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:

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

We have solutions

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

and

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

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

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

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

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

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

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

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 $p(\mathbf{x}|\boldsymbol{\theta})$.

(a) We denote $\mathbf{s} = (s_1, \cdots, s_d)^t$ as the sum of the n samples. If we denote $\mathbf{x}_k = (x_{k1}, \cdots, x_{kd})^t$ for $k = 1, \ldots, n$, then $s_i = \sum_{k=1}^{n} x_{ki}, i = 1, \ldots, d$, and the likelihood is

$$
\begin{aligned}
P(\mathcal{D}|\boldsymbol{\theta}) &= P(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}) = \underbrace{\prod_{k=1}^{n} P(\mathbf{x}_k|\boldsymbol{\theta})}_{\mathbf{x}_k \text{ are indep.}} \\
&= \prod_{k=1}^{n}\prod_{i=1}^{d} \theta_i^{x_{ki}}(1-\theta_i)^{1-x_{ki}} \\
&= \prod_{i=1}^{d} \theta_i^{\sum_{k=1}^{n} x_{ki}}(1-\theta_i)^{\sum_{k=1}^{n}(1-x_{ki})} \\
&= \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}.
\end{aligned}
$$

(b) We assume an (unnormalized) uniform prior for $\boldsymbol{\theta}$, that is, $p(\boldsymbol{\theta}) = 1$ for $0 \leq \theta_i \leq 1$ for $i = 1, \cdots, d$, and have by Bayes' Theorem

$$
p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}.
$$

From part (a), we know that $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta)^{n-s_i}$, and therefore the probability density of obtaining data set $\mathcal{D}$ is

$$
\begin{aligned}
p(\mathcal{D}) &= \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\boldsymbol{\theta} \\
&= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta_1 d\theta_2 \cdots d\theta_d \\
&= \prod_{i=1}^{d} \int_0^1 \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta_i.
\end{aligned}
$$

Now $s_i = \sum_{k=1}^{n} x_{ki}$ takes values in the set $\{0, 1, \ldots, n\}$ for $i = 1, \ldots, d$, and if we use the identity

$$
\int_0^1 \theta^m(1-\theta)^n d\theta = \frac{m!n!}{(m+n+1)!},
$$

and substitute into the above equation, we get

$$
p(\mathcal{D}) = \prod_{i=1}^{d} \int_0^1 \theta_i^{s_i}(1-\theta_i)^{n-s_i} d\theta_i = \prod_{i=1}^{d} \frac{s_i!(n-s_i)!}{(n+1)!}.
$$

We consolidate these partial results and find

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &= \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \\ &= \frac{\prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}}{\prod_{i=1}^{d} s_i!(n-s_i)!/(n+1)!} \\ &= \prod_{i=1}^{d} \frac{(n+1)!}{s_i!(n-s_i)!}\theta_i^{s_i}(1-\theta_i)^{n-s_i}. \end{aligned}$$

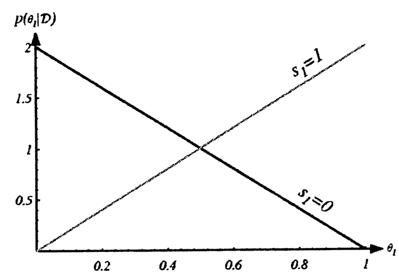
(c) We have $d=1, n=1$, and thus

$$p(\theta_1|\mathcal{D}) = \frac{2!}{s_1!(n-s_1)!}\theta_1^{s_1}(1-\theta_1)^{n-s_1} = \frac{2}{s_1!(1-s_1)!}\theta_1^{s_1}(1-\theta_1)^{1-s_1}.$$

Note that s_1 takes the discrete values 0 and 1. Thus the densities are of the form

$$\begin{aligned} s_1 = 0 &: \quad p(\theta_1|\mathcal{D}) = 2(1-\theta_1) \\ s_1 = 1 &: \quad p(\theta_1|\mathcal{D}) = 2\theta_1, \end{aligned}$$

for $0 \le \theta_1 \le 1$, as shown in the figure.



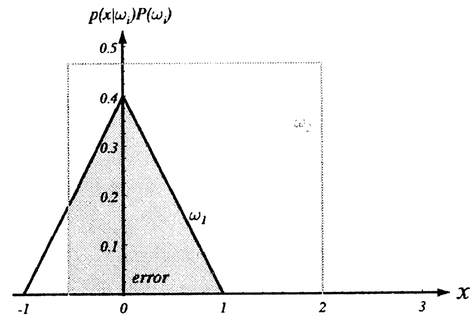
18. Consider how knowledge of an invariance can guide our choice of priors.

(a) We are given that s is actually the number of times that $x=1$ in the first n tests. Consider the $(n+1)$st test. If again $x=1$, then there are $\binom{n+1}{s+1}$ permutations of 0s and 1s in the $(n+1)$ tests, in which the number of 1s is $(s+1)$. Given the assumption of invariance of exchangeablility (that is, all permutations have the same chance to appear), the probability of each permutation is

$$P_{instance} = \frac{1}{\binom{n+1}{s+1}}.$$

Therefore, the probability of $x=1$ after n tests is the product of two probabilities: one is the probability of having $(s+1)$ number of 1s, and the other is the probability for a particular instance with $(s+1)$ number of 1s, that is,

$$\Pr[x_{n+1}=1|\mathcal{D}^n] = \Pr[x_1+\cdots+x_n = s+1]\cdot P_{instance} = \frac{p(s+1)}{\binom{n+1}{s+1}}.$$



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

Note that the probability density of the parameter $\boldsymbol{\theta}$ is fully specified by the sufficient statistic; the data gives no further information, and this implies

$$p(\theta|\mathbf{s},\mathcal{D}) = p(\theta|\mathbf{s}).$$

Since $p(\boldsymbol{\theta}|\mathbf{s}) \neq 0$, we can write

$$\begin{aligned} p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta}) &= \frac{p(\boldsymbol{\theta}|\mathbf{s},\mathcal{D})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})} \\ &= \frac{p(\boldsymbol{\theta}|\mathbf{s})p(\mathcal{D}|\mathbf{s})}{p(\boldsymbol{\theta}|\mathbf{s})} \\ &= p(\mathcal{D}|\mathbf{s}), \end{aligned}$$

which does not involve $\boldsymbol{\theta}$. Thus, $p(\mathcal{D}|\mathbf{s},\boldsymbol{\theta})$ is indeed independent of $\boldsymbol{\theta}$.

24. To obtain the maximum-likelihood estimate, we must maximize the likelihood function $p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x}_1,\ldots,\mathbf{x}_n|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. However, by the Factorization Theorem (Theorem 3.1) in the text, we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = g(\mathbf{s},\boldsymbol{\theta})h(\mathcal{D}),$$

where $\mathbf{s}$ is a sufficient statistic for $\boldsymbol{\theta}$. Thus, if we maximize $g(\mathbf{s},\boldsymbol{\theta})$ or equivalently $[g(\mathbf{s},\boldsymbol{\theta})]^{1/n}$, we will have the maximum-likelihoood solution we seek.

For the Rayleigh distribution, we have from Table 3.1 in the text,

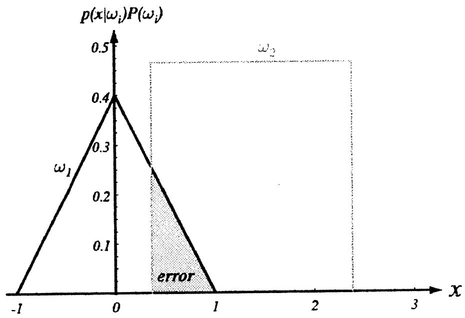
$$[g(s,\theta)]^{1/n} = \theta e^{-\theta s}$$

for $\theta > 0$, where

$$s = \frac{1}{n}\sum_{k=1}^{n} x_k^2.$$

Then, we take the derivative with respect to θ and find

$$\nabla_\theta [g(s,\theta)]^{1/n} = e^{-\theta s} - s\theta e^{-\theta s}.$$



We set this to 0 and solve to get

$$e^{-\hat{\theta}s} = s\hat{\theta}e^{-\hat{\theta}s},$$

which gives the maximum-likelihood solution,

$$\hat{\theta} = \frac{1}{s} = \left(\frac{1}{n}\sum_{k=1}^{n} x_k^2\right)^{-1}.$$

We next evaluate the second derivative at this value of $\hat{\theta}$ to see if the solution represents a maximum, a minimum, or possibly an inflection point:

$$\begin{aligned}\nabla_\theta^2[g(s,\theta)]^{1/n}\Big|_{\theta=\hat{\theta}} &= -se^{-\theta s} - se^{-\theta s} + s^2\theta e^{-\theta s}\Big|_{\theta=\hat{\theta}} \\ &= e^{-\hat{\theta}s}(s^2\hat{\theta} - 2s) = -se^{-1} < 0.\end{aligned}$$

Thus $\hat{\theta}$ indeed gives a maximum (and not a minimum or an inflection point).

25. The maximum-likelihood solution is obtained by maximizing $[g(\mathbf{s},\theta)]^{1/n}$. From Table 3.1 in the text, we have for a Maxwell distribution

$$[g(s,\theta)]^{1/n} = \theta^{3/2}e^{-\theta s}$$

where $s = \frac{1}{n}\sum\limits_{k=1}^{n} x_k^2$. The derivative is

$$\nabla_\theta[g(s,\theta)]^{1/n} = \frac{3}{2}\theta^{1/2}e^{-\theta s} - s\theta^{3/2}e^{-\theta s}.$$

We set this to zero to obtain

$$\frac{3}{2}\theta^{1/2}e^{-\theta s} = s\theta^{3/2}e^{-\theta s},$$

and thus the maximum-likelihood solution is

$$\hat{\theta} = \frac{3/2}{s} = \frac{3}{2}\left(\frac{1}{n}\sum_{k=1}^{n} x_k^2\right)^{-1}.$$

We next evaluate the second derivative at this value of $\hat{\theta}$ to see if the solution represents a maximum, a minimum, or possibly an inflection point:

$$\nabla_\theta^2[g(s,\theta)]^{1/n}\Big|_{\theta=\hat{\theta}} = \frac{3}{2}\frac{1}{2}\theta^{1/2}e^{-\theta s} - \frac{3}{2}\theta^{1/2}se^{-\theta s} - \frac{3}{2}\theta^{1/2}se^{-\theta s} + s^2\theta^{3/2}e^{-\theta s}\Big|_{\theta=\hat{\theta}}$$

where $\mathbf{u} = \mathbf{C}_n^{-1}(\mathbf{x}_{n+1} - \mathbf{m}_n)$ is of $O(d^2)$ complexity, given that $\mathbf{C}_n^{-1}, \mathbf{x}_{n+1}$ and $\mathbf{m}_n$ are known. Hence, clearly $\mathbf{C}_n^{-1}$ can be computed from $\mathbf{C}_{n-1}^{-1}$ in $O(d^2)$ operations, as $\mathbf{u}\mathbf{u}^t, \mathbf{u}^t(\mathbf{x}_{n+1} - \mathbf{m}_n)$ is computed in $O(d^2)$ operations. The complexity associated with determining $\mathbf{C}_n^{-1}$ is $O(nd^2)$.

37. We assume the symmetric non-negative covariance matrix is of otherwise general form:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}.$$

To employ shrinkage of an assumed common covariance toward the identity matrix, then Eq. 77 requires

$$\mathbf{\Sigma}(\beta) = (1-\beta)\mathbf{\Sigma} + \beta\mathbf{I} = \mathbf{I},$$

and this implies $(1-\beta)\sigma_{ii} + \beta \cdot 1 = 1$, and thus

$$\sigma_{ii} = \frac{1-\beta}{1-\beta} = 1$$

for all $0 < \beta < 1$. Therefore, we must first normalize the data to have unit variance.

Section 3.8

38. Note that in this problem our densities need not be normal.

(a) Here we have the criterion function

$$J_1(\mathbf{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}.$$

We make use of the following facts for $i = 1, 2$:

$$\begin{aligned} y &= \mathbf{w}^t\mathbf{x} \\ \mu_i &= \frac{1}{n_i}\sum_{y\in\mathcal{Y}_i} y = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i}\mathbf{w}^t\mathbf{x} = \mathbf{w}^t\boldsymbol{\mu}_i \\ \sigma_i^2 &= \sum_{y\in\mathcal{Y}_i}(y-\mu_i)^2 = \mathbf{w}^t\left[\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^t\right]\mathbf{w} \\ \mathbf{\Sigma}_i &= \sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{x}-\boldsymbol{\mu}_i)^t. \end{aligned}$$

We define the within- and between-scatter matrices to be

$$\begin{aligned} \mathbf{S}_W &= \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 \\ \mathbf{S}_B &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t. \end{aligned}$$

Then we can write

$$\begin{aligned} \sigma_1^2 + \sigma_2^2 &= \mathbf{w}^t\mathbf{S}_W\mathbf{w} \\ (\mu_1 - \mu_2)^2 &= \mathbf{w}^t\mathbf{S}_B\mathbf{w}. \end{aligned}$$

The criterion function can be written as

$$J_1(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}.$$

For the same reason Eq. 103 in the text is maximized, we have that $J_1(\mathbf{w})$ is maximized at $\mathbf{w}\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. In sum, that $J_1(\mathbf{w})$ is maximized at $\mathbf{w} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

(b) Consider the criterion function

$$J_2(\mathbf{w}) = \frac{(\mu_1 - \mu_2)^2}{P(\omega_1)\sigma_1^2 + P(\omega_2)\sigma_2^2}.$$

Except for letting $\mathbf{S}_W = P(\omega_1)\boldsymbol{\Sigma}_1 + P(\omega_2)\boldsymbol{\Sigma}_2$, we retain all the notations in part (a). Then we write the criterion function as a Rayleigh quotient

$$J_2(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}.$$

For the same reason Eq. 103 is maximized, we have that $J_2(\mathbf{w})$ is maximized at

$$\mathbf{w} = (P(\omega_1)\boldsymbol{\Sigma}_1 + P(\omega_2)\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

(c) Equation 96 of the text is more closely related to the criterion function in part (a) above. In Eq. 96 in the text, we let $\tilde{m}_i = \mu_i$, and $\tilde{s}_i^2 = \sigma_i^2$ and the statistical meanings are unchanged. Then we see the exact correspondence between $J(\mathbf{w})$ and $J_1(\mathbf{w})$.

39. The expression for the criterion function

$$J_1 = \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}(y_i - y_j)^2$$

clearly measures the total within-group scatter.

(a) We can rewrite J_1 by expanding

$$\begin{aligned}
J_1 &= \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}[(y_i - m_1) - (y_j - m_2) + (m_1 - m_2)]^2 \\
&= \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}[(y_i - m_1)^2 + (y_j - m_2)^2 + (m_1 - m_2)^2 \\
&\qquad +2(y_i - m_1)(y_j - m_2) + 2(y_i - m_1)(m_1 - m_2) + 2(y_j - m_2)(m_1 - m_2)] \\
&= \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}(y_i - m_1)^2 + \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}(y_j - m_2)^2 + (m_1 - m_2)^2 \\
&\qquad +\frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}2(y_i - m_1)(y_j - m_2) + \frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}2(y_i - m_1)(m_1 - m_2) \\
&\quad +\frac{1}{n_1 n_2}\sum_{y_i\in\mathcal{Y}_1}\sum_{y_j\in\mathcal{Y}_2}2(y_j - m_2)(m_1 - m_2) \\
&= \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2 + (m_1 - m_2)^2,
\end{aligned}$$

(c) We make the following definitions:

$$\begin{aligned}\tilde{\mathbf{W}}^t &= \mathbf{QDW}^t\\ \tilde{\tilde{\mathbf{S}}}_W &= \tilde{\mathbf{W}}^t\mathbf{S}_W\tilde{\mathbf{W}} = \mathbf{QDW}^t\mathbf{S}_W\mathbf{WDQ}^t.\end{aligned}$$

Then we have $|\tilde{\tilde{\mathbf{S}}}_W| = |\mathbf{D}|^2$ and

$$\tilde{\tilde{\mathbf{S}}}_B = \tilde{\mathbf{W}}^t\mathbf{S}_B\tilde{\mathbf{W}} = \mathbf{QDW}^t\mathbf{S}_B\mathbf{WDQ}^t = \mathbf{QD}\tilde{\mathbf{S}}_B\mathbf{DQ}^t,$$

then $|\tilde{\tilde{\mathbf{S}}}_B| = |\mathbf{D}|^2\lambda_1\lambda_2\cdots\lambda_n$. This implies that the criterion function obeys

$$J = \frac{|\tilde{\tilde{\mathbf{S}}}_B|}{|\tilde{\tilde{\mathbf{S}}}_W|},$$

and thus J is invariant to this transformation.

41. Our two Gaussian distributions are $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1, 2$. We denote the samples after projection as $\tilde{\mathcal{D}}_i$ and the distributions

$$p(y|\tilde{\boldsymbol{\theta}}_i) = \frac{1}{\sqrt{2\pi}\tilde{s}}\exp[-(y-\tilde{\mu})^2/(2\tilde{s}^2)],$$

and $\tilde{\boldsymbol{\theta}}_i = \binom{\tilde{\mu}_i}{\tilde{s}}$ for $i = 1, 2$. The log-likelihood ratio is

$$\begin{aligned}
r &= \frac{\ln p(\tilde{\mathcal{D}}|\tilde{\boldsymbol{\theta}}_1)}{\ln p(\tilde{\mathcal{D}}|\tilde{\boldsymbol{\theta}}_2)} = \frac{\ln\left[\prod\limits_{k=1}^{n} p(y_k|\tilde{\boldsymbol{\theta}}_1)\right]}{\ln\left[\prod\limits_{k=1}^{n} p(y_k|\tilde{\boldsymbol{\theta}}_2)\right]}\\
&= \frac{\sum\limits_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\exp\left[\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}\right]\right]}{\sum\limits_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\exp\left[\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}\right]\right]} = \frac{\sum\limits_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\right] + \sum\limits_{k=1}^{n}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}}{\sum\limits_{k=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\tilde{s}}\right] + \sum\limits_{k=1}^{n}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}}\\
&= \frac{c_1 + \sum\limits_{y_k\in\mathcal{D}_1}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2} + \sum\limits_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}}{c_1 + \sum\limits_{y_k\in\mathcal{D}_1}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2} + \sum\limits_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}}\\
&= \frac{c_1 + \frac{1}{2} + \sum\limits_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_1)^2}{2\tilde{s}^2}}{c_1 + \frac{1}{2} + \sum\limits_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_2)^2}{2\tilde{s}^2}} = \frac{c_1 + \frac{1}{2} + \sum\limits_{y_k\in\mathcal{D}_2}\frac{(y_k-\tilde{\mu}_2)+(\tilde{\mu}_2-\tilde{\mu}_1))^2}{2\tilde{s}^2}}{c_1 + \frac{1}{2} + \sum\limits_{y_k\in\mathcal{D}_1}\frac{(y_k-\tilde{\mu}_2)+(\tilde{\mu}_2-\tilde{\mu}_1))^2}{2\tilde{s}^2}}\\
&= \frac{c_1 + \frac{1}{2} + \frac{1}{2\tilde{s}^2}\sum\limits_{y_k\in\tilde{\mathcal{D}}_2}\left((y_k-\tilde{\mu}_2)^2 + (\tilde{\mu}_2-\tilde{\mu}_1)^2 + 2(y_k-\tilde{\mu}_2)(\tilde{\mu}_2-\tilde{\mu}_1)\right)}{c_1 + \frac{1}{2} + \frac{1}{2\tilde{s}^2}\sum\limits_{y_k\in\tilde{\mathcal{D}}_1}\left((y_k-\tilde{\mu}_1)^2 + (\tilde{\mu}_1-\tilde{\mu}_2)^2 + 2(y_k-\tilde{\mu}_1)(\tilde{\mu}_1-\tilde{\mu}_2)\right)}\\
&= \frac{c_1 + 1 + \frac{1}{2\tilde{s}^2}n_2(\tilde{\mu}_2-\tilde{\mu}_1)^2}{c_1 + 1 + \frac{1}{2\tilde{s}^2}n_1(\tilde{\mu}_1-\tilde{\mu}_2)^2} = \frac{c + n_2J(\mathbf{w})}{c + n_1J(\mathbf{w})}.
\end{aligned}$$

Thus we can write the criterion function as

$$J(\mathbf{w}) = \frac{rc - c}{n_2 - rn_1}.$$

This implies that the Fisher linear discriminant can be derived from the negative of the log-likelihood ratio.

42. Consider the criterion function $J(\mathbf{w})$ required for the Fisher linear discriminant.

(a) We are given Eqs. 96, 97, and 98 in the text:

$$\begin{aligned} J_1(\mathbf{w}) &= \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \quad (96) \\ \mathbf{S}_i &= \sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \quad (97) \\ \mathbf{S}_W &= \mathbf{S}_1 + \mathbf{S}_2 \quad (98) \end{aligned}$$

where $y = \mathbf{w}^t\mathbf{x}$, $\tilde{m}_i = 1/n_i \sum_{y\in\mathcal{Y}_i} y = \mathbf{w}^t\mathbf{m}_i$. From these we can write Eq. 99 in the text, that is,

$$\begin{aligned} \tilde{s}_i^2 &= \sum_{y\in\mathcal{Y}_i} (y - \tilde{m}_i)^2 \\ &= \sum_{\mathbf{x}\in\mathcal{D}} (\mathbf{w}^t\mathbf{x} - \mathbf{w}^t\mathbf{m}_i)^2 \\ &= \sum_{\mathbf{x}\in\mathcal{D}} \mathbf{w}^t(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t\mathbf{w} \\ &= \mathbf{w}^t\mathbf{S}_i\mathbf{w}. \end{aligned}$$

Therefore, the sum of the scatter matrixes can be written as

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^t\mathbf{S}_W\mathbf{w} \qquad (100)$$

$$\begin{aligned} (\tilde{m}_1 - \tilde{m}_2)^2 &= (\mathbf{w}^t\mathbf{m}_1 - \mathbf{w}^t\mathbf{m}_2)^2 \qquad (101) \\ &= \mathbf{w}^t(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\mathbf{w} \\ &= \mathbf{w}^t\mathbf{S}_B\mathbf{w}, \end{aligned}$$

where $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$, as given by Eq. 102 in the text. Putting these together we get Eq. 103 in the text,

$$J(\mathbf{w}) = \frac{\mathbf{w}^t\mathbf{S}_B\mathbf{w}}{\mathbf{w}^t\mathbf{S}_W\mathbf{w}}. \qquad (103)$$

(b) Part (a) gave us Eq. 103. It is easy to see that the $\mathbf{w}$ that optimizes Eq. 103 is not unique. Here we optimize $J_1(\mathbf{w}) = \mathbf{w}^t\mathbf{S}_B\mathbf{w}$ subject to the constraint that $J_2(\mathbf{w}) = \mathbf{w}^t\mathbf{S}_W\mathbf{w} = 1$. We use the method of Lagrange undetermined multipliers and form the functional

$$g(\mathbf{w}, \lambda) = J_1(\mathbf{w}) - \lambda(J_2(\mathbf{w}) - 1).$$

We set its derivative to zero, that is,

$$\begin{aligned} \frac{\partial g(\mathbf{w}, \lambda)}{\partial w_i} &= \left(\mathbf{u}_i^t\mathbf{S}_B\mathbf{w} + \mathbf{w}^t\mathbf{S}_B\mathbf{u}_i\right) - \lambda\left(\mathbf{u}_i^t\mathbf{S}_W\mathbf{w} + \mathbf{w}^t\mathbf{S}_w\mathbf{u}_i\right) \\ &= 2\mathbf{u}_i^t(\mathbf{S}_B\mathbf{w} - \lambda\mathbf{S}_W\mathbf{w}) = 0, \end{aligned}$$

where $\mathbf{u}_i = (0 \ 0 \ \cdots \ 1 \ \cdots \ 0 \ 0)^t$ is the n-dimensional unit vector in the ith direction. This equation implies

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w}.$$

$\alpha_i(t)$'s and $P(V^T|M)$ are computed by the `Forward Algorithm`, which requires $O(c^2T)$ operations. The $\beta_i(t)$'s can be computed recursively as follows:

For t=T to 1 (by -1)
For i=1 to c
$\beta_i(t) = \sum_j a_{ij}b_{jk}v(t+1)\beta_j(t+1)$
End

This requires $O(c^2T)$ operations.

Similarly, γij's can be computed by $O(c^2T)$ operations given $\alpha_i(t)$'s, a_{ij}'s, b_{ij}'s, $\beta_i(t)$'s and $P(V^T|M)$. So, $\gamma_{ij}(t)$'s are computed by

$$\underbrace{O(c^2T)}_{\alpha_i(t)\text{'s}} + \underbrace{O(c^2T)}_{\beta_i(t)\text{'s}} + \underbrace{O(c^2T)}_{\gamma_{ij}(t)\text{'s}} = O(c^2T)\text{operations.}$$

Then, given $\hat{\gamma}_{ij}(t)$'s, the $\hat{a}_{ij}$'s can be computed by $O(c^2T)$ operations and $\hat{b}_{ij}$'s by $O(c^2T)$ operations. Therefore, a single revision requires $O(c^2T)$ operations.

50. The standard method for calculating the probability of a sequence in a given HMM is to use the forward probabilities $\alpha_i(t)$.

(a) In the forward algorithm, for $t = 0, 1, \ldots, T$, we have

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \text{ and } j \neq \text{initial status} \\ 1 & t = 0 \text{ and } j = \text{initial status} \\ \sum_{i=1}^{c} \alpha_i(t-1)a_{ij}b_{jk}v(t) & \text{otherwise.} \end{cases}$$

In the backward algorithm, we use for $t = T, T-1, \ldots, 0$,

$$\beta_j(t) = \begin{cases} 0 & t = T \text{ and } j \neq \text{final status} \\ 1 & t = T \text{ and } j = \text{final status} \\ \sum_{i=1}^{c} \beta_i(t+1)a_{ij}b_{jk}v(t+1) & \text{otherwise.} \end{cases}$$

Thus in the forward algorithm, if we first reverse the observed sequence $\mathbf{V}^T$ (that is, set $b_{jk}v(t) = b_{jk}(T+1-t)$ and then set $\beta_j(t) = \alpha_j(T-t)$, we can obtain the backward algorithm.

(b) Consider splitting the sequence $\mathbf{V}^T$ into two parts — $\mathbf{V}_1$ and $\mathbf{V}_2$ — before, during, and after each time step T' where $T' < T$. We know that $\alpha_i(T')$ represents the probability that the HMM is in hidden state ω_i at step T', having generated the firt T' elements of $\mathbf{V}^T$, that is $\mathbf{V}_1$. Likewise, $\beta_i(T')$ represents the probability that the HMM given that it is in ω_i at step T' generates the remaining elements of $\mathbf{V}^T$, that is, $\mathbf{V}_2$. Hence, for the complete sequence we have

$$\begin{aligned} P(\mathbf{V}^T) &= P(\mathbf{V}_1, \mathbf{V}_2) = \sum_{i=1}^{c} P(\mathbf{V}_1, \mathbf{V}_2, \text{hidden state } \omega_i \text{ at step } T') \\ &= \sum_{i=1}^{c} P(\mathbf{V}_1, \text{hidden state } \omega_i \text{ at step } T') P(\mathbf{V}_2 | \text{hidden state } \omega_i \text{ at step } T') \\ &= \sum_{i=1}^{c} \alpha_i(T')\beta_i(T'). \end{aligned}$$

(c) At $T' = 0$, the above reduces to $P(\mathbf{V}^T) = \sum_{i=1}^{c} \alpha_i(0)\beta_i(0) = \beta_j(0)$, where j is the known initial state. This is the same as line 5 in Algorithm 3. Likewise, at $T' = T$, the above reduces to $P(\mathbf{V}^T) = \sum_{i=1}^{c} \alpha_i(T)\beta_i(T) = \alpha_j(T)$, where j is the known final state. This is the same as line 5 in Algorithm 2.

51. From the learning algorithm in the text, we have for a giveen HMM with model parameters $\boldsymbol{\theta}$:

$$\gamma_{ij}(t) = \frac{\alpha_i(t-1)a_{ij}b_{jk}v(t)\beta_j(t)}{P(\mathbf{V}^T|\boldsymbol{\theta})} \qquad (*)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T}\gamma_{ij}(t)}{\sum_{t=1}^{T}\sum_{k=1}^{c}\gamma_{ik}(t)}. \qquad (**)$$

For a new HMM with $a_{i'j'} = 0$, from $(*)$ we have $\gamma_{i'j'} = 0$ for all t. Substituting $\gamma_{i'j'}(t)$ into $(**)$, we have $\hat{a}_{i'j'} = 0$. Therefore, keeping this substitution throughout the iterations in the learning algorithm, we see that $\hat{a}_{i'j'} = 0$ remains unchanged.

52. Consider the decoding algorithm (Algorithm 4).

(a) the algorithm is:

Algorithm 0 (Modified decoding)

begin initialize Path $\leftarrow \{\}, t \leftarrow 0$
  for $t \leftarrow t+1$
    $j \leftarrow 0;\ \delta_0 \leftarrow 0$
    for $j \leftarrow j+1$
      $\delta_j(t) \leftarrow \min_{1\le i\le c}[\delta_i(t-1) - \ln(a_{ij})] - \ln[b_{jk}v(t)]$
    until $j = c$
    $j' \leftarrow \arg\min_j[\delta_j(t)]$
    Append $\omega_{j'}$ *to* Path
  until $t = T$
  return Path
end

(b) Taking the logarithm is an $O(c^2)$ computation since we only need to calculate $\ln a_{ij}$ for all $i, j = 1, 2, \ldots, c$, and $\ln[b_{jk}v(t)]$ for $j = 1, 2, \ldots, c$. Then, the whole complexity of this algorithm is $O(c^2T)$.