A Solution Manual and Study Guide for: Pattern Recognition: A Statistical Approach by Pierre A. Devijver and Josef Kittler.

John L. Weatherwax*

January 1, 2005

*wax@alum.mit.edu

Chapter 5 (Introduction to Feature Selection and Extraction)

Problem Solutions

Problem 2 (the divergence and statistically independent features)

Consider the divergence d_{ij} defined by

$$d_{ij} = \int_{\mathbf{x}} (p(\mathbf{x}|\omega_i) - p(\mathbf{x}|\omega_j)) \ln\left(\frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)}\right) d\mathbf{x} \,. \tag{1}$$

Then if the features are statistically independent in each class we have

$$p(\mathbf{x}|\omega_i) = \prod_{k=1}^{d} p(x_k|\omega_i) \,.$$

Thus the logarithmic term above becomes

$$\begin{aligned}\ln\left(\frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)}\right) d\mathbf{x} &= \ln\left(\prod_{k=1}^{d} \frac{p(x_k|\omega_i)}{p(x_k|\omega_j)}\right) \\ &= \sum_{k=1}^{d} \ln\left(\frac{p(x_k|\omega_i)}{p(x_k|\omega_j)}\right) \,.\end{aligned}$$

Then we get for d_{ij} is

$$d_{ij} = \sum_{k=1}^{d} \int_{\mathbf{x}} (p(\mathbf{x}|\omega_i) - p(\mathbf{x}|\omega_j)) \ln\left(\frac{p(x_k|\omega_i)}{p(x_k|\omega_j)}\right) d\mathbf{x} \,.$$

Since the logarithmic term only depends on x_k (and not the other k's) we can integrate out them by performing the $\int_{\mathbf{x}}$ integration for all variables but x_k. This then gives

$$d_{ij} = \sum_{k=1}^{d} \int_{x_k} (p(x_k|\omega_i) - p(x_k|\omega_j)) \ln\left(\frac{p(x_k|\omega_i)}{p(x_k|\omega_j)}\right) dx_k \,,$$

which is the sum of d scalar divergences each one over a different variable.

Chapter 6 (Interclass Distance Measures In Feature Selection and Extraction)

Notes On The Text

Notes On Discriminant Analsis: The Criterion $J_3(W)$

We want to evaluate

$$\frac{\partial}{\partial W}\text{trace}\{(W^T S_1 W)^{-1}(W^T S_2 W)\}$$

The algebraic procedure for computing derivatives like $\frac{\partial}{\partial W}\text{trace}\{F(W)\}$ where $F(\cdot)$ is a matrix function of a matrix argument is discussed on Page 14. The basic procedure is the following. We consider the matrix derivative as several scalar derivative (one derivative for each component w_{kl} of W). We pass the derivative of w_{kl} though the trace operation and take the scalar derivative of various matrix expressions i.e. $\frac{\partial F(W)}{\partial w_{kl}}$. Taking these derivatives is easier if we introduce the matrix $V(k,l)$ which is a matrix of all zeros except for a single one at the location (k,l). This is a helpful matrix to have since

$$\frac{\partial}{\partial w_{kl}} W = V(k,l)\,.$$

Once we have computed the derivative of the argument of the trace $F(W)$ with respect to w_{kl} we need to write it in the form

$$\frac{\partial F(W)}{\partial w_{kl}} = \sum_i g_i(W)V(k,l)h_i(W)\,.$$

We can then take the trace of the above expression and use the permutability of matrices in the argument of the trace to write

$$\begin{aligned}
\text{trace}\left\{\frac{\partial F(W)}{\partial w_{kl}}\right\} &= \text{trace}\left\{\sum_i g_i(W)V(k,l)h_i(W)\right\} \\
&= \sum_i \text{trace}\left\{g_i(W)V(k,l)h_i(W)\right\} \\
&= \sum_i \text{trace}\left\{h_i(W)g_i(W)V(k,l)\right\}\,. \qquad (2)
\end{aligned}$$

Finally we use the property of the trace to conclude that for any $n \times n$ matrix M

$$MV(k,l) = \left[\begin{array}{ccc} \mathbf{0} & \left[\begin{array}{c} m_{1k} \\ m_{2k} \\ \vdots \\ m_{n-1,k} \\ m_{nk} \end{array}\right] & \mathbf{0} \end{array}\right],$$

or a matrix with the kth column of M in the lth column. Since the only nonzero column is the lth, to take the trace of this matrix, we need to find what the element of the lth row in that column is. From the above we see that this element is m_{lk}. Thus we have just argued that

$$\text{trace}\,\{MV(k,l)\} = M(l,k)\,.$$

When we reassemble all elements, from this result, to compute the full matrix derivative of trace$\{MW\}$ we see that

$$\frac{\partial}{\partial W}\text{trace}\,\{MW\} = M^T\,.$$

Back to Equation 2 we can use the above to get the full matrix derivative

$$\frac{\partial}{\partial W}\text{trace}\{F(W)\} = \sum_i (h_i(W)g_i(W))^T\,. \tag{3}$$

For this problem we now implement this procedure.

To begin we evaluate the w_{kl} derivative of $(W^TS_1W)^{-1}(W^TS_2W)$. From the product rule we have

$$\frac{\partial}{\partial w_{kl}}\left[(W^TS_1W)^{-1}(W^TS_2W)\right] = \left[\frac{\partial}{\partial w_{kl}}(W^TS_1W)^{-1}\right](W^TS_2W) + (W^TS_1W)^{-1}\left[\frac{\partial}{\partial w_{kl}}(W^TS_2W)\right]\,.$$

To evaluate the w_{kl} derivative of $(W^TS_1W)^{-1}$ recall that if $F(W) = G^{-1}(W)$ then

$$\frac{\partial F(W)}{\partial w_{kl}} = -G^{-1}(W)\frac{\partial G(W)}{\partial w_{kl}}G^{-1}(W)\,. \tag{4}$$

Thus we get

$$\frac{\partial (W^TS_1W)^{-1}}{\partial w_{kl}} = -(W^TS_1W)^{-1}\frac{\partial (W^TS_1W)}{\partial w_{kl}}(W^TS_1W)^{-1}\,.$$

Thus we need to evaluate the derivative of W^TS_1W (a similar needed derivative is of W^TS_2W). We get

$$\frac{\partial (W^TS_1W)}{\partial w_{kl}} = V^T(k,l)S_1W + W^TS_1V(k,l)\,.$$

Combining these results we get

$$\begin{aligned}\frac{\partial}{\partial w_{kl}}(W^TS_1W)^{-1}(W^TS_2W) = &-(W^TS_1W)^{-1}\left[V^T(k,l)S_1W + W^TS_1V(k,l)\right](W^TS_1W)^{-1}(W^TS_2W)\\ &+ (W^TS_1W)^{-1}\left[V^T(k,l)S_2W + W^TS_2V(k,l)\right]\,.\end{aligned}$$

Then for each term (there are four of them) once we take the trace we can write each one as $g_i(W)V(k,l)h_i(W)$ for functions $g_i(\cdot)$ and $h_i(\cdot)$ for $i = 1, 2, 3, 4$ by using

$$\text{trace}(W^T) = \text{trace}(W)\,,$$

if needed. We will need to use that identity for the first and third terms. We get

$$\begin{aligned}
g_1(W) &= -(W^TS_2W)(W^TS_1W)^{-1}W^TS_1\,, \quad \text{and} \quad h_1(W) = (W^TS_1W)^{-1}\\
g_2(W) &= -(W^TS_1W)^{-1}W^TS_1\,, \quad \text{and} \quad h_2(W) = (W^TS_1W)^{-1}(W^TS_2W)\\
g_3(W) &= W^TS_2\,, \quad \text{and} \quad h_3(W) = (W^TS_1W)^{-1}\\
g_4(W) &= (W^TS_1W)^{-1}W^TS_2\,, \quad \text{and} \quad h_4(W) = I\,.
\end{aligned}$$

Once we have done this we use Equation 3 (but without the transpose yet) to get

$$\begin{aligned}\left(\frac{\partial}{\partial W}\text{trace}\left\{(W^T S_1 W)^{-1}(W^T S_2 W)\right\}\right)^T &= -(W^T S_1 W)^{-1}(W^T S_2 W)(W^T S_1 W)^{-1} W^T S_1 \\ &- (W^T S_1 W)^{-1}(W^T S_2 W)(W^T S_1 W)^{-1} W^T S_1 \\ &+ (W^T S_1 W)^{-1} W^T S_2 \\ &+ (W^T S_1 W)^{-1} W^T S_2 \,.\end{aligned}$$

Thus taking the transpose of both sides we finally find

$$\begin{aligned}\frac{\partial}{\partial W}\text{trace}\left\{(W^T S_1 W)^{-1}(W^T S_2 W)\right\} &= -2 S_1 W (W^T S_1 W)^{-1}(W^T S_2 W)(W^T S_1 W)^{-1} \\ &+ 2 S_2 W (W^T S_1 W)^{-1} \,,\end{aligned}$$

as we were to show.

Chapter 7 (Probabilistic Separability Measures in Feature Selection)

Notes On The Text

Notes on Probabalistic Distance Measures: the divergence for Gaussians

When the conditional densities are Gaussian we have

$$\begin{aligned} p(x|\omega_i) &\sim N(\mu_i, \Sigma_i) \\ p(x|\omega_j) &\sim N(\mu_j, \Sigma_j) \,. \end{aligned}$$

Then to compute the divergence d_{ij} given by

$$d_{ij} = \int_{-\infty}^{\infty} (p(x|\omega_i) - p(x|\omega_j)) \ln\left(\frac{p(x|\omega_i)}{p(x|\omega_j)}\right) dx \,, \tag{5}$$

we first need to compute the log term $\ln\left(\frac{p(x|\omega_i)}{p(x|\omega_j)}\right)$, where we find

$$\ln\left(\frac{p(x|\omega_i)}{p(x|\omega_j)}\right) = -\frac{1}{2}\left[(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - (x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right] + \frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right) \,.$$

When we expand the quadratics above we get

$$\begin{aligned} \ln\left(\frac{p(x|\omega_i)}{p(x|\omega_j)}\right) &= -\frac{1}{2}\left[x^T\Sigma_i^{-1}x - x^T\Sigma_j^{-1}x - 2\mu_i^T\Sigma_i^{-1}x + 2\mu_j^T\Sigma_j^{-1}x\right] \\ &- \frac{1}{2}\left[\mu_i^T\Sigma_i^{-1}\mu_i - \mu_j^T\Sigma_j^{-1}\mu_j\right] + \frac{1}{2}\ln\left(\frac{|\Sigma_j|}{|\Sigma_i|}\right) \,. \end{aligned}$$

Only the first four terms depend on x while the remaining terms are independent of x and can be represented by a constant C. Because the densities $p(x|\cdot)$ are normalized we note that

$$\int_{-\infty}^{\infty} (p(x|\omega_i) - p(x|\omega_j)) C dx = C(1-1) = 0 \,,$$

and these terms do not affect the divergence. Thus we only need to worry about how to integrate the first four terms. To do these lets first consider the integral of these terms against $p(x|\omega_i)$ (integrating against $p(x|\omega_j)$ will be similar). To do these integral we will use Equation 30 from Appendix D to evaluate the integral of the terms like $x^T\Sigma^{-1}x$, against $p(x|\omega_i)$. When we do that we find the integral of the log ratio term expressed above is given by (multiplied by $-1/2$)

$$\begin{aligned} -2\int_{-\infty}^{\infty} \ln\left(\frac{p(x|\omega_i)}{p(x|\omega_j)}\right) p(x|\omega_i) dx &= \mu_i^T\Sigma_i^{-1}\mu_i + \text{trace}(\Sigma_i\Sigma_i^{-1}) - \mu_i^T\Sigma_j^{-1}\mu_i - \text{trace}(\Sigma_i\Sigma_j^{-1}) \\ &- 2\mu_i^T\Sigma_i^{-1}\mu_i + 2\mu_j^T\Sigma_j^{-1}\mu_i \\ &= -\mu_i^T\Sigma_i^{-1}\mu_i - \mu_i^T\Sigma_j^{-1}\mu_i + 2\mu_j^T\Sigma_j^{-1}\mu_i \\ &+ \text{trace}(I) - \text{trace}(\Sigma_i\Sigma_j^{-1}) \,. \end{aligned}$$

In the same way the integral of the log ratio term against $p(x|\omega_j)$ is given by

$$\begin{aligned}
2\int_{-\infty}^{\infty} \ln\left(\frac{p(x|\omega_i)}{p(x|\omega_j)}\right) p(x|\omega_j)dx &= -\mu_j^T\Sigma_i^{-1}\mu_j - \text{trace}(\Sigma_j\Sigma_i^{-1}) + \mu_j^T\Sigma_j^{-1}\mu_j + \text{trace}(\Sigma_j\Sigma_j^{-1}) \\
&+ 2\mu_i^T\Sigma_i^{-1}\mu_j - 2\mu_j^T\Sigma_j^{-1}\mu_j \\
&= -\mu_j^T\Sigma_j^{-1}\mu_j - \mu_j^T\Sigma_i^{-1}\mu_j + 2\mu_i^T\Sigma_i^{-1}\mu_j \\
&+ \text{trace}(I) - \text{trace}(\Sigma_j\Sigma_i^{-1})\,.
\end{aligned}$$

If we take -1 of the first and second expression and add them together we get two types of terms. Terms involving the trace operation and terms that don't depend on the trace. The trace terms add to give

$$\begin{aligned}
\text{trace terms} &= -\text{trace}(I) + \text{trace}(\Sigma_i\Sigma_j^{-1}) - \text{trace}(I) + \text{trace}(\Sigma_j\Sigma_i^{-1}) \\
&= -2\text{trace}(I) + \text{trace}(\Sigma_i\Sigma_j^{-1}) + \text{trace}(\Sigma_j\Sigma_i^{-1})\,.
\end{aligned}$$

The non-trace terms add together to give

$$\begin{aligned}
\text{non-trace terms} &= \mu_i^T\Sigma_i^{-1}\mu_i + \mu_i^T\Sigma_j^{-1}\mu_i - 2\mu_j^T\Sigma_j^{-1}\mu_i \\
&+ \mu_j^T\Sigma_j^{-1}\mu_j + \mu_j^T\Sigma_i^{-1}\mu_j - 2\mu_i^T\Sigma_i^{-1}\mu_j \\
&= \mu_i^T(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_i - 2\Sigma_j^{-1}\mu_j - 2\Sigma_i^{-1}\mu_j) + \mu_j^T(\Sigma_i^{-1}\mu_j + \Sigma_j^{-1}\mu_j) \\
&= \mu_i^T((\Sigma_i^{-1} + \Sigma_j^{-1})\mu_i - 2(\Sigma_i^{-1} + \Sigma_j^{-1})\mu_j) + \mu_j^T(\Sigma_i^{-1} + \Sigma_j^{-1})\mu_j \\
&= \mu_i^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) - \mu_i^T(\Sigma_i^{-1} + \Sigma_j^{-1})\mu_j + \mu_j^T(\Sigma_i^{-1} + \Sigma_j^{-1})\mu_j \\
&= \mu_i^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) - (\mu_i^T - \mu_j^T)(\Sigma_i^{-1} + \Sigma_j^{-1})\mu_j \\
&= (\mu_i - \mu_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})\mu_i - (\mu_i - \mu_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})\mu_j \\
&= (\mu_i - \mu_j)(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)\,.
\end{aligned}$$

In total when we divide by 2 and add together the trace and the non-trace expressions we get

$$d_{ij} = \frac{1}{2}(\mu_i - \mu_j)(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) + \frac{1}{2}\text{trace}(\Sigma_i\Sigma_j^{-1} + \Sigma_j\Sigma_i^{-1} - 2I)\,, \tag{6}$$

for the expression for the divergence between two Gaussians.

Chapter 8 (Feature Extraction Methods based on Probabilistic Separability Measures)

Notes On The Text

The Chernoff Probabilistic Distance Measure

To begin, recall that our linearly transformed vectors x are obtained from our raw input vectors y via the linear mapping $x = W^t y$. Now as $J_C(W)$ is a scalar we can take the trace of its expression and use the properties of the trace operator to simplify some resulting expressions. Taking this trace, the Chernoff separability measure can be transformed as

$$\begin{aligned} J_C(W) &= \frac{1}{2}s(1-s)\text{tr}\left\{(\mu_2-\mu_1)^t W\left[(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W\right]^{-1} W^t(\mu_2-\mu_1)\right\} \quad (7)\\ &+ \frac{1}{2}\ln|(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W| \\ &- \frac{1-s}{2}\ln|W^t\Sigma_1 W| - \frac{s}{2}\ln|W^t\Sigma_2 W|\,. \end{aligned}$$

To simplify the notation we will define the matrix L as

$$L \equiv \left[(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W\right]^{-1}\,.$$

Since we can cyclically permute the arguments of a trace we have that the first term in Equation 7 can be written as

$$\text{tr}\left\{(\mu_2-\mu_1)^t WLW^t(\mu_2-\mu_1)\right\} = \text{tr}\left\{W^t(\mu_2-\mu_1)(\mu_2-\mu_1)^t WL\right\}\,.$$

As in the book lets also define the matrix M as $M \equiv (\mu_2-\mu_1)(\mu_2-\mu_1)^t$. When we do these two steps we have the books equation 21. To evaluate the value of W at which we will maximize the value of $J_C(W)$ we need to compute the first derivative of the scalar $J_C(W)$ with respect to the matrix W. To compute this lets take the derivative of each term one at a time. To take the derivative of the first term using the results from Page 14 onward where we evaluate the derivative of a trace with respect to a matrix. We first take the derivative of $W^t MWL$ with respect to the element w_{kl}. Using the product rule we find

$$\begin{aligned} \frac{\partial(W^tMWL)}{\partial w_{kl}} &= V^t(k,l)MWL + W^tMV(k,l)L + W^tMW\frac{\partial L}{\partial w_{kl}} \\ &= V^t(k,l)MWL + W^tMV(k,l)L \\ &- W^tMW \\ &\times L\left\{(1-s)V^t(k,l)\Sigma_1 W + (1-s)W^t\Sigma_1 V(k,l) + sV^t(k,l)\Sigma_2 W + sW^t\Sigma_2 V(k,l)\right\}L\,. \end{aligned}$$

Defining d_{kl} as $d_{kl} \equiv \frac{\partial\,\text{tr}(W^tMWL)}{\partial w_{kl}}$ then d_{kl} is given by taking the trace of the above expression. After taking this trace we cyclically permute the L matrix in the second and third trace to find

$$\begin{aligned} d_{kl} &= \text{tr}(V^t(k,l)MWL) + \text{tr}(LW^tMV(k,l)) \\ &- \text{tr}\left[LW^tMWL\left\{(1-s)V^t(k,l)\Sigma_1 W + (1-s)W^t\Sigma_1 V(k,l) + sV^t(k,l)\Sigma_2 W + sW^t\Sigma_2 V(k,l)\right\}\right]\,. \end{aligned}$$

We can now use the two facts that traces are invariant to transposes and to cyclic permutations of their arguments followed by the trace element selection lemmas given by Equations 28 and 29 to simplify the above expression. The first two terms transform as

$$\begin{aligned} \operatorname{tr}(V^t(k,l)MWL) + \operatorname{tr}(LW^tMV(k,l)) &= (MWL)_{kl} + (LW^tM)_{lk} \\ &= (MWL)_{kl} + (M^tWL^t)_{kl} \,. \end{aligned} \tag{8}$$

The last term in d_{kl} (temporally denoted L_{kl}) transforms as

$$\begin{aligned} L_{kl} &= -(1-s)\operatorname{tr}(LW^tMWLV^t(k,l)\Sigma_1 W) - (1-s)\operatorname{tr}(LW^tMWLW^t\Sigma_1 V(k,l)) \\ &- s\operatorname{tr}(LW^tMWLV^t(k,l)\Sigma_2 W) - s\operatorname{tr}(LW^tMWLW^t\Sigma_2 V(k,l)) \\ &= -(1-s)\operatorname{tr}(V^t(k,l)\Sigma_1 WLW^tMWL) - (1-s)(LW^tMWLW^t\Sigma_1)_{lk} \\ &- s\operatorname{tr}(V^t(k,l)\Sigma_2 WLW^tMWL) - s(LW^tMWLW^t\Sigma_2)_{lk} \\ &= -(1-s)(\Sigma_1 WLW^tMWL)_{kl} - (1-s)(\Sigma_1^t WL^tW^tM^tWL^t)_{kl} \\ &- s(\Sigma_2 WLW^tMWL)_{kl} - s(\Sigma_2^t WL^tW^tM^tW)_{kl} \,. \end{aligned}$$

Now Σ_1 and Σ_2 are covariance matrices and so are symmetric. Because of this L is symmetric also and M is symmetric by its definition. Thus using the symmetry of the matrices involved L_{kl} becomes

$$L_{kl} = -2(1-s)(\Sigma_1 WLW^tMWL)_{kl} - 2s(\Sigma_2 WLW^tMWL)_{kl} \,.$$

Including the first two terms given by Equation 8 we have

$$\frac{\partial \operatorname{tr}(W^tMWL)}{\partial w_{kl}} = 2(MWL)_{kl} - 2(1-s)(\Sigma_1 WLW^tMWL)_{kl} - 2s(\Sigma_2 WLW^tMWL)_{kl} \,.$$

So the *matrix* derivative then becomes (which is obtained by removing the index notation of w_{kl} above)

$$\begin{aligned} \frac{\partial \operatorname{tr}(W^tMWL)}{\partial W} &= 2MWL - 2(1-s)\Sigma_1 WLW^tMWL - 2s\Sigma_2 WLW^tMWL \\ &= 2MWL - 2((1-s)\Sigma_1 W + s\Sigma_2 W)LW^tMWL \,. \end{aligned}$$

Next we need to take the derivatives with respect to W of the expressions

$$\ln|(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W| \,, \ln|W^t\Sigma_1 W| \,, \text{and } \ln|W^t\Sigma_2 W| \,.$$

Since we can write

$$(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W = W^t((1-s)\Sigma_1 + s\Sigma_2)W \,,$$

all of these expressions have similar derivatives. Taking the derivative of $\ln|W^t\Sigma_2 W|$ to see that these equal we have

$$\begin{aligned} \frac{\partial \ln|W^t\Sigma_2 W|}{\partial W} &= \frac{1}{|W^t\Sigma_2 W|}\frac{\partial |W^t\Sigma_2 W|}{\partial W} \\ &= \frac{1}{|W^t\Sigma_2 W|}|W^t\Sigma_2 W| \left\{ \Sigma_2 W(W^t\Sigma_2 W)^{-1} + {\Sigma_2}^t W[(W^t\Sigma_2 W)^{-1}]^t \right\} \\ &= 2\Sigma_2 W(W^t\Sigma_2 W)^{-1} \,. \end{aligned}$$

All the other derivatives are similar. For example

$$\begin{aligned}\frac{\partial \ln |(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W|}{\partial W} &= 2((1-s)\Sigma_1 + s\Sigma_2)W(W^t((1-s)\Sigma_1 + s\Sigma_2)W)^{-1} \\ &= 2((1-s)\Sigma_1 W + s\Sigma_2 W)((1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W)^{-1} \\ &= 2((1-s)\Sigma_1 W + s\Sigma_2 W)L\,.\end{aligned}$$

Thus combining everything we finally find that our first derivative of $J_C(W)$ is given by

$$\begin{aligned}\frac{\partial J_C(W)}{\partial W} &= s(1-s)\left\{MWL - ((1-s)\Sigma_1 W + s\Sigma_2 W)LW^tMWL\right\} \\ &+ ((1-s)\Sigma_1 W + s\Sigma_2 W)L \\ &- (1-s)\Sigma_1 W(W^t\Sigma_1 W)^{-1} - s\Sigma_2 W(W^t\Sigma_2 W)^{-1}\,,\end{aligned}$$

which is the books equation 23. To find the maximum of $J_C(W)$ we set $J_C'(W)$ equal to zero to derive an equation for W which would then need to be solved for W. Assuming L is non-singular we can premultiply by L^{-1} and divide by $s(1-s)$ to get the following equation equivalent to $J_C'(W) = 0$

$$MW - [(1-s)\Sigma_1 W + s\Sigma_2 W]LW^tMW + \frac{1}{s(1-s)}[(1-s)\Sigma_1 W + s\Sigma_2 W] \tag{9}$$

$$- \frac{1}{s}\Sigma_1 W(W^t\Sigma_1 W)^{-1}[(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W] \tag{10}$$

$$- \frac{1}{1-s}\Sigma_2 W(W^t\Sigma_2 W)^{-1}[(1-s)W^t\Sigma_1 W + sW^t\Sigma_2 W] = 0\,. \tag{11}$$

Note that in the above the last two terms can be expanded and written as

$$-\frac{1}{s}\Sigma_1 W[(1-s)I + s(W^t\Sigma_1 W)^{-1}W^t\Sigma_2 W] - \frac{1}{1-s}\Sigma_2 W[(1-s)(W^t\Sigma_2 W)^{-1}W^t\Sigma_1 W + sI]\,,$$

which once done gives the books equation 24. We can further combine these terms with $\frac{1}{s(1-s)}[(1-s)\Sigma_1 W + s\Sigma_2 W]$ the third term on line 9 above to get

$$\begin{aligned}& \Sigma_1 W + \Sigma_2 W - \Sigma_1 W(W^t\Sigma_1 W)^{-1}W^t\Sigma_2 W - \Sigma_2 W(W^t\Sigma_2 W)^{-1}W^t\Sigma_1 W \\ =\ & \Sigma_1 W[I - (W^t\Sigma_1 W)^{-1}W^t\Sigma_2 W] + \Sigma_2 W[I - (W^t\Sigma_2 W)^{-1}W^t\Sigma_1 W]\,,\end{aligned}$$

which gives the books equation 25 repeated here for convenience

$$\begin{aligned}& MW - [(1-s)\Sigma_1 W + s\Sigma_2 W]LW^tMW \\ +\ & \Sigma_1 W[I - (W^t\Sigma_1 W)^{-1}W^t\Sigma_2 W] + \Sigma_2 W[I - (W^t\Sigma_2 W)^{-1}W^t\Sigma_1 W] = 0\end{aligned} \tag{12}$$

In the special case where both Gaussians have the *same* covariance matrix then $\Sigma_1 = \Sigma_2 = \Sigma$ and the last two terms in Equation 12 vanish so that Equation 12 becomes

$$MW - \Sigma W(W^t\Sigma W)^{-1}W^tMW = 0\,, \tag{13}$$

which is the books equation 28. Performing an eigenvector-eigenvalue decomposition of $(W^t\Sigma W)^{-1}W^tMW$ to write this matrix as $U\Lambda U^{-1}$ we have

$$MW - \Sigma WU\Lambda U^{-1} = 0\,,$$

or

$$\Sigma^{-1}MWU - WU\Lambda = 0\,. \tag{14}$$

Now since we don't explicitly know the value W we don't explicitly know the values of Λ or U and they are effectively functions of the matrix W. We will see below how to compute these matrices. On grouping some terms together we find

$$\Sigma^{-1}M(WU) = (WU)\Lambda\,.$$

Since Λ is a *diagonal* matrix containing the eigenvalues of $(W^t\Sigma W)^{-1}W^tMW$, multiplication on the right by this matrix Λ is equivalent to multiplying each column of the matrix WU by the corresponding eigenvalue. Comparing each side of this equation column by column we see that the columns of WU must be the eigenvectors of the matrix $\Sigma^{-1}M$. From this we can compute the eigenvectors of $\Sigma^{-1}M$ and place them as columns of the matrix say V. Then since $V = WU$ for some as yet unknown U, W would be given by $W = VU^{-1}$. The point to note now is that in fact the multiplication of V by an invertible matrix U^{-1} does not in fact change the value of J_C and it can be ignored. The fact that multiplication by U^{-1} on the left does not change the value of J_C can be seen by first considering $J_C(W)$ with equal covariance matrices. We find

$$\begin{aligned}
J_C(W) &= \frac{1}{2}s(1-s)\mathrm{tr}\left\{W^tMW(W^t\Sigma W)^{-1}\right\} \\
&+ \frac{1}{2}\ln|W^t\Sigma W| - \frac{1-s}{2}\ln|W^t\Sigma W| - \frac{s}{2}\ln|W^t\Sigma W| \\
&= \frac{1}{2}s(1-s)\mathrm{tr}\left\{W^tMW(W^t\Sigma W)^{-1}\right\}\,.
\end{aligned}$$

If we consider $J_C(WU^{-1})$ we find

$$\begin{aligned}
J_C(WU^{-1}) &= \frac{1}{2}s(1-s)\mathrm{tr}\left\{U^{-t}W^tMWU^{-1}(U^{-t}W^t\Sigma WU^{-1})^{-1}\right\} \\
&= \frac{1}{2}s(1-s)\mathrm{tr}\left\{U^{-t}W^tMWU^{-1}U(W^t\Sigma W)^{-1}U^t\right\} \\
&= \frac{1}{2}s(1-s)\mathrm{tr}\left\{U^tU^{-t}W^tMW(W^t\Sigma W)^{-1}\right\} = J_C(W)\,.
\end{aligned}$$

Since the value of U^{-1} does not matter we can effectively ignore this matrix (take $U = I$). Then the matrix W has columns that are simply the eigenvectors of $\Sigma^{-1}M$. When we use the decomposition $(W^t\Sigma W)^{-1}W^tMW = U\Lambda U^{-1}$ in the above expression for $J_C(W)$ we find

$$\begin{aligned}
J_C(W) &= \frac{1}{2}s(1-s)\mathrm{tr}\left\{W^tMW(W^t\Sigma W)^{-1}\right\} \\
&= \frac{1}{2}s(1-s)\mathrm{tr}\left\{U\Lambda U^{-1}\right\} = \frac{1}{2}s(1-s)\mathrm{tr}\left\{U^{-1}U\Lambda\right\} \\
&= \frac{1}{2}s(1-s)\mathrm{tr}\left\{\Lambda\right\}\,.
\end{aligned}$$

Since M is of rank one the product $\Sigma^{-1}M$ will also be of rank one and only have *one* non-zero eigenvalue. Thus the matrix W will thus in fact be only a column vector and not a matrix. To find its value we could explicitly compute the non-zero eigenvector of $\Sigma^{-1}M$, but an easier method it to write Equation 14 when W is a column vector say v_1 as

$$\Sigma^{-1}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^t v_1 = \lambda v_1\,.$$

Since $(\mu_2 - \mu_1)^t v_1$ is an inner product and is therefore a scalar we can factor it out to get

$$\left((\mu_2 - \mu_1)^t v_1\right) \Sigma^{-1}(\mu_2 - \mu_1) = \lambda v_1 \, .$$

Comparing vectors (and the corresponding multiplying scalars) on each side of this expression we see that

$$\begin{aligned} v_1 &= \Sigma^{-1}(\mu_2 - \mu_1) \\ \lambda &= (\mu_2 - \mu_1)^t v_1 = (\mu_2 - \mu_1)^t \Sigma^{-1}(\mu_2 - \mu_1) \, . \end{aligned} \tag{15}$$

Equation 15 is the optimal Chernoff feature transformation

$$x = v_1^t y = (\mu_2 - \mu_1)^t \Sigma^{-1} y \, , \tag{16}$$

for the case when the covariance matrices of the two densities are equal.

Chapter 9 (Feature Extraction based on the Karhunen-Loeve Expansion)

Problem Solutions

Problem 6 (a constrained optimization problem)

We want to find the extrema of $f = \prod_{i=1}^d \sigma_i$ for $0 \le \sigma_i \le 1$ subject to the constraint that $\sum_{i=1}^d \sigma_i = 1$. Maximizing f is equivalent to maximizing $\log(f)$ or $\sum_{i=1}^d \log(\sigma_i)$, since $\log(\cdot)$ is a monotone increasing function. Since this is a constrained optimization problem we will use the method of Lagrange multipliers to find its solution. We first form the Lagrangian

$$\mathcal{L}(\{\sigma_i\};\lambda) = \sum_{i=1}^d \log(\sigma_i) - \lambda\left(\sum_{i=1}^d \sigma_i - 1\right) ,$$

and then look for stationary points with respect to $(\{\sigma_i\},\lambda)$. We have

$$\frac{\partial \mathcal{L}}{\partial \sigma_i} = \frac{1}{\sigma_i} - \lambda = 0 \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum_{i=1}^d \sigma_i - 1 = 0\,. \tag{18}$$

Using Equation 17 we have $\sigma_i = \frac{1}{\lambda}$, which when we put into Equation 17 gives

$$\frac{d}{\lambda} - 1 = 0 \quad \text{or} \quad \lambda = d\,.$$

From this we then have that

$$\sigma_i = \frac{1}{d}\,.$$

Technically the optimization procedure we did above is valid for *unconstrained* problems where there are no restrictions on the values of σ_i or λ. In this problem since $0 \le \sigma_i \le 1$ the solution we obtain above is only valid if the solution we obtain *also* satisfies these constraints. In this case it does and we have in fact found a *global* minimum.

Differentiation of Scalar Functions of a Matrix Variable

Notes On The Text

The Derivative of the Trace Operator: Example 3: $J(w) = \mathrm{tr}(W^t M W)$

From earlier in this appendix we have that the w_{kl} derivative of $W^t M W$ is given by

$$\frac{\partial (W^t M W)}{\partial w_{kl}} = V^t(k,l) M W + W^t M V(k,l) \,. \tag{19}$$

I found it difficult to directly use the results from the text as there seems no easy way to write this right-hand-side in the required stated form of

$$\frac{\partial F(W)}{\partial w_{kl}} = \sum_i g_i(W) V(k,l) h_i(W) \,, \tag{20}$$

due to the $V^t(k,l)$ term on the right-hand-side of Equation 19. Writing $\frac{\partial F(W)}{\partial w_{kl}}$ exactly as Equation 20, however, does not seem to be the correct requirement, since what we really need is to be able to write $\frac{\partial J(W)}{\partial w_{kl}}$ as

$$\frac{\partial J(W)}{\partial w_{kl}} = \mathrm{tr}\left(\sum_i g_i(W) V(k,l) h_i(W) \right) \,. \tag{21}$$

Note that that $J(W)$ is the argument of the derivative and the *trace* operation in Equation 21 which is not present in Equation 20. For the example of $J(W) = \mathrm{tr}(W^t M W)$ given here, this can be done using some properties of the trace operator as follows

$$\begin{aligned}
\frac{\partial J(W)}{\partial w_{kl}} &= \frac{\partial\, \mathrm{tr}(W^t M W)}{\partial w_{kl}} \\
&= \mathrm{tr}\left(\frac{\partial (W^t M W)}{\partial w_{kl}} \right) \\
&= \mathrm{tr}(V^t(k,l) M W) + \mathrm{tr}(W^t M V(k,l)) \\
&= \mathrm{tr}(W^t M^t V(k,l)) + \mathrm{tr}(W^t M V(k,l)) \\
&= \mathrm{tr}(M^t V(k,l) W^t) + \mathrm{tr}(M V(k,l) W^t) \\
&= \mathrm{tr}\left(M^t V(k,l) W^t + M V(k,l) W^t \right) \,,
\end{aligned}$$

which is in the form needed by Equation 21. In this situation we have

$$\begin{aligned}
g_1(W) &= M^t \\
h_1(W) &= W^t \\
g_2(W) &= M \\
h_2(W) &= W^t \,,
\end{aligned}$$

so that

$$\begin{aligned}
f_1(W) &= h_1(W) g_1(W) = W^t M^t \\
f_2(W) &= h_2(W) g_2(W) = W^t M \,.
\end{aligned}$$

Then from our theorem that

$$\frac{\partial J}{\partial W} = \left(\sum_i f_i(W) \right)^t = \left(\sum_i h_i(W) g_i(W) \right)^t , \tag{22}$$

in this case gives

$$\begin{aligned} \frac{\partial \text{tr}(W^t M W)}{\partial W} &= \left(W^t M^t + W^t M \right)^t = MW + M^t W \\ &= (M + M^t) W \, , \end{aligned} \tag{23}$$

which is the expression given in the book.

Using the above logic we can also derive derivative results for similar expressions. For example a slight variation on the above derivative would be the evaluation of

$$\frac{\partial \text{tr}(W M W^t)}{\partial W} .$$

Here our objective function $J(\cdot)$ is given by $J(W) \equiv \text{tr}(WMW^t)$. Note the different location of the transpose operation "t" between these two examples. The first example above was computing the derivative of the "inner product" form $W^t M W$, while this example is to compute the derivative of the "outer product" form WMW^t.

Following the steps above from earlier in this appendix we have that the w_{kl} derivative of the expression WMW^t is given by

$$\frac{\partial (W M W^t)}{\partial w_{kl}} = V(k,l) M W^t + W M V(k,l)^t \, .$$

From this we have

$$\begin{aligned} \frac{\partial \, \text{tr}(WMW^t)}{\partial w_{kl}} &= \text{tr} \left(\frac{\partial (WMW^t)}{\partial w_{kl}} \right) \\ &= \text{tr}(V(k,l) M W^t) + \text{tr}(W M V(k,l)^t) \\ &= \text{tr}(W^t V(k,l) M) + \text{tr}(V(k,l) M^t W^t) \\ &= \text{tr}(W^t V(k,l) M) + \text{tr}(W^t V(k,l) M^t) \, . \end{aligned}$$

This is in the form needed by Equation 21. In this situation we have

$$\begin{aligned} g_1(W) &= W^t \\ h_1(W) &= M \\ g_2(W) &= W^t \\ h_2(W) &= M^t \, , \end{aligned}$$

so that

$$\begin{aligned} f_1(W) &= h_1(W) g_1(W) = M W^t \\ f_2(W) &= h_2(W) g_2(W) = M^t W^t \, . \end{aligned}$$

Using these and applying our theorem given by Equation 22 we have that

$$\begin{aligned}\frac{\partial \mathrm{tr}(WMW^t)}{\partial W} &= \left(MW^t + M^tW^t\right)^t = WM^t + WM \\ &= W(M + M^t)\,. \qquad (24)\end{aligned}$$

Note that this expression is *not* simply the transpose of Equation 23 if W is not symmetric.

The Derivative of the Trace Operator: Example 4: $J(W) = \mathrm{tr}((W^tMW)^{-1})$

From example 2 in this appendix we have that

$$\frac{\partial\left((W^t\Sigma W)^{-1}\right)}{\partial w_{kl}} = -(W^t\Sigma W)^{-1}V^t(k,l)\Sigma W(W^t\Sigma W)^{-1} - (W^t\Sigma W)^{-1}W^t\Sigma V(k,l)(W^t\Sigma W)^{-1}\,.$$

So the derivative of the *trace* of this expression denoted as D so that

$$D \equiv \frac{\partial\,\mathrm{tr}((W^t\Sigma W)^{-1})}{\partial w_{kl}}\,,$$

then becomes

$$\begin{aligned}D &= -\mathrm{tr}((W^t\Sigma W)^{-1}V^t(k,l)\Sigma W(W^t\Sigma W)^{-1}) - \mathrm{tr}((W^t\Sigma W)^{-1}W^t\Sigma V(k,l)(W^t\Sigma W)^{-1}) \\ &= -\mathrm{tr}((W^t\Sigma W)^{-1}W^t\Sigma^t V(k,l)(W^t\Sigma W)^{-1}) - \mathrm{tr}((W^t\Sigma W)^{-1}W^t\Sigma V(k,l)(W^t\Sigma W)^{-1}) \\ &= -\mathrm{tr}((W^t\Sigma^t W)^{-2}W^t\Sigma^t V(k,l)) - \mathrm{tr}((W^t\Sigma W)^{-2}W^t\Sigma V(k,l))\,.\end{aligned}$$

So matching terms with Equation 21 we have in this case that

$$\begin{aligned}g_1(W) &= -(W^t\Sigma^t W)^{-2}W^t\Sigma^t \\ h_1(W) &= I \\ g_2(W) &= -(W^t\Sigma^t W)^{-2}W^t\Sigma \\ h_2(W) &= I\,,\end{aligned}$$

so that

$$\begin{aligned}f_1(W) &= h_1(W)g_1(W) = -(W^t\Sigma^t W)^{-2}W^t\Sigma^t \\ f_2(W) &= h_2(W)g_2(W) = -(W^t\Sigma^t W)^{-2}W^t\Sigma\,.\end{aligned}$$

Thus our theorem given by Equation 22 we have that

$$\begin{aligned}\frac{\partial J(W)}{\partial W} &= -\left((W^t\Sigma^t W)^{-2}W^t\Sigma^t + (W^t\Sigma^t W)^{-2}W^t\Sigma\right)^t \\ &= -\Sigma W(W^t\Sigma^t W)^{-2} - \Sigma^t W(W^t\Sigma^t W)^{-2}\,,\end{aligned}$$

which is the negative of the result given in the book but I believe is correct.

The Derivative of the Determinant

From the notes in this section it is clear that when $J(W) = |F(W)|$ with $F(W)$ a matrix valued function of a matrix W that the derivative with respect to the element w_{kl} is given by

$$\frac{\partial J(W)}{\partial w_{kl}} = \sum_{i=1}^{r}\sum_{j=1}^{r} F^*_{ij}\frac{\partial F_{ij}(W)}{\partial w_{kl}} \,. \tag{25}$$

From the definition of the adjoint matrix, $F^*(W)$, the (j,j)th element of the matrix product $F^*(W)\frac{\partial F}{\partial w_{kl}}$ is given by

$$\sum_{n=1}^{r} F^*_{nj}\frac{\partial F_{nj}(W)}{\partial w_{kl}} \,. \tag{26}$$

Note that in the way that F^* is defined, the sum over n is done with respect to the *first* index of the matrix $F^*(W)$. This is a somewhat different ordering that what might be expected for the (j,j)th element arising from traditional matrix indexing. The trace of the matrix product $F^*(W)\frac{\partial F(W)}{\partial w_{kl}}$ is given by summing over j in Equation 26, which by using Equation 25 gives

$$\frac{\partial J(W)}{\partial w_{kl}} = \sum_{j=1}^{r}\sum_{n=1}^{r} F^*_{nj}\frac{\partial F_{nj}(W)}{\partial w_{kl}} = \operatorname{tr}\left[F^*(W)\frac{\partial F(W)}{\partial w_{kl}}\right] \,.$$

Since $F^*(W) = |F(W)|F^{-1}(W)$, this can be written as

$$\frac{\partial J(W)}{\partial w_{kl}} = |F(W)|\operatorname{tr}\left[F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}}\right] \,. \tag{27}$$

Thus the problem of differentiating the determinant has been reduced to that of finding the *trace* of a matrix function.

The Derivative of the Determinant Operator: Example 5: $J(w) = |W^tMW|$

Then for this example we have

$$\begin{aligned} F^{-1}(W) &= (W^tMW)^{-1} \quad \text{and} \\ \frac{\partial F(W)}{\partial w_{kl}} &= V^t(k,l)MW + W^tMV(k,l) \,, \end{aligned}$$

so that the product $F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}}$ becomes simply

$$F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}} = F^{-1}(W)V^t(k,l)MW + F^{-1}(W)W^tMV(k,l) \,,$$

so distributing the trace operation the trace of this expression is given by

$$\operatorname{tr}\left(F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}}\right) = \operatorname{tr}(F^{-1}(W)V^t(k,l)MW) + \operatorname{tr}(F^{-1}(W)W^tMV(k,l)) \,.$$

To further evaluate this expression we recall the result that if we take the trace of a matrix product of the form $AV(k,l)$ then we get the (l,k)th element of A. This is the ***trace element selection lemma*** or equation 17 in the book's appendix and is given by

$$\mathrm{tr}(AV(k,l)) = A_{lk}\,, \tag{28}$$

note the *order* of the subscripts in this last expression. Some of the terms in the above expression for $\mathrm{tr}\left(F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}}\right)$ are directly of this form. The others will be transformed into this form. By taking a transpose and then cyclically permuting the matrices in the first trace expression followed by using Equation 28 on the second trace expression, we find

$$\begin{aligned}
\mathrm{tr}\left(F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}}\right) &= \mathrm{tr}(W^tM^tV(k,l)F^{-t}(W)) + (F^{-1}(W)W^tM)_{lk} \\
&= \mathrm{tr}(F^{-t}(W)W^tM^tV(k,l)) + (F^{-1}(W)W^tM)_{lk} \\
&= (F^{-t}(W)W^tM^t)_{lk} + (F^{-1}(W)W^tM)_{lk} \\
&= (F^{-t}(W)W^tM^t + F^{-1}(W)W^tM)_{lk}\,.
\end{aligned}$$

Based on observing these manipulations we can use properties of the trace operator to derive the following alternative form to the above lemma (here the matrix V is on the left of A)

$$\begin{aligned}
\mathrm{tr}(V^t(k,l)A) &= \mathrm{tr}(A^tV(k,l)) \\
&= (A^t)_{lk} \\
&= A_{kl}\,,
\end{aligned} \tag{29}$$

using this form, with $V^t(k,l)$ on the left of A (if we can recognize it when it appears) can save us some algebraic manipulations. Since we ultimately want $\frac{\partial J(W)}{\partial w_{kl}}$ using Equation 27 we have

$$\frac{\partial J(W)}{\partial w_{kl}} = |W^tMW|(F^{-t}(W)W^tM^t + F^{-1}(W)W^tM)_{lk}\,.$$

So computing $\frac{\partial J(W)}{\partial W}$, where the denominator now is a matrix we need to take the transpose of the right-hand-side to get the subscripts to be correct. We find

$$\begin{aligned}
\frac{\partial J(W)}{\partial W} &= |W^tMW|(MWF^{-1}(W) + M^tWF^{-t}(W)) \\
&= |W^tMW|\left(MW(W^tMW)^{-1} + M^tW[(W^tMW)^{-1}]^t\right)\,,
\end{aligned}$$

the same as the expression given in the book.

The Derivative of the Determinant Operator: Example 6: $J(w) = |(W^tMW)^{-1}|$

When $J(W) = |(W^tMW)^{-1}|$ we have $F(W) = (W^t\Sigma W)^{-1}$ and we first need to evaluate $\frac{\partial F(W)}{\partial w_{kl}}$. From earlier we have

$$\frac{\partial F(W)}{\partial w_{kl}} = -[W^t\Sigma W]^{-1}(V^t(k,l)\Sigma W + W^t\Sigma V(k,l))[W^t\Sigma W]^{-1}\,,$$

so the required trace expression from Equation 27 becomes

$$\begin{aligned}\mathrm{tr}\left(F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}}\right) &= -\mathrm{tr}(V^t(k,l)\Sigma W F(W)) - \mathrm{tr}(W^t\Sigma V(k,l)F(W)) \\ &= -\mathrm{tr}(F^t(W)W^t\Sigma^t V(k,l)) - \mathrm{tr}(F(W)W^t\Sigma V(k,l))\,.\end{aligned}$$

Using the trace element selection lemma Equation 28 we have

$$\mathrm{tr}\left(F^{-1}(W)\frac{\partial F(W)}{\partial w_{kl}}\right) = -(F^t(W)W^t\Sigma^t)_{lk} - (F(W)W^t\Sigma)_{lk}\,.$$

so that $\frac{\partial J(W)}{\partial w_{kl}}$ is given by Equation 27 as

$$\frac{\partial J(W)}{\partial w_{kl}} = -|F(W)|(F^t(W)W^t\Sigma^t + F(W)W^t\Sigma)_{lk}\,.$$

To get $\frac{\partial J(W)}{\partial W}$ where the denominator is a matrix so the entire expression is a matrix using the component derivatives computed above we need to transpose the above right-hand-side matrix to get

$$\begin{aligned}\frac{\partial J(W)}{\partial W} &= -|F(W)|(\Sigma W F(W) + \Sigma^t W F(W)^t) \\ &= -|[W^t\Sigma W]^{-1}|\left(\Sigma W(W^t\Sigma W)^{-1} + \Sigma^t W[(W^t\Sigma W)^{-1}]^t\right)\,,\end{aligned}$$

which is the result given in the book.

Hints from Probability and Statistics

Moments of a Quadratic Form

Suppose x is a $l \times 1$ random vector with $E[x] = \mu$ and $\mathrm{Cov}(x) = \Sigma$ and let A be a $l \times l$ symmetric matrix not dependent on x then the quadratic expectation $E[x^T A x]$ is given by

$$E[x^T A x] = \mu^T A \mu + \mathrm{trace}(\Sigma A)\,. \tag{30}$$

References

[1] W. G. Kelley and A. C. Peterson. *Difference Equations. An Introduction with Applications.* Academic Press, New York, 1991.