Chapter 3

Maximum likelihood and Bayesian parameter estimation

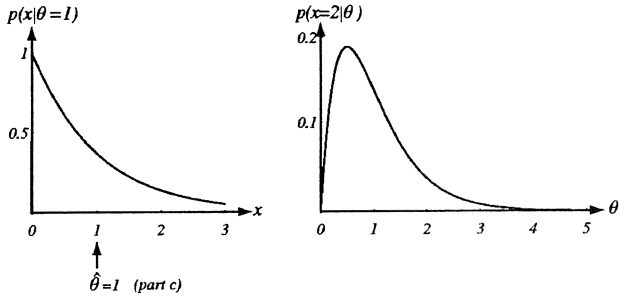
Problem Solutions

Section 3.2

1. Our exponential function is:

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) SEE FIGURE. Note that $p(x = 2|\theta)$ is not maximized when $\theta = 2$ but instead for a value less than 1.0.



(b) The log-likelihood function is

$$l(\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta) = \sum_{k=1}^{n} [\ln \theta - \theta x_k] = n \ln \theta - \theta \sum_{k=1}^{n} x_k.$$

We solve $\nabla_\theta l(\theta) = 0$ to find $\hat{\theta}$ as

$$\begin{aligned}\nabla_\theta l(\theta) &= \frac{\partial}{\partial\theta}\left[n \ln \theta - \theta \sum_{k=1}^{n} x_k\right] \\ &= \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0.\end{aligned}$$

Thus the maximum-likelihood solution is

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{k=1}^{n} x_k}.$$

(c) Here we approximate the mean

$$\frac{1}{n}\sum_{k=1}^{n} x_n$$

by the integral

$$\int_0^\infty x p(x) dx,$$

which is valid in the large n limit. Noting that

$$\int_0^\infty x e^{-x} dx = 1,$$

we put these results together and see that $\hat{\theta} = 1$, as shown on the figure in part (a).

2. Our (normalized) distribution function is

$$p(x|\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

(a) We will use the notation of an *indicator function* $I(\cdot)$, whose value is equal to 1.0 if the logical value of its argument is TRUE, and 0.0 otherwise. We can write the likelihood function using $I(\cdot)$ as

$$\begin{aligned}p(\mathcal{D}|\theta) &= \prod_{k=1}^{n} p(x_k|\theta) \\ &= \prod_{k=1}^{n} \frac{1}{\theta} I\left(0 \le x_k \le \theta\right) \\ &= \frac{1}{\theta^n} I\left(\theta \ge \max_k x_k\right)\ I\left(\min_k x_k \ge 0\right).\end{aligned}$$

We note that $1/\theta^n$ decreases monotonically as θ increases but also that $I(\theta \ge \max_k x_k)$ is 0.0 if θ is less than the maximum value of x_k. Therefore, our likelihood function is maximized at $\hat{\theta} = \max_k x_k$.

We solve this equation and find

$$(1-\hat{P}(\omega_i))\sum_{k=1}^{n} z_{ik} = \hat{P}(\omega_i)\sum_{k=1}^{n}(1-z_{ik}),$$

which can be rewritten as

$$\sum_{k=1}^{n} z_{ik} = \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik} + n\hat{P}(\omega_i) - \hat{P}(\omega_i)\sum_{k=1}^{n} z_{ik}.$$

The final solution is then

$$\hat{P}(\omega_i) = \frac{1}{n}\sum_{k=1}^{n} z_{ik}.$$

That is, the estimate of the probability of category ω_i is merely the probability of obtaining its indicatory value in the training data, just as we would expect.

4. We have n samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from the discrete distribution

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}.$$

The likelihood for a particular sequence of n samples is

$$P(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ki}}(1-\theta_i)^{1-x_{ki}},$$

and the log-likelihood function is then

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n}\sum_{i=1}^{d} x_{ki} \ln \theta_i + (1-x_{ki}) \ln (1-\theta_i).$$

To find the maximum of $l(\boldsymbol{\theta})$, we set $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0}$ and evaluate component by component ($i = 1, \ldots, d$) and get

$$\begin{aligned}
[\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})]_i &= \nabla_{\theta_i} l(\theta) \\
&= \frac{1}{\theta_i}\sum_{k=1}^{n} - \frac{1}{1-\theta_i}\sum_{k=1}^{n}(1-x_{ki}) \\
&= 0.
\end{aligned}$$

This implies that for any i

$$\frac{1}{\hat{\theta}_i}\sum_{k=1}^{n} x_{ki} = \frac{1}{1-\hat{\theta}_i}\sum_{k=1}^{n}(1-x_{ki}),$$

which can be rewritten as

$$(1-\hat{\theta}_i)\sum_{k=1}^{n} x_{ki} = \hat{\theta}_i\left(n - \sum_{k=1}^{n} x_{ki}\right).$$

The final solution is then

$$\hat{\theta}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}.$$

Since this result is valid for all $i = 1, \ldots, d$, we can write this last equation in vector form as

$$\hat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k.$$

Thus the maximum-likelihood value of $\boldsymbol{\theta}$ is merely the sample mean, just as we would expect.

5. The probability of finding feature x_i to be 1.0 in category ω_1 is denoted p:

$$p(x_i = 1|\omega_1) = 1 - p(x_i = 0|\omega_1) = p_{i1} = p > \frac{1}{2},$$

for $i = 1, \ldots, d$. Moreover, the normalization condition gives $p_{i2} = p(x_i|\omega_2) = 1 - p_{i1}$.

(a) A single observation $\mathbf{x} = (x_1, \ldots, x_d)$ is drawn from class ω_1, and thus have

$$p(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p(x_i|\omega_1) = \prod_{i=1}^{d} p^{x_i}(1-p)^{1-x_i},$$

and the log-likelihood function for p is

$$l(p) = \ln p(\mathbf{x}|\omega_1) = \sum_{i=1}^{d}[x_i \ln p + (1 - x_i) \ln (1-p)].$$

Thus the derivative is

$$\nabla_p l(p) = \frac{1}{p}\sum_{i=1}^{d} x_i - \frac{1}{(1-p)}\sum_{i=1}^{d}(1 - x_i).$$

We set this derivative to zero, which gives

$$\frac{1}{\hat{p}}\sum_{i=1}^{d} x_i = \frac{1}{1-\hat{p}}\sum_{i=1}^{d}(1 - x_i),$$

which after simple rearrangement gives

$$(1-\hat{p})\sum_{i=1}^{d} x_i = \hat{p}\left(d - \sum_{i=1}^{d} x_i\right).$$

Thus our final solution is

$$\hat{p} = \frac{1}{d}\sum_{i=1}^{d} x_i.$$

That is, the maximum-likelihood estimate of the probability of obtaining a 1 in any position is simply the ratio of the number of 1's in a single sample divided by the total number of features, given that the number of features is large.

where we have assumed $d\tau/dx \neq 0$ at $\theta = \hat{\theta}$. In short, then, the maximum-likelihood value of $\tau(\theta)$ is indeed $\hat{\theta}$. In practice, however, we must check whether the value of $\hat{\theta}$ derived this way gives a maximum or a minimum (or possibly inflection point) for $p(\tau|\theta)$.

10. Consider the novel method of estimating the mean of a set of points as taking its first value, which we denote $\mathbf{M} = \mathbf{x}_1$.

(a) Clearly, this unusual estimator of the mean is unbiased, that is, the expected value of this statistic is equal to the true value. In other words, if we repeat the selection of the first point of a data set we have

$$bias = \mathcal{E}[\mathbf{M}] - \boldsymbol{\mu} = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \mathbf{M}(k) - \boldsymbol{\mu} = \mathbf{0},$$

where $\mathbf{M}(k)$ is the first point in data set k drawn from the given distribution.

(b) While the unusual method for estimating the mean may indeed be unbiased, it will generally have large variance, and this is an undesirable property. Note that $\mathcal{E}[(x_i - \mu)^2] = \sigma^2$, and the RMS error, σ, is independent of n. This undesirable behavior is quite different from that of the measurement of

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where we see

$$\begin{aligned}\mathcal{E}[(\bar{x} - \mu)^2] &= \mathcal{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathcal{E}[(x_i - \mu)^2]\right] \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Thus the RMS error, $\sigma/\sqrt{n}$, approches 0 as $1/\sqrt{n}$. Note that there are many superior methods for estimating the mean, for instance the sample mean. (In Chapter 9 we shall see other techniques — ones based on resampling — such as the so-called "bootstrap" and "jackknife" methods.)

11. We assume $p_2(\mathbf{x}) \equiv p(\mathbf{x}|\omega_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ but that $p_1(\mathbf{x}) \equiv p(\mathbf{x}|\omega_1)$ is arbitrary. The Kullback-Leibler divergence from $p_1(\mathbf{x})$ to $p_2(\mathbf{x})$ is

$$D_{KL}(p_1, p_2) = \int p_1(\mathbf{x}) \ln p_1(\mathbf{x}) d\mathbf{x} + \frac{1}{2}\int p_1(\mathbf{x})\left[d\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + (\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] d\mathbf{x},$$

where we used the fact that p_2 is a Gaussian, that is,

$$p_2(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right].$$

We now seek $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to minimize this "distance." We set the derivative to zero and find

$$\frac{\partial}{\partial \boldsymbol{\mu}} D_{KL}(p_1, p_2) = -\int \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) p_1(\mathbf{x}) d\mathbf{x} = \mathbf{0},$$

and this implies

$$\boldsymbol{\Sigma}^{-1} \int p_1(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x} = \mathbf{0}.$$

We assume $\boldsymbol{\Sigma}$ is non-singular, and hence this equation implies

$$\int p_1(\mathbf{x})(\mathbf{x}-\boldsymbol{\mu})d\mathbf{x} = \mathcal{E}_1[\mathbf{x}-\boldsymbol{\mu}] = \mathbf{0},$$

or simply, $\mathcal{E}_1[\mathbf{x}] = \boldsymbol{\mu}$. In short, the mean of the second distribution should be the same as that of the Gaussian.

Now we turn to the covariance of the second distribution. Here for notational convenience we denote $\mathbf{A} = \boldsymbol{\Sigma}$. Again, we take a derivative of the Kullback-Leibler divergence and find:

$$\frac{\partial}{\partial \mathbf{A}} D_{KL}(p_1, p_2) = \mathbf{0} = \int p_1(\mathbf{x}) \left[-\mathbf{A}^{-1} + (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] d\mathbf{x},$$

and thus

$$\mathcal{E}_1 \left[\boldsymbol{\Sigma} - (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right],$$

or

$$\mathcal{E}_1 \left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\right] = \boldsymbol{\Sigma}.$$

In short, the covariance of the second distribution should indeed match that of the Gaussian.

Note that above, in taking the derivative above,

$$\frac{\partial |\mathbf{A}|}{\partial \mathbf{A}} = |\mathbf{A}|\mathbf{A}^{-1}$$

we relied on the fact that $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ is symmetric since $\boldsymbol{\Sigma}$ is a covariance matrix. More generally, for an arbitrary non-singular matrix we would use

$$\frac{\partial |\mathbf{M}|}{\partial \mathbf{M}} = |\mathbf{M}|(\mathbf{M}^{-1})^t.$$

Section 3.3

12. In the text we saw the following results:

1. The posterior density can be computed as

$$p(\mathbf{x}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}.$$

2. $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) = p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D})p(\boldsymbol{\theta}|\mathcal{D}))$.

3. $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) = p(\mathbf{x}|\boldsymbol{\theta})$, that is, the distribution of $\mathbf{x}$ is known completely once we know the value of the parameter vector, regardless of the data $\mathcal{D}$.

4. $p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}$.

These are justified as follows:

1. This statement reflects the conceptual difference between the maximum-likelihood estimator and Bayesian estimator. The Bayesian learning method considers the parameter vector $\boldsymbol{\theta}$ to be a random variable rather than a fixed value, as in maximum-likelihood estimation. The posterior density $p(\mathbf{x}|\mathcal{D})$ also depends upon the probability density $p(\boldsymbol{\theta})$ distributed over the entire $\boldsymbol{\theta}$ space instead of a single value. Therefore, the $p(\mathbf{x}|\mathcal{D})$ is the integration of $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ over the entire parameter space. The maximum-likelihood estimator can be regarded as a special case of Bayesian estimator, where $p(\boldsymbol{\theta})$ is uniformly distributed so that its effect disappears after the integration.

2. The $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ implies two steps in computation. One is the computation of the probability density $\boldsymbol{\theta}$ given the data set $\mathcal{D}$, that is, $p(\boldsymbol{\theta}|\mathcal{D})$. The other is the computation of the probability density of $\mathbf{x}$ given $\boldsymbol{\theta}$, that is, $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D})$. The above two steps are independent of each other, and thus $p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D})$ is the product of the results of the two steps.

3. As mentioned in the text, the selection of $\mathbf{x}$ and that of the training samples $\mathcal{D}$ is done independently, that is, the selection of $\mathbf{x}$ does not depend upont $\mathcal{D}$. Therefore we have $p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{D}) = p(\mathbf{x}|\boldsymbol{\theta})$.

4. We substitute the above relations into Eq. 24 in the text and get Eq. 25.

Section 3.4

13. We seek a novel approach for finding the maximum-likelihood estimate for $\boldsymbol{\Sigma}$.

(a) We first inspect the forms of a general vector $\mathbf{a}$ and matrix $\mathbf{A}$:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \text{ and } \mathbf{A} = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{pmatrix}.$$

Consider the scalar

$$\mathbf{a}^t\mathbf{A}\mathbf{a} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_j A_{ij} a_j.$$

The (i,i)th element of this scalar is $\sum_{j=1}^{n} A_{ij}a_j a_i$, and the trace of $\mathbf{A}\mathbf{a}\mathbf{a}^t$ is the sum of these diagonal elements, that is,

$$\text{tr}\,(\mathbf{A}\mathbf{a}\mathbf{a}^t) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}a_j a_i = \mathbf{a}^t\mathbf{A}\mathbf{a}.$$

(b) We seek to show that the likelihood function can be written as

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}^{-1}|^{n/2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right].$$

We note that $p(\mathbf{x}|\boldsymbol{\Sigma}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is known and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent observations from $p(\mathbf{x}|\boldsymbol{\Sigma})$. Therefore the likelihood is

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) &= \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}^{-1}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})\right] \\ &= \frac{|\boldsymbol{\Sigma}|^{-n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})\right]. \end{aligned}$$

From the results in part (a), with $\mathbf{a} = \mathbf{x}_k - \boldsymbol{\mu}$ and $|\mathbf{A}| = |\boldsymbol{\Sigma}^{-1}|$, we have

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) &= \frac{|\boldsymbol{\Sigma}|^{-n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n} \text{tr}\ \left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right] \\ &= \frac{|\boldsymbol{\Sigma}^{-1}|^{-n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\text{tr}\ \left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right], \end{aligned}$$

where we used $\sum_{k=1}^{n} \text{tr}\ (A_k) = \text{tr}\ \left(\sum_{k=1}^{n} A_k\right)$ and $|\boldsymbol{\Sigma}^{-1}| = |\boldsymbol{\Sigma}|^{-1}$.

(c) Recall our definition of the sample covariance matrix:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t.$$

Here we let $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$, which easily leads to the following equalities

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &= \mathbf{A}\hat{\boldsymbol{\Sigma}}^{-1}, \\ |\boldsymbol{\Sigma}^{-1}| &= |\mathbf{A}\hat{\boldsymbol{\Sigma}}^{-1}| = |\mathbf{A}||\hat{\boldsymbol{\Sigma}^{-1}}| \\ &= |\mathbf{A}||\hat{\boldsymbol{\Sigma}}|^{-1} = \frac{|\mathbf{A}|}{|\hat{\boldsymbol{\Sigma}}|} \\ &= \frac{\lambda_1\lambda_2\cdots\lambda_d}{|\hat{\boldsymbol{\Sigma}}|}, \end{aligned}$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$. We substitute these into our result in part (b) to get

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) &= \frac{|\boldsymbol{\Sigma}^{-1}|^{n/2}}{(2\pi)^{nd/2}} \exp\left[-\frac{1}{2}\text{tr}\ \left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu})(\mathbf{x}_k - \boldsymbol{\mu})^t\right)\right] \\ &= \frac{(\lambda_1 \ldots \lambda_n)^{n/2}}{(2\pi)^{nd/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} \exp\left[-\frac{1}{2}\text{tr}\left(n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\right)\right]. \end{aligned}$$

Note, however, that $\text{tr}[n\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}] = n[\text{tr}\ (\mathbf{A})] = n(\lambda_1 + \cdots + \lambda_d)$, and thus we have

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{(\lambda_1 \ldots \lambda_n)^{n/2}}{(2\pi)^{nd/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} \exp\left[-\frac{n}{2}(\lambda_1 + \cdots + \lambda_d)\right].$$

(d) The expression for $p(\mathbf{x}_1, \cdots, \mathbf{x}_n|\boldsymbol{\Sigma})$ in part (c) depends on $\boldsymbol{\Sigma}$ only through $\lambda_1, \cdots, \lambda_d$, the eigenvalues of $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$. We can write our likelihood, then, as

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{nd/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} \left[\prod_{i=1}^{d} \lambda_i e^{-\lambda_i}\right]^{n/2}.$$

Maximizing $p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ is equivalent to maximizing $\lambda_i e^{-\lambda_i}$ with respect to λ_i. We do this by setting the derivative to zero, that is,

$$\frac{\partial[\lambda_i e^{-\lambda_i}]}{\partial \lambda_i} = e^{-\lambda_i} + \lambda_i(-e^{-\lambda_i}) = 0,$$

which has solution $\lambda_i = 1$. In short, $p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\Sigma})$ is maximized by choosing $\lambda_1 = \lambda_2 = \cdots = \lambda_n = 1$. This means that $\mathbf{A} = \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}$, or $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$, as expected.

14. First we note that $p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. We have also $l_k = i$ if the state of nature for $\mathbf{x}_k$ was ω_i.

(a) From Bayes' Rule we can write

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n, l_1, \ldots, l_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) = p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, l_1, \ldots, l_n, \boldsymbol{\Sigma})p(l_1, \ldots, l_n$$

Because the distribution of $l_1, \ldots, l_n$ does not depend on $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c$ or $\boldsymbol{\Sigma}$, we can write

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}, l_1, \ldots, l_n) \\ &= \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}, l_k) \\ &= \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})\right]. \end{aligned}$$

The l_i are independent, and thus the probability density of the ls is a product,

$$p(l_1, \ldots, l_n) = \prod_{k=1}^{n} p(l_k) = \prod_{k=1}^{n} p(\omega_{l_k}).$$

We combine the above equations and get

$$\begin{aligned} &p(\mathbf{x}_1, \ldots, \mathbf{x}_n, l_1, \ldots, l_n|\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) \\ = &\frac{\prod_{k=1}^{n} P(\omega_{l_k})}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})\right]. \end{aligned}$$

(b) We sum the result of part (a) over n samples to find

$$\sum_{k=1}^{n}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{l_k}) = \sum_{i=1}^{l}\sum_{l_k=1}(\mathbf{x}_k - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_i).$$

Here μ_o is formed by averaging n_o fictitious samples x_k for $k = -n_o+1, -n_o+2, \ldots, 0$. Thus we can write

$$\mu_o = \frac{1}{n_o}\sum_{k=-n_o+1}^{0} x_k,$$

and

$$\begin{aligned}\mu_n &= \frac{\sum_{k=1}^{n} x_k}{n+\sigma^2/\sigma_o^2} + \frac{\sigma^2/\sigma_o^2}{\sigma^2/\sigma_o^2+n}\frac{1}{n_o}\sum_{k=-n_o+1}^{0} x_k \\ &= \frac{\sum_{k=1}^{n} x_k}{n+n_o} + \frac{n_o}{n+n_o}\frac{1}{n_o}\sum_{k=1-n_o}^{0} x_k.\end{aligned}$$

We can use the fact that $n_o = \sigma^2/\sigma_o^2$ to write

$$\mu_n = \frac{1}{n+n_o}\sum_{k=-n_o+1}^{n} x_k.$$

Likewise, we have

$$\begin{aligned}\sigma_n^2 &= \frac{\sigma^2\sigma_o^2}{n\sigma_o^2+\sigma^2} \\ &= \frac{\sigma^2}{n+\sigma^2/\sigma_o^2} = \frac{\sigma^2}{n+n_o}.\end{aligned}$$

(b) The result of part (a) can be interpreted as follows: For a suitable choice of the prior density $p(\mu) \sim N(\mu_o, \sigma_o^2)$, maximum-likelihood inference on the "full" sample on $n+n_o$ observations coincides with Bayesian inference on the "second sample" of n observations. Thus, by suitable choice of prior, Bayesian learning can be interpreted as maximum-likelihood learning and here the suitable choice of prior in Bayesian learning is

$$\begin{aligned}\mu_o &= \frac{1}{n_o}\sum_{k=-n_o+1}^{0} x_k, \\ \sigma_o^2 &= \frac{\sigma^2}{n_o}.\end{aligned}$$

Here μ_o is the sample mean of the first n_o observations and σ_o^2 is the variance based on those n_o observations.

16. We assume that $\mathbf{A}$ and $\mathbf{B}$ are non-singular matrices of the same order.

(a) Consider Eq. 44 in the text. We write

$$\begin{aligned}\mathbf{A}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{B} &= \mathbf{A}[(\mathbf{A}+\mathbf{B})^{-1}(\mathbf{B}^{-1})^{-1}] = \mathbf{A}[\mathbf{B}^{-1}(\mathbf{A}+\mathbf{B})]^{-1} \\ &= \mathbf{A}[\mathbf{B}^{-1}\mathbf{A}+\mathbf{I}]^{-1} = (\mathbf{A}^{-1})^{-1}(\mathbf{B}^{-1}\mathbf{A}+\mathbf{I})^{-1} \\ &= [(\mathbf{B}^{-1}\mathbf{A}+\mathbf{I})\mathbf{A}^{-1}]^{-1} = (\mathbf{B}^{-1}+\mathbf{A}^{-1})^{-1}.\end{aligned}$$

We interchange the roles of $\mathbf{A}$ and $\mathbf{B}$ in this equation to get our desired answer:

$$\mathbf{B}(\mathbf{A}+\mathbf{B})^{-1}\mathbf{A} = (\mathbf{A}^{-1}+\mathbf{B}^{-1})^{-1}.$$

(b) Recall Eqs. 41 and 42 in the text:

$$\begin{aligned}\boldsymbol{\Sigma}_n^{-1} &= n\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_o^{-1}\\ \boldsymbol{\Sigma}_n^{-1}\boldsymbol{\mu}_n &= n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_n+\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o.\end{aligned}$$

We have solutions

$$\boldsymbol{\mu}_n = \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)\boldsymbol{\mu}_n+\frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_o,$$

and

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma}.$$

Taking the inverse on both sides of Eq. 41 in the text gives

$$\boldsymbol{\Sigma}_n = \left(n\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}_o^{-1}\right)^{-1}.$$

We use the result from part (a), letting $\mathbf{A}=\frac{1}{n}\boldsymbol{\Sigma}$ and $\mathbf{B}=\boldsymbol{\Sigma}_o$ to get

$$\begin{aligned}\boldsymbol{\Sigma}_n &= \frac{1}{n}\boldsymbol{\Sigma}\left(\frac{1}{n}\boldsymbol{\Sigma}+\boldsymbol{\Sigma}_o\right)^{-1}\\ \boldsymbol{\Sigma}_o &= \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma},\end{aligned}$$

which proves Eqs. 41 and 42 in the text. We also compute the mean as

$$\begin{aligned}\boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n(n\boldsymbol{\Sigma}^{-1}\mathbf{m}_n+\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o)\\ &= \boldsymbol{\Sigma}_n n\boldsymbol{\Sigma}^{-1}\mathbf{m}_n+\boldsymbol{\Sigma}_n\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o\\ &= \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\frac{1}{n}\boldsymbol{\Sigma}n\boldsymbol{\Sigma}^{-1}\mathbf{m}_n+\frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma}_o\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o\\ &= \boldsymbol{\Sigma}_o\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\mathbf{m}_n+\frac{1}{n}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}_o+\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\mu}_o.\end{aligned}$$

Section 3.5

17. The Bernoulli distribution is written

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{d}\theta_i^{x_i}(1-\theta_i)^{1-x_i}.$$

Let $\mathcal{D}$ be a set of n samples $\mathbf{x}_1,\ldots,\mathbf{x}_n$ independently drawn according to $p(\mathbf{x}|\boldsymbol{\theta})$.