A}{K}, \quad \Leftrightarrow \\
\alpha &= \frac{K}{A}
\end{aligned}$$

By symmetry, we get

$$\beta = \frac{N}{B}$$

Hence, the update equations will be $\alpha^{j+1} = K/B$ and $\beta^{j+1} = N/B$.