02471 Machine Learning for Signal Processing

Week 9: Bayesian inference and the EM algorithm

This exercise is based on S. Theodoridis: Machine Learning, A Bayesian and Optimization Perspective 2nd edition, section(s) 3.10–3.11, 12.1–12.2, 12.4–12.5.

The objective of this exercise is to establish the connection between the way we have performed parameter estimation so far and the probabilistic framework. Previously we have designed different cost functions based on what "made sense", but it turns out we can arrive to the same cost functions in a probabilistic framework, thus establishing a unified and more interpretable way to make modeling decisions.

Overview

The exercise have the following structure:

9.1 will establish the connection between the cost functions we have used so far, maximum likelihood and Bayes inference. This exercise is mostly for students who haven't had the Bayesian machine learning course. If you had this course, you should probably skip this exercise.

9.2 will derive updates for EM using Bayes linear regression as the example model.

9.3 will experiment with the derived formulas, and get acquainted with how Bayes linear regression performs.

You can choose to start or end with exercise 9.3.

Notation

- $J(\boldsymbol{\theta})$ is a cost function that we are seeking to minimize with respect to $\boldsymbol{\theta}$.
- $\mathcal{N}(\mu, \sigma^2)$ denotes a normal (Gaussian) distribution with mean value μ and variance σ^2.
- $U(a, b)$ denotes a uniform distribution with a as lower limit and b as upper limit.
- $\mathbf{x} \sim$ distribution$(\cdot)$ means that $\mathbf{x}$ is a random variable that follows the associated distribution.
- $\|\cdot\|_\ell$ denotes the ℓ-norm. If ℓ is absent, we are implicitly referring to the $\ell = 2$ norm, i.e. $\|\cdot\| = \|\cdot\|_2$.

Code

The code can be found in the .m and .py files named in the same way as exercises, ie. the code for exercise 9.x.y is in the file 9_x_y.m (or .py).

For coding exercises that requires implementation we will usually write `complete this line` where the implementing should be done.

Solutions

The solution is provided for all derivation exercises, and often hints are provided at the end of the document. If you get stuck, take a look at the hints, and if you are still stuck, take a look in the solution to see the approach being taken. Then try to do it on your own.

Solutions are also provided for some coding exercises. If you get stuck, take a look at the solution, and then try to implement it on your own.

9.1 Cost functions, Maximum Likelihood and Bayesian Inference

Consider the following model for the supervised learning problem

$$\mathrm{y} = f(\mathbf{x}, \boldsymbol{\theta}) + \eta$$

where, when we have observations we have $(y_n, \boldsymbol{x}_n)$, $n = 1, \cdots, N$, and $\hat{y}_n = f(\boldsymbol{x}_n, \boldsymbol{\theta})$ is the model that maps input data to output data (or, in machine learning terms, maps features to the target variable). In that case, we can write the model as

$$y = f(X, \boldsymbol{\theta}) + \eta$$

To shed light on the relationships between cost functions used in previous weeks, maximum likelihood and Bayes' inference, we consider Bayes' formula, where we condition on an observed dataset X:

$$p(\boldsymbol{\theta}|\boldsymbol{y}, X) = \frac{p(\boldsymbol{y}, \boldsymbol{\theta}|X)}{p(\boldsymbol{y}|X)}$$

To avoid notational clutter, we drop the condition on X and rewrite the joint probability $p(\boldsymbol{y}, \boldsymbol{\theta}|X)$ using the product rule (eq. 2.12):

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

The numerator is often called the complete likelihood. Taking the log of the above expression we get

$$\ln p(\boldsymbol{\theta}|\boldsymbol{y}) = \ln p(\boldsymbol{y}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{y})$$

To apply Bayes' formula for machine learning problems, we need to specify two distributions (models); $p(\boldsymbol{y}|\boldsymbol{\theta})$, which maps parameters and data together, and $p(\boldsymbol{\theta})$ which encodes our prior beliefs. If we consider the above as an optimization problem and only consider the terms that include $\boldsymbol{\theta}$, we get

$$\ln p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \ln p(\boldsymbol{y}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}), \quad \text{also called:}$$
$$\log \text{posterior} \propto \log \text{likelihood} + \log \text{prior}$$

In this exercise we will derive three results; 1) the connection between the log likelihood using a Gaussian distribution and the mean-squared error cost function $J(\boldsymbol{\theta})$, 2) the connection between the MAP estimate when using a Gaussian prior and Ridge regression, and 3) the connection between the MAP estimate when using a Laplacian prior and LASSO.

Exercise 9.1.1

First we will show that the mean squared error cost function corresponds to having a improper flat prior (then $\ln p(\boldsymbol{\theta})$ is constant) and a normal likelihood. The multivariate normal

distribution (or multivariate Gaussian distribution) is

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

As can be seen, the two terms $\boldsymbol{\mu}_y$ and Σ_y enters the equation as parameters. To make the formula applicable we need expressions for these two terms.

Show that $\boldsymbol{\mu}_y = f(X,\boldsymbol{\theta}) + \mathbb{E}[\eta]$. In the sequel we assume we have zero-mean noise ($\mathbb{E}[\eta] = 0$). Show that $\Sigma_y = \Sigma_\eta$. Finally, by substituting the found expressions into the multivariate Gaussian distribution, show, using the rules $\ln ab = \ln a + \ln b$ and $\ln a^b = b\ln a$, that

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

Exercise 9.1.2

Now we assume that we have collected N samples, and that the noise is statistically independent from sample to sample (e.g white noise), so we write $\Sigma_\eta = \sigma^2 I$. In that case, we have $|\Sigma_\eta| = |\sigma^2 I| = \sigma^{2N}$, and $\Sigma_\eta^{-1} = (\sigma^2 I)^{-1} = \frac{1}{\sigma^2}I$. Under this assumption, show that

$$\ln p(\boldsymbol{y}|\boldsymbol{\theta};\boldsymbol{\mu}_y,\Sigma_y) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2$$

Exercise 9.1.3

Next, we consider the MAP optimization problem (maximize the posterior) with this particular likelihood w.r.t. $\boldsymbol{\theta}$, hence we can disregard all terms that is not dependent on $\boldsymbol{\theta}$. Show that this optimization problem can be written as

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} -\frac{1}{2\sigma^2}\|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2 + \ln p(\boldsymbol{\theta})$$

and then demonstrate that, if the flat improper prior ($p(\boldsymbol{\theta})$ is constant), we get

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y}-f(X,\boldsymbol{\theta})\|^2$$

Thus we can see that maximizing the derived log-likelihood corresponds to optimizing the mean-squared error, and hence we have readily established the link and now we can understand that the likelihood is essentially the model for our noise. We now have a framework in place, where, if we desire another noise model than e..g white noise, we simply go through the previous steps to derive the cost function.

Additionally, we can see that log-prior that added to the cost function, based on what we have learned so far, could serve as a regularization term.

Exercise 9.1.4

Let us now consider the case where we use a normal prior, with $\boldsymbol{\theta} \in \mathbb{R}^K$, and the weights are statistically independent ($\Sigma_\theta = \sigma_\theta^2 I$) and zero-mean. Formally, $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta};0,\sigma^2 I)$. Reuse the expression from the previous exercise to get

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

Combine this result with the log-likelihood we derived in the last exercise to obtain

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - f(X, \boldsymbol{\theta})\|^2 + \frac{1}{2\sigma_\theta^2}\|\boldsymbol{\theta}\|^2$$

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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - f(X, \boldsymbol{\theta})\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

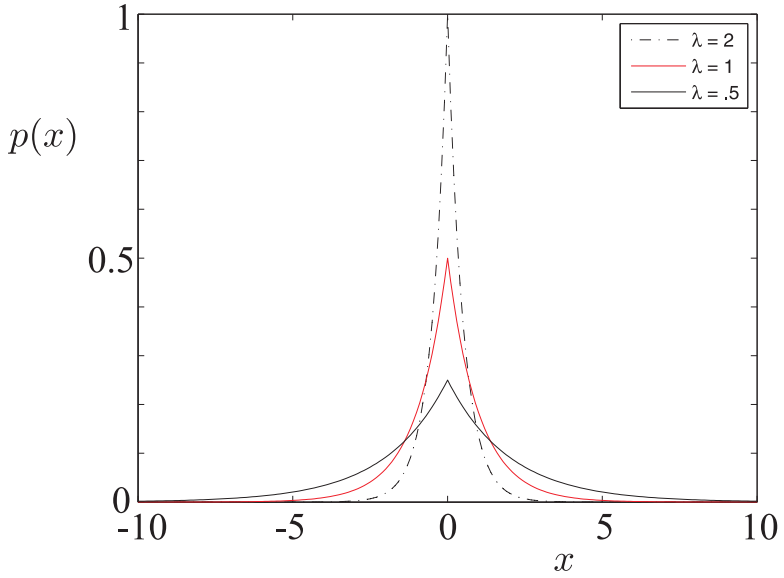
We recognize this expression as the Ridge regression if we use a linear model for $f(X, \boldsymbol{\theta})$, hence we have developed this cost function from a probabilistic perspective and shown that Ridge regression is indeed the same as maximizing the log-posterior if we use a normal log-likelihood and a normal log-prior.

Exercise 9.1.5

Let us now consider the log to the univariate Laplacian distribution, defined as

$$p(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

And looks like (on the figure λ corresponds to b)



Show that the log to the Laplace distribution is

$$\ln p(x|\mu, b) = -\ln 2 - \ln b - \frac{1}{b}|x - \mu|$$

Exercise 9.1.6

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

$$\ln p(\boldsymbol{\theta}|0, b) = -l\ln 2 - l\ln b - \frac{1}{b}\|\boldsymbol{\theta}\|_1$$

Exercise 9.1.7

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

$$\ln p(\boldsymbol{\theta}, \boldsymbol{y}|X) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\|(\boldsymbol{y} - f(X,\boldsymbol{\theta}))\|^2 - l\ln 2 - l\ln b - \frac{1}{b}\|\boldsymbol{\theta}\|_1$$

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

Before carrying out this exercise, be sure to read ML section 12.4–12.5. In this exercise, we are going to derive then updates necessary for using EM for the general linear regression model.

Using the EM algorithm requires the following:

1. Specification of the complete log likelihood, $\ln p(\boldsymbol{y}, \boldsymbol{\theta})$ (The model)
2. Derive $\mathbb{E}\big[\ln p(\boldsymbol{y}, \boldsymbol{\theta}; \boldsymbol{\xi}^{(j)})\big]$ and denote this term $\mathcal{Q}(\boldsymbol{\xi}, \boldsymbol{\xi}^{(j)})$.
3. Maximize $\mathcal{Q}(\boldsymbol{\xi}, \boldsymbol{\xi}^{(j)})$ in order to get $\boldsymbol{\xi}^{(j+1)}$

We consider a normal likelihood and a normal prior. These expressions (or the logs of them) have already been derived in the earlier exercise. We assume we have N observations and our parameter vector $\boldsymbol{\theta}$ is of length K. Additionally, we choose to simplify the covariance matrices and assume white noise, i.e. $\Sigma_\eta = \beta^{-1}I$, and that the parameters $\boldsymbol{\theta}$ are independent in the prior $\Sigma_{\boldsymbol{\theta}} = \alpha^{-1}I$, where α and β are the precision parameters, i.e $\alpha = \frac{1}{\sigma_{\boldsymbol{\theta}}^2}$ and $\beta = \frac{1}{\sigma_\eta^2}$.

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

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

To employ EM, we need to define $\boldsymbol{\xi}$, the parameter vector that is learned by EM, and derive $\mathbb{E}\big[\ln p(\boldsymbol{y}, \boldsymbol{\theta}; \boldsymbol{\xi}^{(j)})\big]$. We can readily see that $\boldsymbol{\xi} = [\alpha\ \beta]^T$. With respect to the expectation, terms that does not contain $\boldsymbol{\theta}$ are considered constants w.r.t. the expectation, we have two terms for which we need to evaluate, namely $\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2$ and $\|\boldsymbol{\theta}\|^2$.

Exercise 9.2.1

We will start deriving $A := \mathbb{E}[\|\theta\|^2]$.

To compute the expectation we use the following rule $\theta^T\theta = \text{trace}(\theta\theta^T)$ (since θ is a vector), and use that trace is a linear operator i.e. $\text{trace}(A+B) = \text{trace}(A) + \text{trace}(B)$, and $\mathbb{E}[\text{trace}(A)] = \text{trace}(\mathbb{E}[A])$. Then show that

$$A := \mathbb{E}\big[\|\theta\|^2\big] = \text{trace}\big(\mathbb{E}\big[\theta\theta^T\big]\big)$$

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 in the EM

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

Verify this step by yourself. Finally, show that by substitution we obtain

$$A = \text{trace}\Big(\Sigma^{(j)}_{\theta|y}\Big) + \|\boldsymbol{\mu}^{(j)}_{\theta|y}\|^2$$

Hence, we have now found an expression for one of the terms needed. To compute this term, we need expressions for $\boldsymbol{\mu}^{(j)}_{\theta|y}$ and $\Sigma^{(j)}_{\theta|y}$ which we will compute in a moment. First we will derive the expectation of the second term.

Exercise 9.2.2

The other term we need to evaluate is $\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2$. To evaluate, we again use the properties of the trace operator, show that by doing so we can perform the following rewrite

$$\|\boldsymbol{y} - \Phi\boldsymbol{\theta}\|^2 = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\Phi\boldsymbol{\theta} + \text{trace}\big(\Phi\boldsymbol{\theta}\boldsymbol{\theta}^T\Phi^T\big)$$

Exercise 9.2.3

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 $\text{trace}(\cdot)$ is a linear operator show that we get

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

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

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

We have now arrived at the solution for B, so we can now put it all together and obtain

$$\mathcal{Q}\big(\alpha,\beta;\alpha^{(j)},\beta^{(j)}\big) = \mathbb{E}[\ln p(\boldsymbol{y},\boldsymbol{\theta}|\alpha,\beta)] = -\frac{1}{2}(N+K)\ln(2\pi) + \frac{N}{2}\ln\beta - \frac{\beta}{2}B + \frac{K}{2}\ln\alpha - \frac{\alpha}{2}A$$

Exercise 9.2.4

The missing piece for the expectation step is to be able to evaluate A and B. To get those, we evaluate the posterior. Since we are considering a normal likelihood and a normal prior, we can use sec 12.10.4 in the book. Here we have expressions for how to specify the sufficient statistics fo the posterior. From eq. (12.138) and eq. (12.139) we have[1], 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

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

Then we get the following expressions

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

Since we can now completely compute $\mathcal{Q}\big(\alpha, \beta; \alpha^{(j)}, \beta^{(j)}\big)$, we have completed the derivation of the E step.

Exercise 9.2.5

The next step is the maximization step where we maximize $\mathcal{Q}\big(\alpha, \beta; \alpha^{(j)}, \beta^{(j)}\big)$ w.r.t. α and β. We can derive closed form updates for these by taking the derivative to $\mathcal{Q}\big(\alpha, \beta; \alpha^{(j)}, \beta^{(j)}\big)$ w.r.t. α and β respectively and set those to zero. Show that the closed form solutions are

$$\begin{aligned} \alpha &= \frac{K}{A} \\ \beta &= \frac{N}{B} \end{aligned}$$

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

We have now completed the derivation of the update formulas for the EM algorithm for Bayesian linear regression.

9.3 Bayesian linear regression on real data

In this exercise we will try out the formulas in two settings: 1) we know all the parameters (which will not be the case in real-life), and see how that affects the regressions, 2) learn the parameters from data using the EM.

Note, since our model corresponds to Ridge regression an from the first exercise we had $\lambda = \frac{\sigma_\eta}{\sigma_\theta}$, EM is effectively learning the regularization parameter without performing cross-validation!

[1]The book is inconsistent and write $\boldsymbol{u}_{z|t}$ in two different ways in sec 12.5 and 12.10. Both are correct though!. We use the form used in 12.5

Exercise 9.3.1

This example implements example 12.1 from the book, be sure to read that example before carrying out this exercise. Inspect and run the code corresponding to this exercise. Relate the code to the formulas you have derived, and try and reproduce all three plots on Figure 12.1.

Exercise 9.3.2

This example implements example 12.2 from the book, be sure to read that example before carrying out this exercise. Inspect and run the code corresponding to this exercise. Relate the code to the formulas you have derived. Play with different parameters for the noise terms and data set. You can also try to carry out ridge regression using cross-validation, and see to what extent there is correspondence with the estimate of λ by EM and cross-validation.

HINTS

Exercise 9.1.4

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